Universidad de Sevilla

Departamento de Lenguajes y Sistemas Informáticos
Escuela Técnica Superior de Ingeniería Informática

Evolutionary Algorithms to Discover Quantitative Association Rules

Tesis Doctoral

María del Mar Martínez Ballesteros

Sevilla, Diciembre de 2011

UNIVERSIDAD DE SEVILLA

Evolutionary Algorithms to Discover Quantitative Association Rules

MEMORIA QUE PRESENTA

María del Mar Martínez Ballesteros

PARA OPTAR AL GRADO DE DOCTOR EN INFORMÁTICA
CON MENCIÓN DE DOCTORADO INTERNACIONAL
POR LA UNIVERSIDAD DE SEVILLA

DIRECTORES

Dr. José C. Riquelme Santos y
Dra. Alicia Troncoso Lora

Departamento de Lenguajes y Sistemas Informáticos
Escuela Técnica Superior de Ingeniería Informática

Diciembre de 2011

Dr. José C. Riquelme Santos, profesor Catedrático de Universidad adscrito al departamento de Lenguajes y Sistemas Informáticos de la Universidad de Sevilla, y Dra. Alicia Troncoso Lora, profesora Titular de Universidad adscrita al área de Lenguajes y Sistemas Informáticos de la Universidad Pablo de Olavide de Sevilla,

CERTIFICAN QUE:

Dña. María del Mar Martínez Ballesteros, ha realizado bajo su supervisión el trabajo de investigación titulado:

EVOLUTIONARY ALGORITHMS TO DISCOVER
QUANTITATIVE ASSOCIATION RULES

Una vez revisado, autorizan la presentación del mismo como tesis doctoral en la Universidad de Sevilla y estiman oportuna su presentación al tribunal para su valoración. Dicha tesis ha sido realizada dentro del programa de doctorado *Tecnología e Ingeniería del Software,* con mención de excelencia MEE2011-0129 del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Sevilla. Igualmente, autorizan la presentación para obtener la acreditación de Doctorado Internacional.

Sevilla, Diciembre de 2011.

José Cristóbal Riquelme Santos

Alicia Troncoso Lora

Agradecimientos

La presente memoria es el resultado de un gran esfuerzo, dedicación y aprendizaje y por tanto, está dedicada a todas aquellas personas sin las cuales no hubiera sido posible la realización de la misma.

En primer lugar, debo agradecer a mis directores de Tesis, José Riquelme y Alicia Troncoso, la confianza depositada en mí al otorgarme la oportunidad de realizar la presente tesis doctoral, por su capacidad de trabajo, su apoyo, sus valiosos consejos y su constante dedicación, y por supuesto extiendo este agradecimiento a Francisco Martínez, pues sin todos ellos esta tesis no habría sido posible.

Deseo igualmente agradecer a mis compañeros y amigos del Dpto. de Lenguajes y Sistemas Informáticos de la Universidad de Sevilla por la ayuda prestada, por su cariño y amistad brindada desde que comencé a formar parte de esta gran familia, y sobre todo, por compartir risas y penas, por amenizar y hacer más llevaderos los largos días de trabajo en la escuela.

Muy especialmente me gustaría agradecer a todos los integrantes de mi grupo de investigación y en particular, a Cristina e Isa por poder colaborar con ellas y haber alcanzado algunos logros de esta tesis trabajando junto a ellas.

Quisiera hacer una mención a mis profesores de la Universidad de Jaén, por formarme como ingeniera, hacer que despertara en mi la curiosidad por investigar y ahora darme la oportunidad de poder trabajar con algunos de ellos.

También quisiera agradecer a Jaume Bacardit su invitación para realizar una estancia de investigación en la Universidad de Nottingham, y en general, a los miembros de su grupo de investigación, por su gran acogida, por haberme ayudado en todo lo que he necesitado, y por hacer más fácil mi integración en la vida inglesa.

Finalmente y no por ello menos importante, me gustaría agradecer a mi familia, y sobre todo a mis padres, así como a mi hermana, porque todo lo que he conseguido ha sido gracias a ellos, por su apoyo, cariño y compresión en las dificultades por las que he pasado desde que empecé en este viaje, así como agradezco la cantidad de ánimos y consejos que me han dado estando siempre para ayudar a enfrentarme a cualquier obstáculo. Esta memoria es por y para vosotros.

Y por último, a una de las personas más importantes de mi vida, porque a su lado es todo más sencillo, gracias por ser como eres y estar siempre a mi lado. Tú más que nadie sabes todo el trabajo y dedicación que esta memoria lleva detrás. Quiero que sepas que una parte de esta memoria también es tuya.

Gracias a todos

Tesis Doctoral parcialmente subvencionada por el Ministerio de Educación y Ciencia con el proyecto TIN2007-68084-C-00

MINISTERIO
DE EDUCACIÓN
Y CIENCIA

MEC
TIN2007-68084-C-00

Resumen

La investigación que se propone en el trabajo reflejado en esta memoria de tesis doctoral, se incluye dentro de la disciplina del descubrimiento de conocimiento en bases de datos y en concreto, se centra en la etapa de minería de datos que incluye el mismo. La mayoría de las tareas de la minería de datos están basadas en el aprendizaje inductivo y dentro del mismo, el trabajo presentado pertenece al aprendizaje no supervisado. Concretamente, las reglas de asociación será la técnica utilizada para la obtención de conocimiento.

Esta memoria de investigación se centra entonces en la extracción de reglas de asociación cuantitativas basadas en técnicas de computación evolutiva. En concreto, se plantean cuatro propuestas: un algoritmo evolutivo de codificación real denominado Qarga, otro algoritmo evolutivo de codificación real que extiende el conocido algoritmo Chc de codificación binaria, llamado Qarga-Chc. La tercera propuesta, Eqar, es una versión mejorada de Qarga-Chc y por último, la cuarta propuesta denominada Moqar se basa en una optimización multiobjetivo que sigue el esquema del algoritmo Nsga-II.

Existen numerosos algoritmos y herramientas para la extracción de reglas de asociación que trabajan sobre dominios continuos pero que se limitan a discretizar dichos dominios mediante alguna estrategia concreta para tratarlos posteriormente como si fueran discretos. Sin embargo, las reglas de asociación cuantitativas que se han obtenido en los algoritmos propuestos en esta tesis doctoral trabajan con intervalos sobre los atributos sin necesidad de una discretización previa de los datos. Asimismo permiten un cierto grado de libertad para elegir el tipo de reglas que se van a obtener y que las variables no sean forzadas a pertenecer al antecedente o al consecuente.

Las cuatro propuestas desarrolladas en esta tesis doctoral se han aplicado sobre diferentes tipos de datos y dominios. En concreto se han evaluado sobre datos sintéticos bajo diferentes niveles de ruido y bases de datos públicas en las cuales se ha establecido una comparación con otras técnicas de la literatura. Además, se han descubierto reglas de asociación sobre series temporales climatológicas reales con el objetivo de encontrar todas las relaciones existentes entre la contaminación atmosférica y las condiciones climatológicas. También se han extraído reglas en series de datos de ozono procedente de observaciones de satélites en la península ibérica para intentar modelar las relaciones entre el ozono y variables climáticas en distintas ciudades. Finalmente se presenta una aplicación de nuestra herramienta a datos biológicos, en concreto, microarrays para la determinación de relaciones entre genes y sus niveles de expresión.

Palabras Clave

Minería de datos, algoritmos evolutivos, reglas de asociación cuantitativas.

Abstract

The research proposed in this PhD dissertation is included within the discipline of knowledge discovery in databases and in particular, it is focuses on the stage of Data Mining. Most of the Data Mining tasks are based on inductive learning and the work presented belongs to unsupervised learning. Specifically, the association rules is the technique used to obtain knowledge.

Therefore, this dissertation is focused on the extraction of quantitative association rules based on evolutionary computation techniques. Specifically, four proposal are presented: a real-coded evolutionary algorithm called QARGA, another real-coded evolutionary algorithm that extends the well-known binary coded CHC algorithm, named QARGA-CHC. The third proposal, EQAR, is an improved version of QARGA-CHC, and finally, the fourth proposal called MOQAR is based on a multi-objective optimization that follows the scheme of the NSGA-II algorithm.

There are many algorithms and tools for the extraction of association rules working on continuous domains but they discretize these domains using a specific strategy to deal with them as discrete data. However, the quantitative association rules obtained by the proposed algorithms in this PhD dissertation, work with adaptive intervals on the attributes without a previous data discretization. Furthermore, the proposed algorithms allow several degrees of freedom in specifying the user's preference regarding both of the structure of the rules that will be obtained and attributes are not forced to belong to the antecedent or the consequent.

The four proposals developed in this PhD dissertation have been applied on different data types and domains. In particular, they have been evaluated on synthetic data under different noise levels and public databases which a comparison with other techniques from literature have been carried out. Afterwards, association rules have been discovered over real world climatological time series in order to find all the relationships among air pollution and weather conditions. Furthermore, association rules have been discovered in ozone series from satellite observations over the Iberian Peninsula with the aim to model the relationships among the ozone and the climatological variables in different cities. Finally, an application of our tools to biological data has been proposed, in particular, to microarrays for determining relationships among genes and their expression levels.

Keywords Data Mining, Evolutionary Algorithm, Quantitative Association Rules.

Índice general

Índice de figuras

Índice de tablas

Parte I

Memoria

Capítulo 1

Introducción

1.1. Planteamiento del problema

Es habitual encontrar fenómenos naturales en los que unas variables se correlacionan con otras. De esta forma, procesos del mundo real pueden ser modelados con el conocimiento de variables relacionadas que tienen un efecto en tal proceso. Por ejemplo, la existencia de la lluvia ácida no puede entenderse sin la existencia de otros agentes contaminantes como el monóxido de nitrógeno (NO) o el dióxido de azufre (SO_2). En el caso del pronóstico de la lluvia es necesario el análisis de otras variables como la temperatura, la humedad o la presión atmosférica. En otras palabras, conocer cómo algunas variables pueden afectar a otras puede ser relevante para obtener modelos de comportamiento precisos. Además un análisis diligente de las variables correlacionadas en el caso de datos climatológicos puede conducir al descubrimiento de cómo una variable puede comportarse en relación a otras.

El objetivo del proceso de extracción de las reglas de asociación (a partir de ahora AR, del inglés *Association Rules*) consiste precisamente en descubrir la presencia de conjunciones de atributos pertenecientes a un valor o un intervalo que aparecen en un conjunto de datos con cierta frecuencia.

Una rigurosa revisión de los trabajos publicados recientemente revela que la extracción de AR con atributos numéricos es un tema emergente desde que fue introducido por primera vez en el año 1993. Debido a que se trata de un tema en auge y bastante interesante, nos hemos centrado en la extracción de AR cuantitativas (a partir de ahora QAR, del inglés *Quantitative Association Rules*) en bases de datos con atributos continuos sin un paso previo de discretización, es decir, sin establecer a priori una división del dominio de los datos en intervalos. En la literatura estudiada hemos encontrado un gran número de técnicas y algoritmos cuyo objetivo es la extracción de AR. En el ámbito de las QAR, algunas técnicas trabajan sobre datos reales que discretizan realizando una división previa del dominio de los atributos en intervalos, como ocurre con el conocido algoritmo APRIORI [Agrawal and Srikant, 1994]. Análogamente existe un gran número de algoritmos donde se establecen umbrales mínimos de soporte y confianza. Otras técnicas dividen el proceso de extracción de reglas en dos fases: primero obtienen el conjunto de elementos con valores frecuentes en los datos, y segundo, construyen las reglas a partir de estos conjuntos frecuentes.

Tras el estudio realizado, se ha visto una carencia considerable de algoritmos que trabajen sobre QAR, sin discretizar los datos, sin dividir el proceso en dos fases y sin establecer umbrales mínimos por parte del usuario. Por tanto, la principal motivación de esta tesis doctoral ha sido el desarrollo de algoritmos evolutivos capaces de encontrar QAR en bases de datos con atributos continuos, evitando la discretización como paso previo del proceso, a diferencia de muchos otros enfoques que discretizan los datos para descubrir reglas [Agrawal et al., 1993, Aumann and Lindell, 2003, Vannucci and Colla, 2004]. Adicionalmente se han realizado experimentos para comprobar la capacidad de los algoritmos propuestos en la extracción de QAR con el fin de mejorar las técnicas existentes.

Los algoritmos que se proponen en esta tesis doctoral proporcionan QAR estableciendo relaciones entre todos los atributos de los conjuntos de datos. Cabe destacar que dichos algoritmos permiten varios grados de libertad en la especificación de las preferencias del usuario en cuanto al número total de atributos, así como atributos en el antecedente y en el consecuente, amplitud de los intervalos o cobertura de los datos.
Asimismo, la búsqueda de ARs no debe confundirse con el descubrimiento de subgrupos (a partir de ahora SD, del inglés *Subgroup Discovery*) [del Jesús et al., 2007]. Las AR son una herramienta de aprendizaje no supervisado, mientras que SD lleva a cabo un aprendizaje supervisado. Tanto AR como SD buscan reglas, pero SD realiza una búsqueda de condiciones de un sólo atributo y AR puede tratar con varios atributos en el antecedente y en el consecuente. Por otra parte, las AR no preestablecen el rango en el que los atributos del consecuente pueden variar.

1.2. Objetivos

La presente memoria de tesis doctoral está organizada en torno a una serie de objetivos que involucra el análisis de las técnicas existentes para extracción de AR, el análisis de las distintas métricas que existen en la literatura para la evaluación de la calidad de las mismas, el desarrollo de algoritmos que sean capaces de llevar a cabo el descubrimiento de AR y la aplicación de los mismos a dominios de diferentes problemas.

Los principales objetivos de esta tesis doctoral se centran en:

- Estudio del entorno del problema: Concretamente, revisión de las técnicas existentes relacionadas con la extracción de AR, determinación de problemas e inconvenientes de las mismas, así como aspectos novedosos y diferenciadores con respecto a los algoritmos ya existentes. Asimismo, realización de un estudio de las diferentes medidas de calidad y recopilación de ejemplos para su posterior evaluación.

- Diseño y desarrollo de algoritmos evolutivos capaces de encontrar AR en bases de datos con atributos continuos. Análisis previo de qué parámetros son necesarios para los mismos y definición de la función de evaluación a optimizar.

- Evaluación de los algoritmos desarrollados mediante la realización de una serie de experimentos que verifiquen y validen el funcionamiento de las propuestas. Para ello, se pretende la aplicación de los mismos a bases de datos sintéticas y datos del mundo real. Especialmente, los algoritmos se aplicarán a variables climatológicas o atmosféricas, así como a datos biológicos procedentes de microarrays con el objetivo de obtener un análisis de agentes contaminantes debido a su relación con el cambio climático y establecer relaciones entre genes.

- Análisis de los resultados obtenidos por los algoritmos y establecer una comparativa entre éstos y los resultantes de otras técnicas existentes en la literatura.

1.3. Estructura de la memoria de la tesis doctoral

Esta memoria de tesis doctoral está dividida en tres partes como se detalla a continuación:

- La **Parte I** se divide en tres capítulos:
  - El **Capítulo 1** está dedicado a la descripción del problema, el contexto de investigación y los objetivos de la tesis doctoral. En la *Sección 1.1* se plantea el problema que hemos abordado y la motivación para el desarrollo de la tesis doctoral y en la *Sección 1.2* se detallan los objetivos a perseguir.

  - El **Capítulo 2** incluye el contexto de investigación en el que se encuadra la tesis doctoral. Las *Secciones 2.1*, *2.2*, *2.3* y *2.4*, respectivamente, describen brevemente el concepto de minería de datos, la computación evolutiva y los problemas entre diversidad y convergencia que presentan, el algoritmo CHC y por último algunos conceptos sobre optimización multiobjetivo. A continuación, la *Sección 2.5* introduce una sección dedicada a la descripción de las AR, incluyendo las medidas de calidad utilizadas y algunos algoritmos existentes en la literatura para su descubrimiento. Por último, la *Sección 2.6* presenta un resumen de las aportaciones que se han realizado durante el proceso de investigación de esta tesis doctoral.

  - El **Capítulo 3** resume las distintas propuestas que se incluyen dentro de las selección de trabajos, así como la discusión conjunta de los resultados de los mismos.
- La **Parte II** contiene el compendio de artículos que comprende la selección de publicaciones que se han desarrollado durante el proceso de investigación llevado a cabo para la realización de esta tesis doctoral. La selección de publicaciones se ha distribuido en dos capítulos que se detallan a continuación: el **Capítulo 4** presenta la selección de artículos en revistas que se incluyen en el ranking JCR de Thomsom-Reuters y el **Capítulo 5** contiene otros trabajos de investigación relevantes que han sido publicados en revistas no indexadas o actas de congresos que se encuadran dentro de la temática de esta tesis doctoral.

- Por último, la **Parte III** concluye la memoria de tesis doctoral, resumiendo las conclusiones obtenidas durante esta investigación, las cuales nos impulsan a considerar ciertas alternativas como trabajos futuros para la mejora de las propuestas desarrolladas.

Capítulo 2

Contexto de investigación

2.1. Minería de Datos

El dominio del trabajo realizado en la tesis doctoral, se encuadra dentro de la disciplina del descubrimiento de conocimiento en bases de datos (a partir de ahora KDD, del inglés *Knowledge Discovery in Databases*) y concretamente en la fase de minería de datos (a partir de ahora DM, del inglés *Data Mining)* [Dunham, 2003, Orallo et al., 2004, Han and Kamber, 2006, Hand, 2007]. La DM es una de las áreas que más éxito ha tenido a la hora de analizar información con el objetivo de extraer nuevo conocimiento. La DM está relacionada con una multitud de disciplinas entre las cuales se encuentra el aprendizaje automático que es el área de la *inteligencia artificial* que se ocupa de desarrollar algoritmos capaces de aprender. Entre otras sofisticadas técnicas, se aplica la inteligencia artificial para encontrar patrones y relaciones dentro de los datos permitiendo la creación de modelos, es decir, representaciones abstractas de la realidad, pero es el KDD el encargado de llevar a cabo la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan un significado a estos patrones encontrados.

El proceso de aprendizaje más utilizado por las técnicas de DM es el denominado *aprendizaje inductivo*. Un tipo de aprendizaje inductivo es el **aprendizaje descriptivo**, también conocido como *aprendizaje no supervisado*, cuyo objetivo es identificar patrones que explican o resumen los datos, es decir, sirven para explorar las propiedades de los datos examinados. Con el aprendizaje descriptivo se pueden establecer relaciones entre variables (atributos) o bien entre individuos (ejemplos), es decir *asociaciones o clustering*, respectivamente.

El concepto de **asociación** es bastante amplio e incluye diferentes formas de representar e interpretar el conocimiento. Las AR han sido una de las representaciones más investigadas y relevantes entre todas las posibles dentro de la DM. Las AR identifican relaciones no explícitas entre atributos de forma que una asociación entre dos atributos ocurre cuando la frecuencia de que dos valores determinados de cada atributo conjuntamente sea relativamente alta. La técnica de AR es muy común en los análisis de cestas de mercados. En la Sección 2.5 las AR serán detalladas con más profundidad.

2.2. Computación Evolutiva

Existen distintas técnicas de *inteligencia artificial* para optimizar el aprendizaje de AR entre las cuales podemos destacar la **computación evolutiva** [Fogel, 1995, Fogel, 1998, Eiben and Smith, 2008]. La computación evolutiva es un subconjunto dentro de las metaheurísticas bioinspiradas y es una de las principales técnicas que participan en lo que se conoce como *Soft Computing* (Figura 2.1). La *Soft Computing* es una rama de la inteligencia artificial centrada en el diseño de sistemas inteligentes capaces de resolver problemas reales imitando la forma en que lo hace el ser humano.

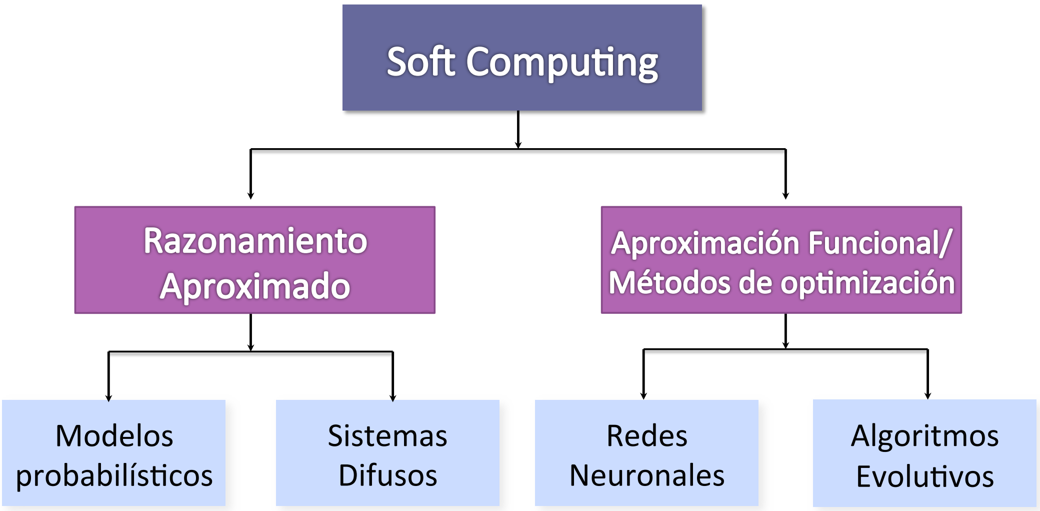


Figura 2.1: Taxonomía de la computación inteligente.

La computación evolutiva constituye una clase de métodos de búsqueda y optimización que imitan los principios de la evolución natural [Goldberg, 1989] basándose en poblaciones cuyos individuos representan soluciones a nuestro problema. Cuando tiene lugar la simulación de estos modelos en un ordenador, se obtiene una técnica de optimización probabilística que en muchos casos mejora a otras heurísticas y/o algoritmos en la resolución de problemas difíciles.

Un **algoritmo evolutivo** [Back, 1996, Back et al., 1997] (a partir de ahora EA, del inglés *Evolutionary Algorithm)* puede ser utilizado para determinar los conjuntos frecuentes y seguidamente los intervalos de los atributos que forman una regla. Por tanto, en esta tesis doctoral se propone optimizar tanto los atributos que componen una regla, como los intervalos de los mismos a partir de técnicas de *computación evolutiva* con nuevos métodos frente a los existentes. En la tesis doctoral además se profundiza sobre este tema describiendo las distintas metodologías y los operadores genéticos desarrollados.

El funcionamiento básico de cualquier EA se basa en mantener una población de posibles soluciones al problema, realizar modificaciones sobre esa población de soluciones y seleccionar un número de individuos según su nivel de adaptación o fitness que se mantendrán en generaciones futuras. La población de individuos irá evolucionando a lo largo de las distintas generaciones con un único objetivo: ir explorando mejores regiones del espacio de búsqueda mediante esas modificaciones y selección para así conseguir llegar a una solución más cercana a la óptima.

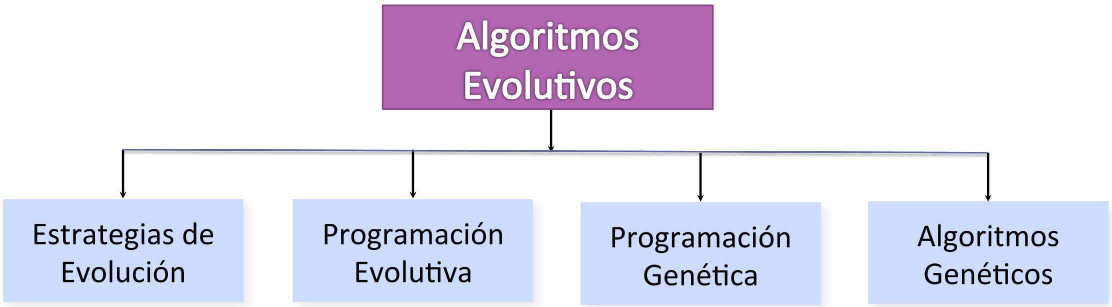


Figura 2.2: Paradigmas básicos de algoritmos evolutivos.

Existen cuatro paradigmas básicos o tipos principales de EA dentro de la *computación evolutiva*, como muestra la Figura 2.2, que comparten las características que se detallan a continuación. Utilizan un aprendizaje colectivo para toda la población, los descendientes se generan mediante procesos no determinísticos o aleatorios, tratando de modelar los procesos de mutación y cruce que se dan en la evolución natural de cualquier especie (la mutación será la autorreplicación errónea de los individuos y el cruce el intercambio de material genético entre individuos). Otra característica común será la medida de calidad o fitness de cada individuo. De todas las técnicas existentes nos centraremos en algoritmos genéticos (a partir de ahora GA, del inglés *Genetic Algorithm*) que utilizan operadores genéticos sobre cromosomas.

Los GA son algoritmos de optimización, búsqueda y aprendizaje inspirados en los procesos de evolución natural y genética. Fueron desarrollados en los EE.UU. en los años 70s por J. Holland y posteriormente por K. DeJong [Jong, 1975], y D. Golberg [Goldberg, 1989]. Se consideran algoritmos muy populares y potentes en muchos campos y en diversas aplicaciones y además proporcionan altas prestaciones a bajo costo. Se aplican tanto a optimización discreta como continua, aunque originalmente fueron desarrollados para la primera [Back et al., 1997, Michalewicz, 1992, Eiben and Smith, 2003, Davis, 1991].

Estos algoritmos se implementan como una simulación por ordenador en el que una población de representaciones abstractas (los cromosomas) de soluciones candidatas (individuos) para un problema de optimización que evolucionan hacia mejores soluciones. En esta tesis doctoral, se utilizan GA de codificación real debido a que el dominio de las AR es continuo y por tanto, los algoritmos propuestos tratan con datos numéricos durante el proceso de extracción de reglas.

Para construir un GA, tal y como muestra la Figura 2.3, es necesario diseñar una representación, decidir cómo inicializar la población, diseñar una correspondencia entre genotipo y fenotipo, diseñar una forma de evaluar un individuo, un operador de mutación y cruce adecuado, establecer cómo seleccionar los individuos para ser padres, cómo reemplazar a los individuos, y la condición de parada. Las propiedades de los GA, en general, hacen de ellos una técnica adecuada para las AR. Las aplicaciones de los algoritmos evolutivos sobre la extracción de AR en bases de datos básicamente se basan en la búsqueda de los conjuntos frecuentes dentro de los datos, así como la optimización de las diferentes variables que evalúan las AR.

Existen varias alternativas del esquema de representación ya que los GA siguen dos enfoques respecto a la forma de codificar reglas dentro de una población de individuos:

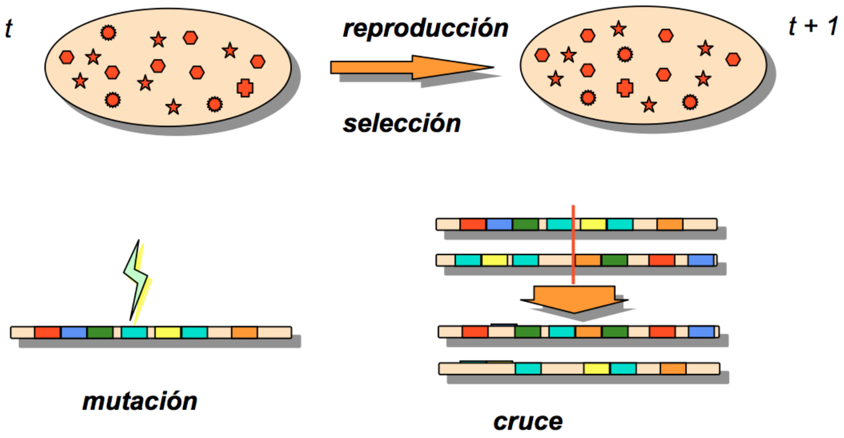


Figura 2.3: Elementos que constituyen un algoritmo genético.

- *Enfoque "Cromosoma = Regla"*, en el que cada individuo codifica una única regla.

- *Enfoque "Cromosoma = Base de Reglas"*, también denominado enfoque **Pittsburgh** [Jong et al., 1993, Janikow, 1993], en el que cada individuo representa un conjunto de reglas.

Dentro del enfoque *"Cromosoma = Regla"* existen dos propuestas genéricas:

- El enfoque **Michigan** [Greene and Smith, 1993, Giordana and Neri, 1995] en el que cada individuo codifica una única regla pero la solución final será la población final o un subconjunto de la misma. En este caso, es necesario evaluar el comportamiento del conjunto de reglas al completo y la aportación de la regla individual al mismo.

- El enfoque **IRL** (del inglés *Iterative Rule Learning*) [Venturini, 1993], en el que cada cromosoma representa una regla, pero la solución del GA es el mejor individuo y la solución global está formada por los mejores individuos de una serie de ejecuciones sucesivas.

La elección del esquema de representación depende, entre otros aspectos, de la tarea a realizar por parte del algoritmo de DM y, por tanto, del tipo de regla a descubrir.

2.3. Equilibrio entre diversidad y convergencia

Existen dos factores contrapuestos que influyen sobre la efectividad de un GA como puede verse en la Figura 2.4. Por un lado, la **convergencia**, centra la búsqueda en regiones prometedoras mediante **presión selectiva**, la cual permite que los mejores individuos sean seleccionados para reproducirse y resulta necesaria para que el proceso de búsqueda no sea aleatorio. Por otro lado, la **diversidad** evita la convergencia prematura, esto es, la rápida convergencia hacia zonas que no contienen el óptimo global y está asociada a las diferencias entre los cromosomas que componen la población. La falta de diversidad genética hace que todos los individuos en la población sean parecidos.

Para solucionar los problemas de poca diversidad y convergencia prematura, ya que en la práctica es irreversible, podrían considerarse dos importantes opciones de mejora como son la inclusión de un mecanismo de diversidad en la evolución y la reinicialización cuando se produce convergencia prematura.

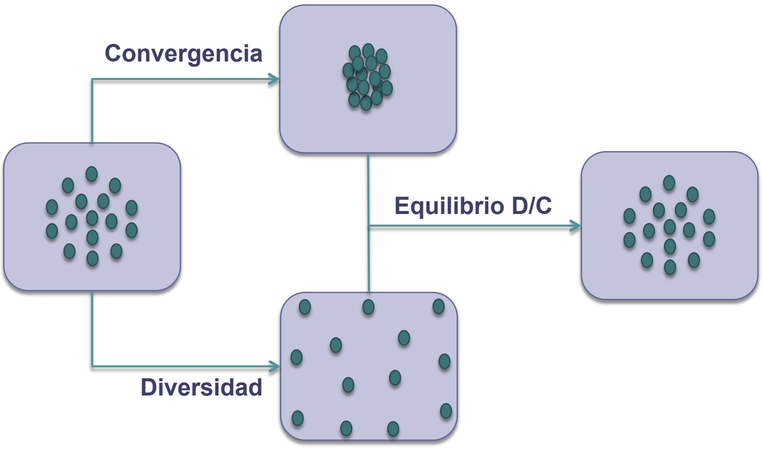


Figura 2.4: Equilibrio entre diversidad y convergencia

Ante los problemas relacionados con el desequilibrio entre *diversidad y convergencia* que influyen negativamente sobre la efectividad de los algoritmos genéticos, Eshelman [Eshelman, 1991] propuso una variante al GA tradicional llamado CHC (del inglés *Cross generational elitist selection, Heterogeneous recombination and Cataclilsm mutation*), que obtiene un gran equilibrio entre diversidad y convergencia.

La idea general de este algoritmo es que utiliza una estrategia de selección elitista, incorpora un mecanismo de restricción al cruzamiento que impide la combinación de individuos muy similares entre si, y además, no utiliza mutación, pues la diversidad se mantiene mediante un mecanismo de reinicialización.

Concretamente las características fundamentales para dotar de diversidad al algoritmo CHC son la introducción de un nuevo cruce llamado *Half Uniform* (a partir de ahora HUX), un mecanismo de prevención de incesto y la reinicialización de la población cuando se dan ciertas condiciones. El CHC original presenta una estrategia elitista para la selección de la población que constituirá la próxima generación produciendo una convergencia elevada. Asimismo, incluye una fuerte diversidad en el proceso evolutivo a través de mecanismos de prevención de incesto y el operador específico de cruce uniforme HUX. Por otra parte, la población se reinicializa cuando su diversidad es pobre. Estas cuatro componentes novedosas se resumen en la Figura 2.5.

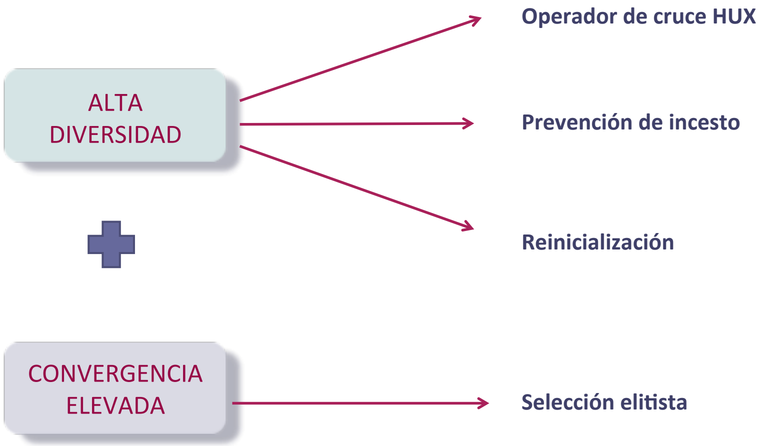


Figura 2.5: Componentes del CHC

Los detalles de las principales características del algoritmo CHC se resumen en los siguientes puntos:

- **Selección elitista**: Este tipo de estrategia garantiza la supervivencia de los mejores individuos. La población actual y sus descendientes se unen y se eligen los mejores individuos (de acuerdo a la función fitness) para componer la población de la siguiente generación.

- **Operador de cruce HUX**: Este operador intercambia exactamente la mitad de los genes de los padres que no coinciden. Este cruce se considera muy destructivo e introduce una cierta diversidad en la población evitándose así la prevención de la convergencia prematura.

- **Prevención de incesto**: En el algoritmo CHC el cruce entre los hermanos está prohibido. Por lo tanto, para evitar esto, se aplica la siguiente regla: Dos individuos sólo se cruzan si su distancia de *Hamming* dividida por dos es mayor que un cierto umbral que se fija en función de la longitud del individuo, como por ejemplo, el número de bits dividido por cuatro. Por consiguiente, sólo padres muy distintos se cruzan. Cuando no hay padres para cruzar debido a que la distancia entre ellos es muy baja, el umbral se decrementa en una unidad. La idea clave es evitar la aplicación del operador de cruce entre individuos similares.

- **Reinicialización**: Cuando el proceso evolutivo converge debido a que los individuos son muy similares, si el umbral definido para la prevención de incesto es negativo la población se reinicializa para proporcionar diversidad a la misma. Generalmente la población se reinicializa con el mejor individuo de la población y variaciones del mejor individuo las cuales consisten en la variación de un porcentaje de los genes con cierta probabilidad.

El algoritmo funciona como sigue. En cada generación, se forman aleatoriamente parejas de todos los individuos de la población. Se analiza si hay diferencias significativas entre los dos individuos de cada pareja (prevención de incesto) y, de ser así, se aplica el operador de cruce para generar dos nuevos descendientes. La población de la siguiente generación se construye de forma elitista, escogiendo las mejores soluciones entre la antigua población y los descendientes generados, manteniendo siempre el mismo tamaño de la población. En caso de que esta nueva población sea igual a la población anterior, se reduce en uno el umbral d definido para la prevención de incesto. Cuando este umbral llega a cero, la población se reinicializa construyendo una nueva población e inicializando dicho umbral.

En general presenta un mejor patrón de explotación que el GA tradicional al impedir que se crucen individuos muy similares. Si el operador de reinicialización es eficaz, las poblaciones resultantes presentan una mayor diversidad que las del GA tradicional. Al proporcionar mayor diversidad genética, es capaz de obtener mejores resultados que el GA tradicional para la resolución de problemas "difíciles" como por ejemplo, aquellos problemas que tengan espacio de búsqueda muy disperso, etc.

2.4. Optimización multiobjectivo

Los EA fueron originalmente diseñados para solucionar problemas de optimización de un sólo objetivo. Sin embargo, muchos problemas de optimización del mundo real se caracterizan por la existencia de múltiples medidas u objetivos que entran unos en conflicto con otros y que deberían optimizarse, o al menos ser satisfechos simultaneamente.

Un problema multiobjetivo consiste en:

$$Maximizar\, o\, Minimizar \;\; z = f(x) = (f_1(x), f_2(x), ..., f_n(x)) \tag{2.1}$$

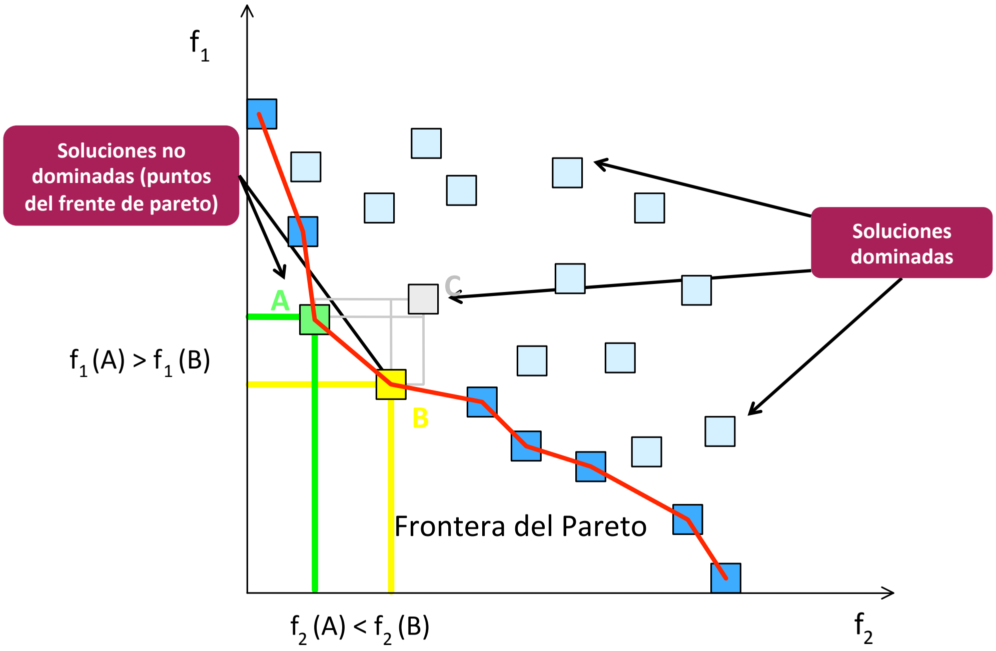


Figura 2.6: Ejemplo de conjunto de soluciones no-dominadas que forma la frontera del pareto en los algoritmos multiobjetivo para un problema de mínimos.

Debido a que la optimización multiobjetivo busca un vector óptimo y no un único valor, no se puede considerar en muchas ocasiones que una solución es mejor que otra, por tanto puede que no exista una única solución óptima sino un conjunto de soluciones optimas, conocido como conjunto óptimo de Pareto [Zitzler and Thiele, 1999]. En la Figura 2.6 puede verse un ejemplo de un conjunto de soluciones no dominadas que forman la frontera del pareto.

Se considera que una solución a domina a otra b, si cada objetivo de a es mejor o igual que el correspondiente objetivo de b y en al menos un objetivo es mejor.

La presencia de múltiples objetivos en conflicto y la necesidad de tener que tomar decisiones provoca que surjan en la práctica un número de escenarios de problemas diferentes. En las últimas dos décadas se ha incrementado el interés en el uso de GA para optimización multiobjetivo.

Existen múltiples propuestas de GA multiobjetivo [Deb, 2001] que se pueden agrupar entorno a tres enfoques:

- Modelos evolutivos que utilizan **pesos** para la combinación de objetivos en una función escalar. La agregación de los objetivos conduce a la obtención de un único punto de equilibrio en la frontera de pareto. Tienen como inconveniente la posible descompensación entre objetivos, el conocimiento profundo requerido sobre el problema y que, en general, no proporcionan una familia de soluciones. VOW-GA (*Variable Objective Weighting Genetic Algorithm*) [Hajela and Lin, 1992] y RW-GA (*Random Weighted Genetic Algorithm*) [Ishibuchi and Murata, 1998] son ejemplos de algoritmos de este tipo.

- Métodos evolutivos basados en **población**, en los que la búsqueda se realiza en diferentes direcciones para generar poblaciones de soluciones no dominadas. Dentro de este enfoque se encuadran, entre otros, los algoritmos MOGA (*Multi-Objective Genetic Algorithm*) [Fonseca and Fleming, 1993], NPGA (*Niched Pareto Genetic Algorithm*) [Horn and Nafpliotis, 1993] y los algoritmos NSGA (*Non Dominated Sorting Genetic Algorithm*) [Srinivas and Deb, 1994] y NSGA-II [Deb et al., 2002].

- Métodos que utilizan **elitismo**, manteniendo una población elite con soluciones no dominadas que intervienen de distinta forma en la evolución. Dentro de este enfoque se incluyen los algoritmos SPEA (*Strength Pareto Evolutionary Algorithm*) [Zitzler and Thiele, 1999], SPEA-II [Zitzler et al., 2001] y μ-λ MEA (*Multiobjective Evolutionary Algorithm*) [Sarker et al., 2002].

2.5. Reglas de Asociación

En el campo de la DM, las AR se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos. A diferencia de la *clasificación*, las AR son *descriptivas* y por tanto, se trata de una técnica de aprendizaje no supervisado. Las tareas de minería descriptivas caracterizan las propiedades generales de los datos [Han and Kamber, 2006].

La extracción de AR encuentra asociaciones interesantes y/o relaciones de correlación entre elementos de conjuntos de datos de gran volumen. Las AR muestran condiciones de valores en los atributos que ocurren juntos con frecuencia en un conjunto de datos. Los primeros estudios sobre las AR se centraron en datos con atributos binarios, principalmente datos organizados en forma transaccional.

Un ejemplo típico y ampliamente utilizado en la extracción de AR es el *análisis de la cesta de la compra* [Agrawal et al., 1993] sobre el cual se han desarrollado una gran variedad de métodos. A partir de los datos procedentes de los artículos comprados por un cliente, resultaba interesante saber si ciertos grupos de artículos se compran juntos constantemente, ya que se podrían utilizar para ajustar el diseño de la tienda, la venta cruzada, promociones, diseño de catálogos e identificar segmentos de clientes basados en los patrones de compra.

Algunos términos básicos en la terminología de las AR son:

- **ítem** corresponde a pares *atributo-valor.*
- **instancia** es un conjunto de ítems, corresponde a un ejemplo o registro de la base de datos.

El objetivo del proceso de extracción de las AR consiste precisamente en descubrir la presencia de pares de conjunciones (atributo - valor) que aparecen en un conjunto de datos con cierta frecuencia para así formular las reglas que describen las relaciones existentes entre los atributos.

Las AR fueron definidas por primera vez en [Agrawal et al., 1993] para atributos discretos como sigue:

Sea $I = \{i_1, i_2, ..., i_n\}$ un conjunto de n ítems y $D = \{t_1, t_2, ..., t_N\}$ un conjunto de N transacciones, donde cada t_j contiene un subconjunto de ítems. Esto es, una regla puede ser definida como $X \Rightarrow Y$, donde $X, Y \subseteq I$ y $X \cap Y = \emptyset$. Finalmente, X y Y se llaman antecedente (o parte izquierda de la regla) y consecuente (o parte derecha de la regla), respectivamente.

En [Agrawal and Srikant, 1994], el problema de descubrir las AR se divide en dos subtareas:

- Encontrar todos los conjuntos que superan el valor mínimo de soporte. Dichos conjuntos se denominan *conjuntos frecuentes.*
- Partiendo de los conjuntos frecuentes encontrados, generar las reglas que superan el valor mínimo de confianza.

La generación de los conjuntos frecuentes es la tarea más importante y la que requiere más tiempo de computación, de ahí que sea la más estudiada entre la comunidad científica. A diferencia de la estrategia que siguen la mayoría de los investigadores en sus herramientas de extracción de AR, los algoritmos que se proponen en esta tesis doctoral obtienen AR sin realizar el proceso de construcción de los conjuntos frecuentes como paso previo.

Las dos medidas más comúnmente utilizadas para medir las relaciones entre antecedente y consecuente son el soporte y la confianza, sin embargo, como veremos en la Sección 2.5.1 no son suficientes para medir la calidad de las AR obtenidas.

Un algoritmo de AR tratará de descubrir todas las reglas que excedan las cotas mínimas especificadas para el soporte y la confianza. La búsqueda exhaustiva de AR consideraría simplemente todas las combinaciones posibles de atributos, poniéndolos como antecedentes y consecuentes, entonces se evaluaría el soporte y la confianza de cada regla, y se descartarían todas las asociaciones que no satisfacen las restricciones. Sin embargo, el número de combinaciones crece rápidamente con el número de atributos, por lo que si hay 1000 atributos, se tendrán 2^{1000} combinaciones (aproximadamente 10^{300}). Así, este procedimiento para la búsqueda de AR es muy costoso computacionalmente, por lo que se necesitan otros procedimientos más eficientes.

Tradicionalmente, se han extraído AR sobre bases de datos nominales o de datos discretos, sin embargo, cuando el dominio es continuo, las AR se conocen como **Reglas de Asociación Cuantitativas** (a partir de ahora QAR, del inglés *Quantitative Association Rules*). En este contexto el conjunto I de la definición de Agrawal es $\Re^n$ con n el número de atributos y D es la tabla de valores de atributos sobre los que queremos obtener las AR.

Formalmente, sea $F = \{F_1, ..., F_n\}$ un conjunto de atributos, con valores en $\mathbb{R}$. Sea A y C dos subconjuntos disjuntos de F, esto es, $A \subset F$, $C \subset F$, y $A \cap C = \emptyset$. Una QAR es una regla $X \Rightarrow Y$, en la que los atributos de A pertenecen al antecedente X, y los atributos de C pertenecen al consecuente Y, tales que:

$$X = \bigwedge_{F_i \in A} F_i \in [l_i, u_i] \tag{2.2}$$

$$Y = \bigwedge_{F_j \in C} F_j \in [l_j, u_j] \tag{2.3}$$

donde l_i y l_j representan el límite inferior de los intervalos para F_i y F_j respectivamente, y u_i y u_j el límite superior.

Por ejemplo, una QAR podría ser numéricamente expresada como:

$$F_1 \in [12, 25] \wedge F_3 \in [5, 9] \Rightarrow F_2 \in [3, 7] \wedge F_5 \in [2, 8] \tag{2.4}$$

donde F_1 y F_3 constituyen los atributos que aparecen en el antecedente y F_2 y F_5 los que aparecen en el consecuente.

Sin embargo, la mayoría de las aplicaciones a datos con atributos numéricos suelen ser abordados mediante una discretización previa [Yang et al., 2010] de estos atributos numéricos. La discretización está dentro de la tarea de transformación de los datos de la etapa de preprocesamiento [Pyle, 1999] del proceso KDD y consiste en dividir el rango del atributo en k intervalos donde cada valor es reemplazado por el intervalo al que pertenece. No obstante, aunque la discretización simplifica el conjunto de datos sobre el cual se aplica, presenta una serie de inconvenientes [del Jesus et al., 2011] como son la perdida de información, la amplitud fija del intervalo y el tamaño de los intervalos que implica un soporte muy alto y la consecuente pérdida de información en las reglas obtenidas cuando hay pocos intervalos y por tanto intervalos de gran amplitud. Por otro lado, si hay un gran número de intervalos y por consiguiente intervalos estrechos, las reglas obtenidas pueden tener un soporte muy bajo y no superar los umbrales mínimos en algunos casos. También pueden existir ciertos problemas con los límites de los intervalos definidos mediante la discretización, ya que valores muy cercanos a los límites (a la izquierda y a la derecha) tienen que pertenecer a la fuerza a más de una regla. Del mismo modo, pequeños cambios realizados en valores muy cercanos al límite de un intervalo, pueden provocar cambios injustificados en el conjunto de reglas activas.

Como se comentó al inicio de esta memoria, las propuestas desarrolladas tratarán los atributos continuos sin necesidad de discretizar los datos previamente.

La revisión del estado del arte que hemos realizado revela que existe un amplio rango de aplicaciones para la extracción de AR. En concreto, se han utilizado AR para extraer asociaciones en series temporales con fines predictivos. Las AR también han sido aplicadas extensamente en el área de la bioinformática, por ejemplo, para analizar datos procedentes de microarrays, proteínas, etc. Pueden ser de utilidad en la gestión y toma de decisiones dentro del proceso de desarrollo de software, en el área de *Web Mining*, Minería de Texto, etc.

2.5.1. Medidas de calidad de las Reglas de Asociación

Los siguientes párrafos detallan las medidas de calidad más utilizadas para evaluar una AR. Hay que tener en cuenta que es muy importante medir la calidad de una regla para evaluar los resultados y seleccionar el mejor conjunto de reglas para los algoritmos propuestos. En el proceso de extracción de AR, se han seleccionado medidas basadas en la probabilidad de que los valores pertenezcan al antecedente y/o al consecuente que evalúan la generalidad y fiabilidad de las AR. En particular, la medida del *soporte* se utiliza para representar la generalidad de la regla y las medidas *confianza*, *lift* y *leverage* se usan normalmente para representar la fiabilidad de una regla [Berzal et al., 2001, Geng and Hamilton, 2006]. Las medidas de interés se definen formalmente para el caso de las QAR como sigue:

- **Soporte(X)** [Geng and Hamilton, 2006]: El *soporte* de un conjunto X de atributos pertenecientes a un intervalo se define como el porcentaje de instancias en el conjunto de datos que satisfacen X. En general, el *soporte* de X es conocido como la probabilidad de X.

  $$sup(X) = P(X) = \frac{n(X)}{N}. \tag{2.5}$$

  donde $n(X)$ es el número de ocurrencias de X en el conjunto de datos, y N es el número de instancias que forman tal conjunto de datos.

- **Soporte**($X \Longrightarrow Y$)[Geng and Hamilton, 2006]: El *soporte* de la regla $X \Longrightarrow Y$ es el porcentaje de instancias del conjunto de datos que satisfacen X e Y simultáneamente.

$$sup(X \Longrightarrow Y) = P(Y \cap X) = \frac{n(XY)}{N}. \qquad (2.6)$$

donde $n(XY)$ es el número de instancias que satisfacen las condiciones para el antecedente X y el consecuente Y al mismo tiempo.

- **Confianza**$(X \Longrightarrow Y)$[Geng and Hamilton, 2006]: La *confianza* es la probabilidad de que las instancias que satisfacen X, también satisfacen Y. En otras palabras, es el *soporte* de la regla dividido entre el *soporte* del antecedente.

$$conf(X \Longrightarrow Y) = P(X \mid Y) = \frac{sup(X \Longrightarrow Y)}{sup(X)} \qquad (2.7)$$

- **Lift**$(X \Longrightarrow Y)$[Brin et al., 1997a]: El *lift* o interés está definido como cuántas veces más X e Y aparecen juntos en el conjunto de datos de lo esperado, asumiendo que la presencia de X e Y son hechos estadísticamente independientes. Mide el grado de dependencia entre el antecedente X y el consecuente Y. Un valor mayor que 1 indica dependencia estadística en ocurrencias simultáneas de X e Y, en otras palabras, la regla proporciona información exitosa relativa a que X e Y aparezcan juntos en el conjunto de datos. Un valor inferior a 1 indica independencia estadística.

$$lift(X \Longrightarrow Y) = \frac{sup(X \Longrightarrow Y)}{sup(X)sup(Y)} = \frac{conf(X \Longrightarrow Y)}{sup(Y)} \qquad (2.8)$$

- **Leverage**$(X \Longrightarrow Y)$[Piatetsky-Shapiro, 1991]: La medida *leverage* mide la proporción de casos adicionales cubiertos tanto por X e Y sobre los esperados si X e Y fueran independientes. El *leverage* toma valores en el intervalo [-1, 1]. Valores iguales o inferiores a 0, indican una fuerte independencia entre antecedente y consecuente. Valores mayores que 0 son deseables. Por otro lado, valores cercanos a 1 son esperados para AR de gran calidad. Además, el *leverage* es una cota inferior para el *soporte*, y por tanto, optimizando sólo el *leverage* se garantiza un cierto mínimo de *soporte* a diferencia de optimizar sólo la *confianza* o el *lift*.

$$lev(X \Longrightarrow Y) = sup(X \Longrightarrow Y) - sup(X)sup(Y) \qquad (2.9)$$

- **Accuracy**$(X \implies Y)$[Geng and Hamilton, 2006]: La medida *accuracy* mide el grado de veracidad o coincidencia de los datos obtenidos sobre los datos reales. Un valor de *accuracy* igual a 100 % significa que los valores medidos son exactamente los mismos que los valores dados. En el campo de la extracción de AR, la *accuracy* mide el porcentaje de instancias del conjunto de datos que cumplen el antecedente y el consecuente más el porcentaje de instancias que no satisfacen ni el antecedente ni el consecuente. La *accuracy* puede tomar valores en el intervalo [0, 1] y se esperan valores cercanos a 1 para una regla de alta calidad y veracidad.

$$Acc(X \implies Y) = sup(X \implies Y) + sup(\neg X \implies \neg Y) \tag{2.10}$$

  donde $\neg$ significa negación, por tanto, $sup(\neg X \implies \neg Y)$ es el porcentaje de instancias en el conjunto de datos que no satisfacen ni X ni Y simultaneamente.

- **Conviction**$(X \implies Y)$[Brin et al., 1997b]: Es una medida que tiene en cuenta tanto el *soporte* del antecedente como el *soporte* del consecuente de la regla. Fue introducido como una alternativa a la *confianza*. Valores en el intervalo [0,1) implican dependencia negativa, valores superiores a 1 significa dependencia positiva, y un valor igual a 1 expresa independencia.

  La *conviction* es direccional y consigue su máximo valor (infinito) cuando la implicación es perfecta, es decir, si cada vez que X ocurre, también ocurre Y. Sin embargo, tiene el problema de no estar acotado al tener como máximo el valor infinito.

$$conv(X \implies Y) = \frac{sup(X)sup(\neg Y)}{sup(X \implies \neg Y)} = \frac{1 - sup(X)}{1 - conf(X \implies Y)} \tag{2.11}$$

- **Gain**$(X \implies Y)$[Brin et al., 1997b]: La medida *gain* se conoce como valor añadido o cambio del *soporte*. Se calcula a partir de la diferencia entre la *confianza* de la regla y el *soporte* del consecuente.

$$Gain(X \implies Y) = conf(X \implies Y) - sup(Y) \tag{2.12}$$

- **Certainty Factor**$(X \Longrightarrow Y)$ [Shortliffe and Buchanan, 1975]: La medida *certainty factor* fue introducido por Shortliffe y Buchanan para representar la incertidumbre de un sistema experto llamado MYCIN. Se interpreta como una medida de la variación de la probabilidad de que Y está en una instancia cuando consideramos sólo aquellas instancias donde X aparece. Una interpretación similar puede aplicarse para un *certainty factor* negativo. Es el valor de la *gain* normalizado en el intervalo [-1,1] y alcanza su máximo valor si y sólo si la regla es totalmente precisa y cierta.

  Si $conf(X \Longrightarrow Y) > sup(Y)$:

$$CF(X \Longrightarrow Y) = \frac{conf(X \Longrightarrow Y) - sup(Y)}{1 - sup(Y)} \tag{2.13}$$

  Si $conf(X \Longrightarrow Y) \leq sup(Y)$:

$$CF(X \Longrightarrow Y) = \frac{conf(X \Longrightarrow Y) - sup(Y)}{sup(Y)} \tag{2.14}$$

En la mayoría de los casos, es suficiente centrarse en una combinación de *soporte*, *confianza*, o bien, *lift* o *leverage* para obtener una buena medida de la calidad de la regla. Sin embargo, cómo de buena es una regla para modelar un conjunto de datos en términos de utilidad es un concepto subjetivo y depende del dominio y del objetivo del problema.

Para una mejor comprensión de las medidas de calidad descritas, en la Tabla 2.1 se muestra un pequeño ejemplo mediante el uso de un conjunto de datos formado por ocho instancias y tres atributos.

Para ello consideremos dos reglas ejemplo, a partir de ahora llamadas Regla 1 y Regla 2, respectivamente:

- Regla **1** : $F_1 \in [30, 38] \wedge F_2 \in [179, 200] \Rightarrow F_3 \in [84, 94]$
- Regla **2** : $F_1 \in [30, 38] \wedge F_2 \in [179, 200] \Rightarrow F_3 \in [46, 94]$

La Regla 1 y la Regla 2 comparten el mismo antecedente aunque el consecuente de ambas es distinto ya que difieren en el intervalo al que pertenece F_3. Intuitivamente, ambas

Tabla 2.1: Ejemplo de un conjunto de datos cuantitativos.

Instancia	F_1	F_2	F_3
t_1	35	183	88
t_2	42	154	47
t_3	37	186	93
t_4	30	199	112
t_5	33	173	83
t_6	24	178	75
t_7	63	177	91
t_8	22	167	60

reglas parecen tener la misma calidad ya que satisfacen el mismo número de instancias del conjunto de datos. Sin embargo, es necesario realizar un análisis de las reglas con el fin de establecer cual es la mejor en base a las medidas de calidad de las AR descritas previamente.

A continuación, se calculan cada una de las medidas de calidad para las reglas 1 y 2.

En la Regla 1, el *soporte* del antecedente es 37.5 %, ya que tres instancias, t_1, t_3 y t_4 satisfacen simultáneamente que F_1 y F_2 pertenecen a los intervalos [30, 38] y [179, 200] respectivamente (tres instancias de ocho, $sup(X) = 0{,}375$). El *soporte* del consecuente es $sup(Y) = 0{,}375$ debido a que las instancias t_1, t_3 y t_7 satisfacen que $F_3 \in [84, 94]$. Con respecto a la *confianza*, sólo dos instancias satisfacen los tres atributos (F_1 y F_2 en el antecedente, y F_3 en el consecuente) que aparecen en la regla y por tanto, $sup(X \Rightarrow Y) = 0{,}25$ y $conf(X \Rightarrow Y) = 0{,}25/0{,}375 = 0{,}66$, es decir, la regla tiene una *confianza* del 66 %. El valor de *lift* es $lift(X \Longrightarrow Y) = 0{,}25/(0{,}375 \cdot 0{,}375) = 1{,}76$, el valor de *leverage* es $lev(X \Longrightarrow Y) = 0{,}25 - (0{,}375 \cdot 0{,}375) = 0{,}23$ y el valor del *accuracy* es $acc(X \Longrightarrow Y) = 0{,}25 + 0{,}5 = 0{,}75$, ya que $sup(X \Rightarrow Y) = 0{,}25$, $sup(\neg X \Longrightarrow \neg Y) = 0{,}5$, $sup(X) = 0{,}375$ and $sup(Y) = 0{,}375$, como se ha mencionado anteriormente. Del mismo modo, $gain(X \Longrightarrow Y) = 0{,}29$, $CF(X \Longrightarrow Y) = 0{,}46$ y $conv(X \Longrightarrow Y) = 1{,}87$.

En la Regla 2, el *soporte* del antecedente es el mismo de la Regla 1, esto es, 37.5 %, ya que tres instancias, t_1, t_3 y t_4 satisfacen simultáneamente que F_1 y F_2 pertenecen a los intervalos [30, 38] y [179, 200] respectivamente. Sin embargo, el *soporte* del consecuente es

$sup(Y) = 0{,}875$ debido a que todas las instancias excepto t_4 satisfacen que $F_3 \in [46, 94]$. La *confianza* en esta regla es la misma que en la Regla 1, sólo dos instancias satisfacen los tres atributos que aparecen en la regla. Por tanto $sup(X \Rightarrow Y) = 0{,}25$ y el valor de la *confianza* es también del 66 %. Con respecto al *lift* o interés, $lift(X \Longrightarrow Y) = 0{,}25/(0{,}375 \cdot 0{,}875) = 0{,}75$, *leverage* es $lev(X \Longrightarrow Y) = 0{,}25 - (0{,}375 \cdot 0{,}875) = -0{,}57$ y el valor del *accuracy* es $acc(X \Longrightarrow Y) = 0{,}25 + 0 = 0{,}25$, ya que $sup(X \Rightarrow Y) = 0{,}25$, $sup(\neg X \Longrightarrow \neg Y) = 0$, $sup(X) = 0{,}375$ y $sup(Y) = 0{,}875$, como se ha mencionado anteriormente. En cuanto al resto de métricas, $gain(X \Longrightarrow Y) = -0{,}21$, $CF(X \Longrightarrow Y) = -0{,}24$ y $conv(X \Longrightarrow Y) = 0{,}37$.

Tabla 2.2: Medidas de calidad para las reglas ejemplo 1 y 2.

Medidas	Regla 1	Regla 2
Soporte del antecedente	0.375	0.375
Soporte del consecuente	0.375	0.875
Soporte de la regla	0.25	0.25
Confianza	0.66	0.66
Lift	1.76	0.75
Leverage	0.23	-0.57
Accuracy	0.75	0.25
Gain	0.29	-0.21
Certainty Factor	0.46	-0.34
Conviction	1.87	0.37

Es de notar que la *confianza* no tiene en cuenta el *soporte* del consecuente de la regla, ya que el valor de la *confianza* es el mismo tanto en la Regla 1 como en la Regla 2 y por tanto no es capaz de detectar dependencias negativas. Las métricas *lift* y *leverage* deberían considerarse para solucionar este inconveniente ya que miden el grado de dependencia entre el antecedente y el consecuente. El valor del *lift* de la Regla 1 y la Regla 2 es 1.76 y 0.75 respectivamente. En este caso, el *lift* de la Regla 2 es inferior a 1 y además menor que el de la Regla 1, lo cual permite intuir que la primera regla es más interesante que la segunda. Del mismo modo sucede con el *leverage* ya que el valor de esta medida para la Regla 2 también es negativo y mayor para la Regla 1. En el caso del valor del *accuracy* también es mayor para la Regla 1. Por tanto, se puede concluir que la primera regla tiene mejor calidad, precisión, interés y fuerte dependencia entre el antecedente y el consecuente que la segunda regla a pesar de que tienen la misma *confianza*.

Hay que destacar que *leverage*, *lift* y *accuracy* en algunos casos no son suficientes ya que también presentan algunos inconvenientes debido a que sólo miden co-ocurrencias y no tienen en cuenta el sentido de la implicación al ser medidas simétricas, es decir, *lift*$(X \Longrightarrow Y)$ es igual que *lift*$(Y \Longrightarrow X)$. *Gain*, *certainty factor* y *conviction* presentan la ventaja de que tienen en cuenta tanto el *soporte* del antecedente como el *soporte* del consecuente y además, toman en consideración el sentido de la implicación de las reglas. En el caso de las dos reglas ejemplo, *gain*, *certainty factor* y *conviction*, también demuestran que la Regla 1 es mejor al presentar valores más altos y positivos, al contrario de lo que sucede con la Regla 2, que tiene valores negativos para *gain* y *certainty factor*.

2.5.2. Algoritmos extracción de Reglas de Asociación

El descubrimiento de AR con atributos numéricos es un tema emergente desde que fue introducido por primera vez en [Srikant and Agrawal, 1996]. La mayoría de los algoritmos para extracción de AR están basados en métodos propuestos por Agrawal et al. tales como AIS [Agrawal et al., 1993] y APRIORI [Agrawal and Srikant, 1994], SETM [Houtsma and Swami, 1995], etc. AIS fue el primer algoritmo que se desarrolló para obtener AR y se caracteriza por extraer reglas de la forma $X \Rightarrow Y$ donde X es un conjunto de ítemsets e Y un único ítem. La principal desventaja que presentaba este algoritmo es que generaba innecesariamente conjuntos candidatos que nunca llegaban a ser relevantes. Sin embargo, APRIORI, que es probablemente el algoritmo más citado y usado en la literatura, se basa en el conocimiento previo o “a priori” de los conjuntos frecuentes utilizando una estrategia de búsqueda basada en anchura. Por otro lado, SETM se deriva del algoritmo AIS y se caracteriza porque fue diseñado para utilizar SQL para la generación de ítemsets relevantes.

Otros de los algoritmos clásicos de la literatura son ECLAT [Zaki, 2000] y FP-GROWTH [Han et al., 2004]. El algoritmo ECLAT se basa en realizar un agrupamiento (clustering) entre los ítems para aproximarse al conjunto de ítems frecuentes maximales y luego emplea algoritmos para generar los ítems frecuentes contenidos en cada grupo. A diferencia de APRIORI, utiliza una estrategia de búsqueda basada en profundidad. Por otro lado, la idea básica del algoritmo FP-Growth (del inglés *frequent pattern growth*) puede ser descrita como un esquema de eliminación recursiva de todos los ítems infrecuentes para construir un árbol de patrones frecuentes (a partir de ahora FP-Tree, del inglés *Frequent Pattern Tree*) para descubrir posteriormente los ítemsets frecuentes.

Sin embargo, al igual que Apriori y los algoritmos mencionados anteriormente, muchos investigadores se centran en bases de datos con atributos discretos, mientras que la mayoría de las bases de datos del mundo real contienen fundamentalmente atributos continuos [Aumann and Lindell, 2003]. La mayoría de las herramientas que aparentemente trabajan en dominios continuos simplemente discretizan los atributos usando una estrategia específica y después, tratan esos atributos como si fueran discretos [Vannucci and Colla, 2004].

A continuación se mencionan algunos de los algoritmos sobre aprendizaje de AR tras una revisión de los trabajos publicados recientemente en la literatura.

Los autores de Genar [Mata et al., 2001] proponen un algoritmo basado en técnicas evolutivas para encontrar QAR, que fue mejorado en [Mata et al., 2002] y llamado Gar. El proceso está dividido en dos fases: en la primera fase se buscan todos los conjuntos de atributos que están presentes frecuentemente en la base de datos y en la segunda fase se extraen reglas a partir de los conjuntos calculados anteriormente.

Otro GA llamado Earmga fue usado en [Yan et al., 2009] para obtener QAR. Sin embargo, el único objetivo a optimizar en la función fitness era la confianza. Para satisfacer este propósito los autores evitan la especificación de umbrales para el soporte mínimo que fue la principal contribución de este trabajo.

Recientemente, en [Alcalá-Fdez et al., 2010] presentaron un estudio sobre tres algoritmos para analizar su efectividad en la extracción de QAR. En particular, los algoritmos citados anteriormente, Earmga [Yan et al., 2009], Gar [Mata et al., 2002] y Genar [Mata et al., 2001] se aplicaron a dos conjuntos de datos del mundo real, mostrando su eficiencia en términos de cobertura y confianza.

En [del Jesus et al., 2011] se hace una revisión sobre el aprendizaje de AR basado en el uso de EA aplicado sobre variables booleanas, categóricas, cuantitativas y difusas, así como las principales aplicaciones de AR.

Algunos GA multiobjetivos para la extracción de AR son los algoritmos introducidos en [Wakabi-Waiswa and Baryamureeba, 2008] y [Ghosh and Nath, 2004]. En el trabajo realizado en [Dehuri et al., 2006], los autores propusieron un algoritmo GA multiobjetivo rápido y escalable para la extracción de AR usando paralelismo y una red homogénea de estaciones de trabajo. La propuesta utiliza la medida confianza, una medida de comprensibilidad basada en el número de atributos que pertenecen a la regla considerando que una regla es más comprensible cuanto menor sea el número de atributos que pertenecen al antecedente y una medida de interés similar al *lift* como funciones objetivo.

MODENAR es un GA multiobjetivo basado en frente de pareto que fue presentado en [Alatas et al., 2008]. La función fitness estaba compuesta de cuatro objetivos diferentes: soporte, confianza, comprensibilidad de la regla (para ser maximizado) y la amplitud de los intervalos que constituyen la regla (para ser minimizado).

En [Qodmanan et al., 2011] fue propuesto otro GA multiobjetivo para extraer AR. Este algoritmo no tiene en cuenta un umbral mínimo para el soporte y la confianza y aplica el algoritmo FP-tree. El objetivo de la función fitness es maximizar la correlación entre el soporte y la confianza.

También se pueden definir AR usando lógica difusa en sustitución de la lógica intervalar. En el trabajo presentado en [Kaya and Alhajj, 2005] se propuso un framework basado en GA para extraer AR difusas. Para ser precisos, se presentó un método de clustering para ajustar los centroides de los clusters y a continuación, proporcionaron un enfoque diferente basado en el conocido algoritmo de clustering CURE [Guha et al., 1998] para generar las funciones de pertenencia.

Más tarde, en [Kaya and Alhajj, 2006] introdujeron un GA para extraer AR ponderadas difusas en el que se optimiza las funciones de pertenencia. Este GA determina automáticamente un umbral mínimo para el soporte y la confianza a partir de los datos. Para conseguir este objetivo, los valores base de las funciones de pertenencia para cada atributo cuantitativo se refinaron mediante la maximización de dos funciones de evaluación diferentes: el número de conjuntos frecuentes grandes y el promedio del intervalo de confianza de las reglas generadas.

Por otra parte, en [Alcalá-Fdez et al., 2009] presentaron un nuevo algoritmo para extraer AR difusas y determinar las funciones de pertenencia por medio de aprendizaje evolutivo basado en el modelo de representación de $2-$tuplas.

El trabajo presentado en [Tong et al., 2005] estudió el conflicto entre problemas de soporte y confianza mínimos cuando se requerían simultáneamente. También se propone un método para encontrar QAR mediante clustering de las transacciones de una base de datos.

Posteriormente, esas agrupaciones se proyectaron en los dominios de los atributos para crear intervalos significativos que podrían ser solapados.

Un sistema clasificador fue presentado en [Orriols-Puig et al., 2008] con el objetivo de extraer QAR sobre flujos de datos (*data streams*) tanto numéricos como categóricos sin etiquetas. La principal novedad de este enfoque radica en la eficiencia y la adaptabilidad de los datos recogidos on-line.

Una optimización metaheurística basada en técnicas de enjambre de partículas aproximado se presentó en [Alatas and Akin, 2008]. En este caso, la singularidad fue la obtención de los valores que determinan los intervalos para AR en lugar de conjuntos de elementos frecuentes. Evaluaron y probaron varios operadores nuevos tales como redondeo, reparación y filtrado en datos sintéticos.

En [Ayubi et al., 2009] propusieron un algoritmo que extraía reglas generales que fueron aplicadas tanto en atributos discretos como en atributos cuantitativos discretizados. Los conjuntos de ítems frecuentes se almacenan en una estructura de árbol de manera que pueden realizarse cálculos a partir de conjuntos más simples mediante recursividad. Esta característica permite obtener beneficios en cuanto a la complejidad computacional, la gestión de memoria y un gran potencial para la paralelización. Los autores también utilizan otros operadores de orden además del operador de igualdad en la parte del antecedente.

Por otro lado, se proponen técnicas de DM en [Bellazzi et al., 2005] para el descubrimiento de AR en series temporales. Los autores extraen satisfactoriamente datos temporales procedentes de múltiples sesiones de hemodiálisis mediante la aplicación de preprocesamiento, reducción de datos y filtrado como un paso previo al proceso de extracción de AR. Finalmente, las AR se obtuvieron siguiendo la conocida estrategia de generación de ítemset de Apriori [Agrawal et al., 1993, Agrawal and Srikant, 1994].

En [Winarko and Roddick, 2007] se introduce un algoritmo para descubrir patrones temporales frecuentes y AR temporales. La propuesta generaliza el algoritmo Memisp (del inglés *Mining Sequential Patterns using I-PrefixSpan*) [Lin and Lee, 2002] que descubre patrones secuenciales mediante técnicas recursivas find-then-index. Fue especialmente notable, la restricción de distancia máxima incluida para eliminar patrones insignificantes y consecuentemente, reducir el número de AR temporales.

En el ámbito de clasificación no supervisada, los autores en [Wan et al., 2007] usaron clustering para discretizar los atributos de series temporales hidrológicas, como un paso previo en la extracción de reglas, que fueron finalmente obtenidas por medio del algoritmo Apriori. Siguiendo con técnicas de clustering, se utilizó clustering difuso en [Chen et al., 2010] con el objetivo de acelerar el cálculo para satisfacer los requerimientos de soportes mínimos múltiples, de forma que se mejoraban los resultados publicados en el trabajo inicial de los mismos autores [Chen et al., 2009].

En [Huang et al., 2008] descubrieron patrones espacio-temporales y presentaron QAR usando los algoritmos PrefixSpan [Pei et al., 2001] y Fiti [Tung et al., 2003] sobre series temporales de datos del océano para descubrir relaciones entre las variaciones de salinidad y temperatura.

El concepto de ventana deslizante ha sido utilizado recientemente [Khan et al., 2010] con el propósito de obtener un bajo uso de memoria y bajo coste computacional basado en el algoritmo Apriori para descubrir ítemsets cuyo soporte aumente a lo largo del tiempo.

Por último, las QAR también han sido usadas extensamente en el área de la bioinformática. En el trabajo de [Georgii et al., 2005] se analizaron datos de microarray utilizando QAR. Para ello, eligieron una variante del algoritmo introducido en [Ruckert et al., 2004], basado en espacios medios o combinaciones lineales de variables delimitadas con una constante.

Por otra parte, en [Gupta et al., 2006] descubrieron AR para secuencias de proteínas mediante un algoritmo que seguía cuatro pasos: primero hacen una partición equidistante de los atributos, en segundo lugar, las particiones se asignan a enteros consecutivos que representan los intervalos; en tercer lugar, encuentran el soporte de todos los intervalos y finalmente usan los conjuntos frecuentes para generar AR.

Dentro de este área, los autores en [Nam et al., 2009] propusieron un nuevo método de extracción de AR temporales basados en el algoritmo Apriori donde se identificaron asociaciones temporales a partir de datos de expresión de genes.

En la siguientes secciones se detallan los algoritmos Apriori, Genar, Modenar y Earmga, los cuales hemos usado para comparar las distintas propuestas de esta tesis doctoral. Cabe destacar que para todos estos algoritmos se han utilizado las implementaciones disponibles en la herramienta Keel [Alcalá-Fdez et al., 2009] salvo para el algoritmo Modenar que se han utilizado los propios resultados publicados en la literatura.

2.5.2.1. Algoritmo APRIORI

El algoritmo Apriori propuesto en [Agrawal and Srikant, 1994] es el algoritmo para encontrar AR más conocido y citado en la literatura en el cual se basan la mayoría de los algoritmos existentes.

La principal mejora respecto a los algoritmos Ais [Agrawal and Srikant, 1994] y Setm [Houtsma and Swami, 1995] se encuentra en la forma de generar los conjuntos candidatos, puesto que obliga a cumplir la propiedad de los conjuntos frecuentes: *cualquier subconjunto de un conjunto frecuente debe ser también un conjunto frecuente*. Mediante esta propiedad se consigue que no se construyan muchos de los conjuntos frecuentes que necesitaban los algoritmos Ais y Setm.

Este algoritmo se basa en el conocimiento previo o "a priori" de los conjuntos frecuentes y utiliza una estrategia de búsqueda basada en anchura. Cabe destacar que este algoritmo obtiene AR discretizando previamente los valores continuos de la base de datos a diferencia de las propuestas desarrolladas en esta tesis doctoral. De este algoritmo existen diferentes variantes como son AprioriTid y AprioriHybrid.

Conceptualmente, el algoritmo Apriori tiene los siguientes pasos para generar los conjuntos de ítems frecuentes:

- Generación de todos los conjuntos de ítems con un elemento. Uso de estos conjuntos para generar los de dos elementos, y así sucesivamente.
- Cálculo del soporte de los conjuntos de ítems de tal forma que se obtiene el conjunto resultante tras eliminar aquellos subconjuntos que no superen el soporte mínimo.

El algoritmo ataca el problema reduciendo el número de conjuntos considerados de forma que el usuario define un soporte mínimo y Apriori genera todos los conjuntos que cumplen con la condición de tener un soporte mayor o igual a ese umbral. Para cada conjunto frecuente X se generan todas las AR $A \Rightarrow C$ tales que $A \cup C = X$ y $A \cap C = \emptyset$. Cualquier regla que no satisfaga las restricciones impuestas por el usuario, como por ejemplo la confianza mínima, se desechan y las reglas que sí lo cumplen se conservan.

2.5.2.2. Algoritmo GENAR

Genar [Mata et al., 2001] usa un modelo evolutivo clásico con elitismo y una función fitness de objetivos agregados. Cada individuo codifica una regla con un número fijo de atributos de forma que sólo el último atributo pertenece al consecuente y el resto al antecedente. A diferencia de Apriori, este algoritmo no discretiza los atributos previamente.

Los operadores genéticos de este método son: operador que selecciona un porcentaje de individuos que tienen mejor fitness, un operador de cruce basado en un punto que intercambia la primera parte del primer individuo padre con la segunda parte del segundo padre para obtener un descendiente. El operador de mutación se basa en alterar uno de los intervalos de la regla.

La función fitness sólo tiene en cuenta el soporte de la regla y una penalización de las reglas que ya han cubierto los mismos registros en la base de datos.

2.5.2.3. Algoritmo MODENAR

Modenar [Alatas et al., 2008] es un algoritmo evolutivo diferencial multiobjetivo basado en el concepto de pareto. Cada individuo del EA es un conjunto de tripletas, donde cada tripleta es un índice a un atributo y dos valores que son los extremos que definen el intervalo de pertenencia de ese atributo. Cada tripleta representa entonces una condición que forma parte del antecedente o el consecuente.

Se caracteriza por eliminar todas las soluciones dominadas y utilizar un operador de filtrado para borrar las soluciones que están mas cercanas a otras en caso de que el número de soluciones no dominadas exceda un determinado umbral. En cuanto a los operadores genéticos, la mutación consiste en realizar perturbaciones de los individuos a partir de la diferencia ponderada de otros dos individuos seleccionados. El operador de cruce genera un nuevo individuo mezclando los individuos mutados con los individuos originales según una probabilidad de distribución.

El reemplazamiento de la población se realiza de la siguiente manera: se seleccionan tres padres para generar un nuevo descendiente que sustituirá al primer individuo padre en la población de la siguiente generación si lo domina. En otro caso, se calcula una función fitness de objetivos ponderados tanto para el descendiente como para el primer padre y se elige para formar parte de la población aquel que tenga mayor valor fitness.

Los objetivos que Modenar intenta maximizar son el soporte, la confianza, la comprensibilidad y la amplitud.

2.5.2.4. Algoritmo EARMGA

El algoritmo Earmga [Yan et al., 2009] está basado en un algoritmo genético para identificar AR sin especificar umbrales de soporte mínimo y está basado en un FP-tree generalizado.

Los individuos de Earmga codifican una regla de $k+1$ genes, donde el gen 0 indica la posición del punto de corte entre el último ítem que pertenece al antecedente y el primer ítem que pertenece al consecuente. El resto de k genes representan cada uno de los ítems que se ordenan de forma ascendiente en cada una de las partes. Al igual que Apriori, Earmga realiza una discretización previa con la salvedad de que agrupa una serie de intervalos base resultantes de la discretización para representar los intervalos de los atributos de una determinada regla.

En cuanto a los operadores genéticos, este algoritmo utiliza una determinada probabilidad para seleccionar los individuos en función de su fitness y el operador de cruce está basado en una estrategia de dos puntos elegidos aleatoriamente para intercambiar cada uno de los segmentos de los genes de los individuos padres para obtener los nuevos descendientes. El operador de mutación también se aplica según una probabilidad y puede modificar tanto el índice del atributo como los intervalos bases asociados a él.

La función fitness utiliza una medida definida como *confianza relativa* y está definida en base al soporte de la regla y el soporte del antecedente y del consecuente.

2.6. Resumen de aportaciones

En esta sección se muestran las principales contribuciones que se han desarrollado durante el proceso de investigación de esta tesis doctoral. Dado que se trata de una tesis doctoral por compendio de artículos, las publicaciones se presentan de acuerdo a la relevancia para la consecución de la misma y según su organización en las distintas partes de esta memoria.

La selección de artículos en revistas que se incluyen en el ranking JCR de Thomsom-Reuters y se presentan en el Capítulo 4 son:

- M. Martínez-Ballesteros, F. Martínez-Álvarez , A. Troncoso Lora, J.C. Riquelme Santos. *An Evolutionary Algorithm to Discover Quantitative Association Rules in Multi-*

dimensional Time Series. Soft Computing (SOCO). Vol. 15, No. 10, pp. 2065-2084, 2011 [Martínez-Ballesteros et al., 2011a] [JCR 2.122, JCR-5 1.672, Q1].

- M. Martínez-Ballesteros, F.Martínez-Álvarez, A. Troncoso Lora, J.C. Riquelme Santos. *Mining quantitative association rules based on evoluationary computation and its application to atmospheric pollution.* Integrated Computer-Aided Engineering (ICAE). Vol. 17, No. 3, pp. 227-242, 2010 [Martínez-Ballesteros et al., 2010] [JCR 1.512, JCR-5 1.350, Q2].

- M. Martínez-Ballesteros, S. Salcedo-Sanz, J. C. Riquelme, C. Casanova-Mateo, J. L. Camacho. *Evolutionary Association Rules for Total Ozone Content Modeling from Satellite Observations.* Chemometrics and Intelligent Laboratory Systems. Vol 109, No. 2, pp. 217-227, December 2011 [Martínez-Ballesteros et al., 2011] [JCR 2.222, JCR-5 2.415, Q1].

Otros trabajos de investigación relevantes para esta tesis publicados en revistas no indexadas o actas de congresos internacionales que se encuadran dentro de la temática de esta tesis doctoral y se han incluido en el Capítulo 5:

- M. Martínez Ballesteros, F. Martínez Álvarez, A. Troncoso Lora, J.C. Riquelme Santos. *Quantitative association rules applied to climatological time series forecasting.* International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'09). Lecture Notes in Computer Science, Vol. 5788, pp. 284-291, 2010 [Martínez-Ballesteros et al., 2009] [Ranking CORE C].

- M. Martínez Ballesteros, C. Rubio Escudero, J.C. Riquelme Santos. *Mining Quantitative Association Rules in Microarray data using Evolutive Algorithms.* Proceedings of 3^{rd} International Conference on Agents and Artificial Intelligence (ICAART 2011). No. 3. Rome, Italy, pp. 574-577, 2011 [Martínez-Ballesteros et al., 2011b] [Ranking CORE C]

- M. Martínez-Ballesteros, J.C. Riquelme. *Analisys of Measures of Quantitative Association Rules.* International Conference on Hybrid Artificial Intelligence Systems (HAIS 2011). Lecture Notes in Artificial Intelligence, Vol. 6679, No. 2, pp. 319-326, 2011[Martínez-Ballesteros and Riquelme, 2011] [Ranking CORE C]

- M. Martínez-Ballesteros, I. Nepomuceno-Chamorro, J.C. Riquelme. *Inferring Gene-Gene Associations from Quantitative Association Rules.* IEEE International Conference on Intelligent Systems Design and Applications (ISDA'11), 2011. [Ranking CORE C]

Otras publicaciones relacionadas con la tesis doctoral que no han sido incluidas en esta memoria son:

- M. Martínez Ballesteros, F.Martínez-Álvarez, A. Troncoso Lora, J.C. Riquelme Santos. *Descubriendo Reglas de Asociación Numéricas entre Series Temporales.* Workshop on Mining of Non-Conventional Data (MINCODA'09). Sevilla 2009.

- M. Martínez Ballesteros, C. Rubio Escudero, J.C. Riquelme Santos. *Extracción de Reglas de Asociación Cuantitativas en datos de Microarray aplicando Algoritmos Evolutivos.* VII Congreso Español sobre Metaheurísticas, Algoritmos Evolutivos y Bioinspirados (MAEB 2010), Valencia, 2010.

Publicaciones no relacionadas con la tesis doctoral que no han sido incluidas en esta memoria son:

- M. Martínez Ballesteros, Víctor M. Rivas. *EvFuzzySystem: Evolución de Sistemas Difusos para Problemas de Regresión Multi-Dimensionales* XV Congreso Español sobre Tecnologías y Lógica Fuzzy (ESTYLF 2010). Huelva, 2010.

- F. Martínez-Álvarez, C. Rubio Escudero, M. Martínez-Ballesteros, J.C. Riquelme Santos. *On the use of algorithms to discover motifs in DNA sequences.* IEEE International Conference on Intelligent Systems Design and Applications (ISDA'11), 2011. [Ranking CORE C]

Capítulo 3

Discusión Conjunta de los Resultados

Esta sección resume las distintas propuestas que se incluyen en la Parte II, y especialmente aquellas que se presentan en el Capítulo 4 de la memoria. El objetivo de este capítulo consiste en dar una visión global de las publicaciones que han motivado la realización de esta tesis doctoral en la modalidad de compendio de artículos, así como una breve discusión conjunta sobre los resultados obtenidos para cada una de ellas.

Hemos de puntualizar que los trabajos presentados en esta memoria están encadenados, por lo que seguidamente describiremos la relación entre las diferentes secciones que se incluyen en la Parte II.

En primer lugar, se ha desarrollado un algoritmo de codificación genética real, en adelante llamado QARGA (del inglés *Quantitative Association Rules by Genetic Algorithm*), para encontrar relaciones existentes entre variables utilizando QAR (Sección 4.1). El algoritmo utiliza una codificación particular de los individuos que no realiza una previa discretización de los atributos y además no obliga a establecer las variables que pertenecen al antecedente o al consecuente.

Posteriormente, se ha desarrollado un algoritmo de codificación real (Sección 4.2) basado en el anterior, a partir de ahora denominado QARGA-CHC , que extiende el esquema general del algoritmo evolutivo de codificación binaria CHC [Eshelman, 1991].

El tercer algoritmo que hemos desarrollado, en adelante llamado EQAR (del inglés *Evolutionary Quantitative Association Rules*), se trata de una versión mejorada de QARGA-CHC

(Sección 4.3) que añade nuevas características. Tanto Qarga-Chc, como Eqar, pretenden establecer un equilibrio entre la diversidad y la convergencia que supone un problema en los algoritmos genéticos.

Las tres propuestas anteriores usan un esquema basado en pesos para la función fitness que implica algunas medidas de evaluación de las AR, sin embargo, utilizando un esquema multiobjetivo basado en soluciones no dominadas podrían reducirse estas medidas y mejorar algunos de los inconvenientes que a veces puede plantear una función lineal de objetivos ponderados.

Por tanto, la última propuesta denominada Moqar (del inglés *Multi-Objective evolutionary algorithm based on Quantitative Association Rules extraction*) ha consistido en un algoritmo evolutivo multiobjetivo (Sección 5.4) que sigue el esquema del algoritmo Nsga-II (del inglés *Non-dominated Sorting Genetic Algorithm II*) [Deb et al., 2002], en el que previamente se han analizado las relaciones entre las diferentes medidas de evaluación de AR (Sección 5.3) como primer paso para seleccionar los mejores objetivos implicados en dicho algoritmo.

Cabe destacar que las propuestas desarrolladas han sido evaluadas en distintos ámbitos de aplicación. Las distintas bases de datos utilizadas se detallan a continuación:

- Bases de datos sintéticas (Secciones 4.1 y 4.2).
- Bases de datos públicas procedentes del repositorio BUFA (del inglés *Bilkent University Function Approximation*) [Guvenir and Uysal, 2000] (Sección 4.1).
- Series temporales climatológicas del mundo real obtenidas de una Estación Meteorológica de Sevilla (Secciones 4.1, 4.2 y 5.1).
- Conjunto de datos de ozono proporcionados por la Agencia Estatal de Meteorología (AEMET) para el modelado de TOC (del inglés *Total Ozone Column*) en la Península Ibérica (Sección 4.3).
- Conjunto de datos procedentes de un experimento de Microarrays relacionados con la inflamación y la inmunidad [Calvano et al., 2005] (Sección 5.2).
- Conjunto de datos de Microarrays del ciclo celular de la levadura (Saccharomyces cerevisiae) [Spellman et al., 1998] y [Cho et al., 1998] (Sección 5.4).

Los detalles relacionados con cada una de las propuestas se describen en la Parte II distribuidas en los Capítulos 4 y 5 para distinguir las publicaciones de revistas indexadas de otras publicaciones en revistas no indexadas y actas de congresos.

A continuación se muestra un resumen de las distintas propuestas que se recogen en la presente memoria, así como una breve discusión sobre los resultados obtenidos por cada una de ellas.

3.1. QARGA

El algoritmo llamado Qarga (del inglés *Quantitative Association Rules by Genetic Algorithm*) es un GA de codificación real cuyo objetivo es encontrar las relaciones existentes entre distintas variables mediante QAR. Este algoritmo utiliza una codificación particular de los individuos para solucionar dos problemas básicos. Por un lado, no lleva a cabo discretización previa de los atributos y en segundo lugar, no es necesario establecer el conjunto de variables que pertenecen al antecedente o al consecuente. Por tanto, es posible descubrir todas las dependencias subyacentes entre diferentes variables. La búsqueda de los intervalos más apropiados se llevan a cabo mediante Qarga de forma que los intervalos se ajustan para encontrar AR con alto soporte y confianza, junto con otras medidas usadas para cuantificar la calidad de las reglas.

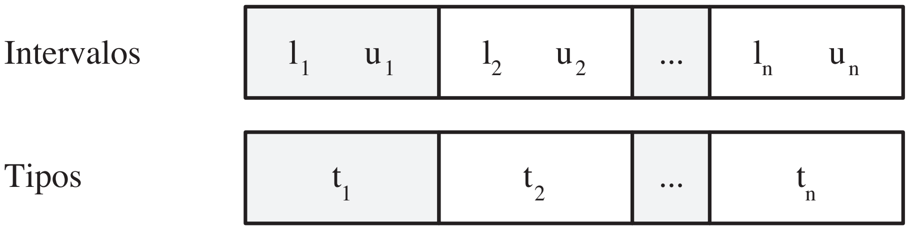


Figura 3.1: Representación de un individuo de la población.

Cada uno de los individuos de la población constituye una regla y cada gen de un individuo representa los límites de los intervalos y el tipo de cada atributo para indicar si pertenece al antecedente, consecuente o no pertenece a la regla. Por lo tanto, la representación de un individuo consiste en dos estructuras de datos como muestra la Figura 3.1. La estructura superior incluye todos los atributos de la base de datos, donde l_j representa el límite inferior del intervalo y u_j representa el límite superior. La estructura inferior indica la pertenencia de un atributo a la regla representada por un individuo. El tipo de cada atributo t_j puede tener tres valores: 0 cuando el atributo no pertenece a la regla, 1 si pertenece al antecedente de la regla y 2 cuando pertenece al consecuente de la regla.

Los individuos de la población se someten a un proceso evolutivo en el cual se aplican operadores de mutación y cruce. Al final del proceso, el individuo que presenta mejor fitness se designa como la mejor regla. Además, la función fitness ha sido provista de un conjunto de parámetros para que el usuario pueda dirigir el proceso de búsqueda dependiendo de las reglas deseadas.

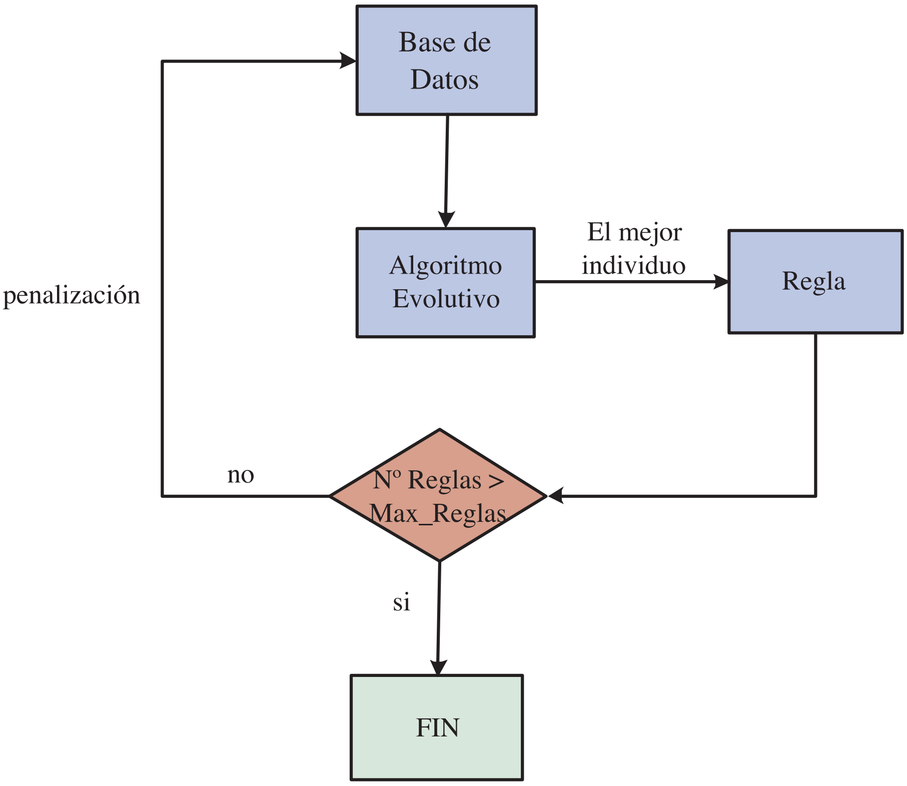


Figura 3.2: Esquema del aprendizaje IRL de Qarga .

El algoritmo propuesto lleva a cabo un proceso de aprendizaje iterativo (IRL, del inglés *Iterative Rule Learning*) [Venturini, 1993] mediante la penalización de las instancias cubiertas por una regla para que las siguientes reglas que QARGA descubra intenten cubrir aquellas instancias que aún no han sido cubiertas. El esquema general de IRL se ilustra en la Figura 3.2 y el esquema evolutivo de QARGA se describe en la Figura 3.3.

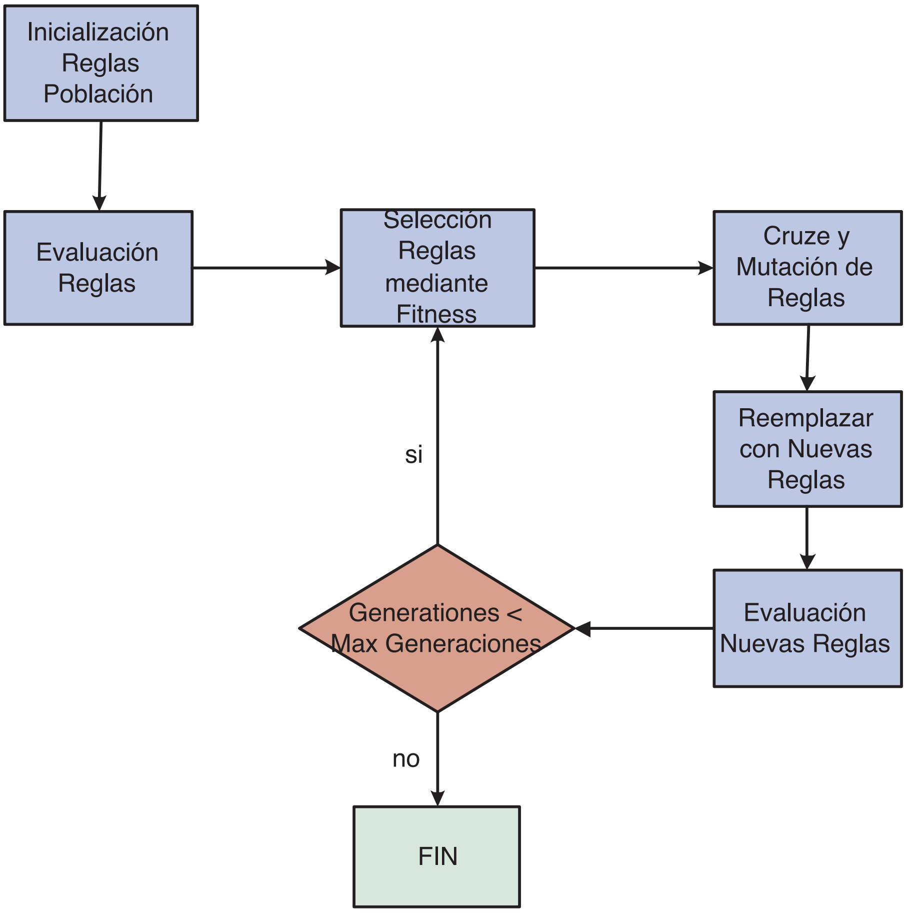

Inicialización Reglas Población
Evaluación Reglas
Selección Reglas mediante Fitness
Cruze y Mutación de Reglas
Reemplazar con Nuevas Reglas
Evaluación Nuevas Reglas
Generationes < Max Generaciones
si
no
FIN

Figura 3.3: Esquema del proceso evolutivo de QARGA .

En cada iteración, el algoritmo evolutivo se ejecuta obteniéndose el individuo que representa la mejor regla y a continuación se marcan las instancias que se cubren mediante la regla elegida. El proceso iterativo termina cuando se encuentra el número máximo de reglas deseadas.

Para evaluar el algoritmo propuesto, se han llevado a cabo varios tipos de experimentos que se describen como sigue:

- En primer lugar, se han analizado varios conjuntos de datos públicos procedentes del repositorio BUFA [Guvenir and Uysal, 2000] (del inglés *Bilkent University Function Approximation*) con el propósito de realizar una comparativa con otros EA existentes. En concreto, Qarga ha sido comparado con los algoritmos EARMGA [Yan et al., 2009] y GENAR [Mata et al., 2001] y los tres algoritmos han sido evaluados sobre 15 bases de datos: Basketball, Bodyfat, Bolts, Kinematics, Longley, Normal Body Temperature, Plastic, Pollution, Pw Linear, Pyramidines, Quake, Schools, Sleep, Stock Price y Vineyard.

- Qarga también ha sido aplicado sobre dos tipos de series temporales multidimensionales sintéticamente generadas (donde las relaciones son conocidas) para analizar su potencial para descubrir reglas en series temporales.

- Finalmente, Qarga ha descubierto QAR en series temporales multidimensionales del mundo real compuestas por variables climatológicas. En concreto, las series temporales están compuestas por la temperatura, humedad, dirección y velocidad del viento, hora del día, día de la semana y finalmente el ozono troposférico O_3 que se trata de un agente contaminante influenciado por las variables anteriores. Todas las variables han sido obtenidas de una estación meteorológica de la ciudad de Sevilla (España) durante los meses de Julio y Agosto de los años 2003 y 2004, periodos con una alta concentración de ozono en la atmósfera. Para fines predictivos, las series temporales climatológicas han sido forzadas a pertenecer al antecedente mientras que el ozono se ha forzado a que pertenezca al consecuente. El algoritmo Apriori [Agrawal and Srikant, 1994] se ha evaluado sobre estas variables climatológicas con el objetivo de establecer una comparativa con Qarga.

Todos los resultados obtenidos de los experimentos anteriores muestran que Qarga actúa notablemente ya que se han obtenido mejores resultados cuando se ha comparado con otros EA como Earmga y Genar en bases de datos públicas del repositorio BUFA. En términos generales, Qarga presenta mejores resultados en cuanto al soporte, confianza y amplitud lo cual conduce a que Qarga presenta reglas más fiables, con menos errores y además, obtiene reglas más comprensibles ya que el número de atributos que aparecen en el antecedente y en el consecuente es más pequeño que en el resto de algoritmos y por tanto proporcionan al usuario una mejor comprensión. En esta experimentación también se han llevado a cabo una serie de análisis estadísticos para evaluar a QARGA siguiendo los procedimientos no paramétricos discutidos en [García et al., 2009]. En concreto se han aplicado los test estadísticos de Friedman e Iman-Davenport, Holm-Hochberg y Bonferroni-Dunn y todos han determinado que QARGA es el mejor algoritmo con respecto a EARMGA y GENAR.

En cuanto a las series temporales multidimensionales sintéticamente generadas, también se han obtenido buenos resultados ya que las reglas extraídas por QARGA han sido bastante similares a las reglas esperadas. Asimismo, han presentado valores altos tanto para confianza como para *lift* por lo que pueden considerarse bastante precisas e interesantes. La amplitud de los intervalos han sido moderados y los límites de los intervalos son similares a los límites de las reglas esperadas.

Para finalizar, QARGA ha obtenido con éxito QAR significativas en series temporales multidimensionales del mundo real. En particular, se han encontrado dependencias relevantes entre la concentración de ozono y otras series temporales climatológicas como son la temperatura, humedad, día de la semana, hora del día, velocidad y dirección del viento. En este tipo de experimentación, podemos concluir que QARGA obtiene mejores resultados cuando es comparado con el algoritmo APRIORI, ya que el soporte, la confianza y el *lift* son mayores y la amplitud de los intervalos es menor, lo cual implica reglas más interesantes, más interpretables y con menos errores.

Otra observación relevante es que APRIORI descubre las reglas con intervalos diferentes para la misma variable en el antecedente, pero igual intervalo en el consecuente y viceversa. Este problema se debe a la discretización de los intervalos ya que en este caso, es necesario utilizar más de una regla para representar una relación entre dos variables, sin embargo, este hecho no ocurre en QARGA. Otro de los problemas que presenta APRIORI se debe a la discretización de la base de datos en tres intervalos. En este caso el algoritmo únicamente encuentra reglas cuando los valores de ozono varían en dos de los tres posibles intervalos, y por consiguiente, le resulta imposible encontrar reglas con altas concentraciones de ozono, a diferencia de QARGA donde los resultados han demostrado que sí es posible.

En otros trabajos, QARGA ha sido aplicado a datos procedentes de un experimento de microarrays relacionados con la inflamación y la inmunidad [Calvano et al., 2005] que se llevó a cabo en la Universidad de St. Louis, Missouri [Calvano et al., 2005], donde la sangre de ocho voluntarios fue analizada y cuatro de ellos fueron tratados con una toxina que producía un proceso de inflamación y los otros cuatro fueron tratados con placebo. Las muestras fueron tomadas en 6 puntos de tiempo en un periodo de 24 horas. En este tipo de experimentación también se ha demostrado que QARGA es un método válido para analizar datos de microarray y además, mediante el programa Onto-CC [Romero-Zaliz et al., 2008] se ha comprobado que las relaciones entre genes que han aparecido en las reglas son relevantes.

Los artículos asociados a esta parte son:

- (Sección 4.1) M. Martínez-Ballesteros, F. Martínez-Álvarez , A. Troncoso Lora, J.C. Riquelme Santos. *An Evolutionary Algorithm to Discover Quantitative Association Rules in Multidimensional Time Series.* Soft Computing. Vol. 15, No. 10, pp. 2065-2084. 2011 [Martínez-Ballesteros et al., 2011a] [JCR 2.122, JCR-5 1.672, Q1].

- (Sección 5.1) M. Martínez Ballesteros, F. Martínez Álvarez, A. Troncoso Lora, J.C. Riquelme Santos. *Quantitative association rules applied to climatological time series forecasting.* International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'09), Lecture Notes in Computer Science, Vol. 5788, pp. 284-291, 2010. [Ranking CORE C] [Martínez-Ballesteros et al., 2009].

- (Sección 5.2) M. Martínez Ballesteros, C. Rubio Escudero, J.C. Riquelme Santos, F. Martínez Álvarez. *Mining Quantitative Association Rules in Microarray data using Evolutive Algorithms.* Proceedings of 3rd International Conference on Agents and Artificial Intelligence (ICAART 2011). Num. 3. Rome, Italy, pp. 574-577, 2011. [Ranking CORE C] [Martínez-Ballesteros et al., 2011a]

- M. Martínez Ballesteros, F.Martínez-Álvarez, A. Troncoso Lora, J.C. Riquelme Santos. *Descubriendo Reglas de Asociación Numéricas entre Series Temporales.* Workshop on Mining of Non-Conventional Data (MINCODA'09). Sevilla 2009.

- M. Martínez Ballesteros, C. Rubio Escudero, J.C. Riquelme Santos. *Extracción de Reglas de Asociación Cuantitativas en datos de Microarray aplicando Algoritmos Evolutivos.* VII Congreso Español sobre Metaheurísticas, Algoritmos Evolutivos y Bioinspirados (MAEB 2010), Valencia, 2010.

3.2. QARGA-CHC

El algoritmo Qarga-Chc extiende al algoritmo Qarga descrito en la Sección 3.1. Se trata de un algoritmo genético con codificación real para obtener AR a partir de conjuntos de datos cuantitativos. Qarga-Chc sigue el esquema general del algoritmo evolutivo de codificación binaria Chc propuesto en [Eshelman, 1991] que se ha detallado en la Sección 2.3.

El esquema Chc original presenta una estrategia elitista para la selección de la población que constituirá la próxima generación e incluye una fuerte diversidad en el proceso evolutivo a través de mecanismos de prevención de incesto y un operador de cruce específico llamado *Half Uniform* (HUX). Por otra parte, la población se reinicializa cuando su diversidad es pobre.

Sin embargo, Qarga-Chc adopta una estrategia de reinicialización más conservadora y un operador de cruce menos perjudicial que el esquema de cruce HUX.

Dado que Qarga-Chc extiende al algoritmo Qarga, comparte la misma codificación de individuos, la función fitness, así como el esquema de aprendizaje IRL para llevar a cabo la penalización de las instancias que ya han sido cubiertas previamente. La Figura 3.4 ilustra como trabaja el algoritmo propuesto Qarga-Chc.

En primer lugar, la población se inicializa y se establece el umbral de cruce (distancia mínima *MinDist*) para la prevención de incesto. En cada iteración de Qarga-Chc, se evalúa la población y se calcula el fitness de cada individuo. A continuación, se aplica el operador de cruce que utiliza Qarga salvo que en este caso sólo se lleva a cabo un número máximo de cruces igual a la mitad de la población. Aquellos padres que superan *MinDist* se cruzan para prevenir el incesto y generar nuevos descendientes. Por lo tanto, se garantiza una distancia mínima entre descendientes y padres. Seguidamente tiene lugar la selección

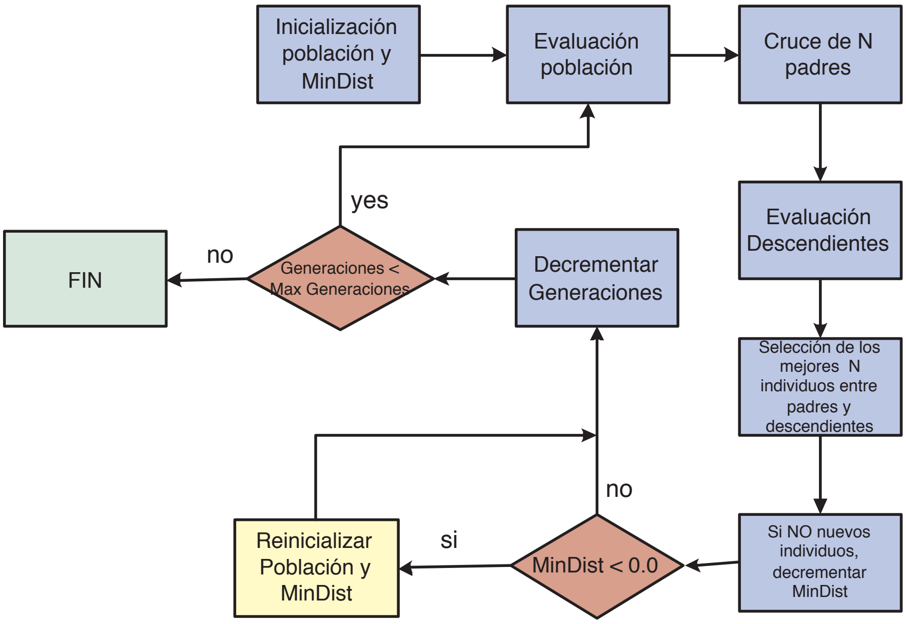

Inicialización población y MinDist
Evaluación población
Cruce de N padres
Evaluación Descendientes
Selección de los mejores N individuos entre padres y descendientes
Si NO nuevos individuos, decrementar MinDist
MinDist < 0.0
no
si
Reinicializar Población y MinDist
Decrementar Generaciones
Generaciones < Max Generaciones
yes
no
FIN

Figura 3.4: Esquema del proceso evolutivo de QARGA-CHC .

elitista de forma que se eligen los mejores N individuos entre la generación actual y los descendientes. Si no se crean nuevos individuos en la generación actual, se decrementa el parámetro *MinDist*. En caso de que el umbral sea menor que cero, la población y el umbral se reinician. Por último, el proceso ha de llevarse a cabo tantas veces como indica el número máximo de generaciones.

Para evaluar QARGA-CHC se ha realizado la siguiente experimentación:

- En primer lugar, QARGA-CHC ha sido aplicado al mismo conjunto de datos sintéticos utilizados en [Alatas et al., 2008] con el objetivo de determinar si es posible encontrar QAR con los valores precisos para los intervalos numéricos a los que pertenece cada atributo de la regla. También se han añadido diferentes niveles de ruido a los conjuntos de datos sintéticos para validar la eficiencia de QARGA-CHC. Los resultados han sido comparados con los obtenidos por el algoritmo MODENAR [Alatas et al., 2008].

- Al igual que QARGA, QARGA-CHC ha sido aplicado sobre las variables climatológicas temperatura, humedad, dirección y velocidad del viento, hora del día y día de la semana, sin embargo, en este caso, además del ozono O_3 también se han incluido otros agentes contaminantes como el monóxido de nitrógeno NO y el dióxido de azufre SO_2. Todas las variables han sido obtenidas de una estación meteorológica de Sevilla (España).

Los resultados obtenidos en los distintos conjuntos de datos han demostrado el buen funcionamiento de Qarga-Chc. En cuanto a las bases de datos sintéticas sin ruido, podemos decir que las reglas encontradas por Qarga-Chc son más precisas que las encontradas por Modenar aunque el soporte y la confianza de ambos algoritmos son los mismos. En cuanto a las bases de datos que contienen ruido, se puede afirmar que Qarga-Chc extrae reglas satisfactoriamente ya que mostró su capacidad para superar los diferentes niveles de ruido, incluso obteniendo una mejora con respecto a las reglas proporcionadas por Modenar en términos de soporte y confianza.

Para las bases de datos de variables climatológicas y agentes contaminantes, Qarga-Chc ha actuado notablemente ya que se han encontrado dependencias relevantes entre la concentración de ozono en la atmósfera y el resto de variables, y de igual forma en los otros agentes contaminantes presentes. Además las relaciones presentadas en las reglas descubiertas concuerdan con los procesos químicos asociados a estos agentes.

El artículo asociado a esta parte es:

- (Sección 4.2) M. Martínez-Ballesteros, F.Martínez-Álvarez, A. Troncoso Lora, J.C. Riquelme Santos. *Mining quantitative association rules based on evoluationary computation and its application to atmospheric pollution.* Integrated Computer-Aided Engineering (ICAE), Vol. 17, No. 3, pages 227-242, 2010 [Martínez-Ballesteros et al., 2010] [JCR 1.512, JCR-5 1.350, Q2].

3.3. EQAR

La tercera propuesta, Eqar (del inglés *Evolutionary Quantitative Association Rules*), también es un EA para extraer QAR a partir de bases de datos cuantitativas. Eqar extiende el algoritmo Qarga-Chc descrito en la Sección 3.2 y por tanto incluye las características anteriormente descritas además de nuevas añadidas para mejorar el funcionamiento del algoritmo y la calidad de las reglas obtenidas. En concreto, se ha modificado la generación de la población inicial del algoritmo evolutivo para beneficiar a los ejemplos que han sido cubiertos por pocas reglas y además la función fitness ha sido ampliada.

En cuanto a la generación de la población inicial, tanto en Qarga como en Qarga-Chc, se llevaba a cabo eligiendo una instancia de la base de datos aleatoriamente para asegurarnos de que las reglas generadas cubran al menos un ejemplo. Sin embargo, en Eqar los ejemplos de la base de datos se seleccionan en función de su nivel dentro de una jerarquía que se organiza según la inversa del número de reglas que cubre un ejemplo. De esta manera los ejemplos cubiertos por pocas reglas tienen más prioridad para ser seleccionados.

Otra de las novedades que presenta Eqar con respecto a Qarga y a Qarga-Chc es la inclusión de una nueva función fitness. Esta mejora se debe a que la función fitness que utilizan Qarga y Qarga-Chc no es del todo apropiada en algunas situaciones ya que la confianza presenta algunos inconvenientes. En concreto, la confianza no tiene en cuenta el soporte del consecuente de la regla y por tanto no es capaz de detectar dependencias negativas entre ítems. Por esta razón, se ha propuesto una nueva función fitness que incluye las medidas *lift* y *leverage* como alternativa a las medidas de soporte y confianza que eran optimizadas en la anterior función fitness.

Eqar ha sido aplicado sobre el problema del modelado de series de ozono procedentes de observaciones de satélites conocido como TOC (del inglés *Total Ozone Column*) que se trata de un indicador más completo para el cambio climático que las series de ozono troposférico. En concreto, se han analizado modelos de series de TOC en la Península Ibérica mediante la extracción de QAR. Los datos han sido proporcionados por la Agencia Estatal de Meteorología (AEMET) y proceden de mediciones de datos de TOMS (del inglés *Total Ozone Mapping Spectrometer*) a bordo del satélite NASA Nimbus-7 medidos en Lisboa, Madrid, Murcia, Arenosillo y Montlouis. Las medidas que se han utilizado como variables de entrada y por tanto forzadas a pertenecer al antecedente son: TP (del inglés *Tropopause height*) medido en hPa, OLR (del inglés *Outgoing longwave radiation*) medido en Wm^{-2}, t_{50} (temperatura a 50 hPa) y ω_{200} (velocidad del viento vertical). La variable TOC medida en DU (Unidades Dobson) se ha forzado a pertenecer al consecuente.

Eqar ha obtenido QAR para las ciudades de Lisboa, Madrid y Murcia y éstas han sido analizadas por el grupo de expertos para seleccionar aquellas que tenían más relevancia meteorológica. El conjunto resultante de QAR ha sido validado sobre las ciudades de Arenosillo y Montlouis utilizando como medida de validación la medida *accuracy*.

Los resultados obtenidos por Eqar han sido bastante buenos en cuanto a los valores de las medidas de calidad de las QAR encontradas, así como los obtenidos en la validación de las mismas. Además el resultado del análisis de las reglas están de acuerdo con los resultados obtenidos en otros trabajos sobre modelización de TOC, por lo que podemos concluir que el uso de AR en modelado de TOC podría ser un método de análisis interesante para el futuro en este problema y otros problemas similares.

El artículo asociado a esta parte es:

- (Sección 4.3) M. Martínez-Ballesteros, S. Salcedo-Sanz, J. C. Riquelme, C. Casanova-Mateo, J. L. Camacho. *Evolutionary Association Rules for Total Ozone Content Modeling from Satellite Observations.* Chemometrics and Intelligent Laboratory Systems. Vol 109, No. 2, pp 217-227, 2011. [Martínez-Ballesteros et al., 2011] [JCR 2.222, JCR-5 2.415, Q1].

3.4. MOQAR

El algoritmo Moqar es una mejora de Qarga, Qarga-Chc y Eqar ya que estos algoritmos utilizan una combinación ponderada de medidas de evaluación en la función fitness, al igual que muchos otros algoritmos de descubrimiento de QAR que están considerados como enfoques multiobjetivo. Sin embargo, una combinación de objetivos suele presentar problemas ya que puede tratarse como una optimización de un sólo objetivo, puede haber desequilibrio entre los distintos objetivos debido a que la optimización de un objetivo puede afectar negativamente a otro. Además no son capaces de identificar todas las soluciones no dominadas, ya que existe una dependencia de los valores de los pesos especificados y son subjetivos pues requieren un conocimiento previo del problema.

Moqar extiende Qarga, Qarga-Chc y Eqar a un enfoque multiobjetivo basado en el algoritmo Nsga-II (del inglés *Non-dominated Sorting Genetic Algorithm II*) [Deb et al., 2002] cuyo principal propósito es la evolución de la población basada en la ordenación de soluciones en frentes de dominancia. El primer frente se compone de soluciones no dominadas por ninguna otra solución de la población (primer frente de Pareto), el segundo se compone de soluciones dominadas por una solución, el tercero por soluciones dominadas por dos soluciones, y así sucesivamente.

Para seleccionar los mejores objetivos, se ha llevado a cabo un estudio estadístico para analizar las relaciones y dependencias entre las medidas de interés de las AR. Concretamente, se han aplicado coeficientes de correlación y análisis de componentes principales entre diferentes medidas con el objetivo de encontrar grupos entre ellas y seleccionar la más representativa para cada grupo.

El esquema del algoritmo se muestra en la Figura 3.5.

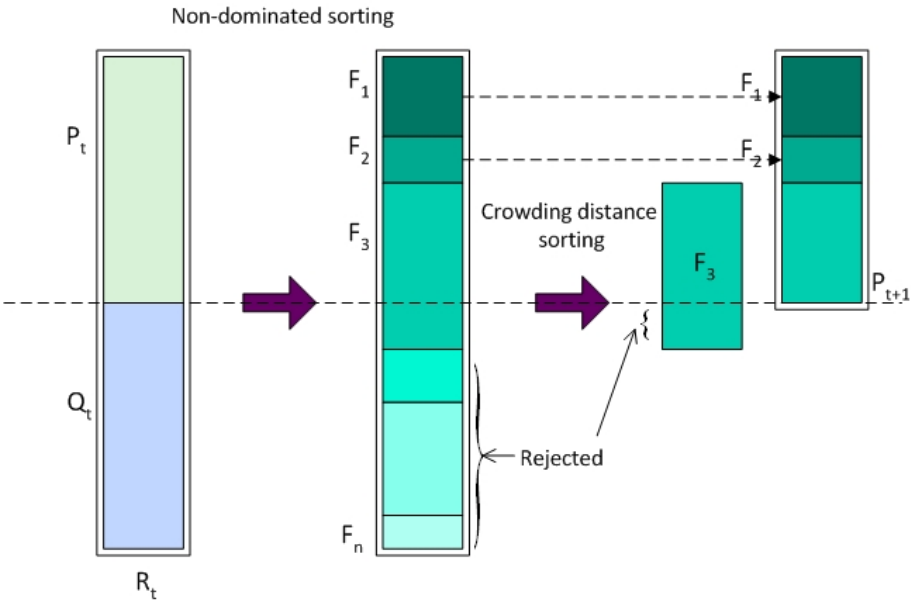


Figura 3.5: Esquema del algoritmo Moqar basado en el esquema de Nsga-II.

Tanto la población como los individuos descendientes se evalúan en función de los objetivos seleccionados para ordenarse en diferentes frentes de pareto según su dominancia (*Fast Nondominated Sorting*) [Deb et al., 2002]. Para elegir los N mejores individuos, se añaden todos los individuos de cada frente hasta completar la población, en caso de que todos los individuos de un determinado frente no puedan añadirse, se seleccionan aquellos que se sitúan más distantes a sus vecinos (*Crowding Distance Sorting*) [Deb et al., 2002].

El proceso se repite hasta alcanzar el número de generaciones máximas, así como el número de reglas deseadas.

Para evaluar el funcionamiento de MoQAR se han utilizado bases de datos de expresión génica procedentes de microarrays para el ciclo celular de la levadura (Saccharomyces cerevisiae) [Spellman et al., 1998] y [Cho et al., 1998]. Se han analizado los mismos datos de entrenamiento que fueron utilizados en [Soinov et al., 2003] para poder realizar una comparativa de ambos métodos. Se han obtenido QAR a partir de estos datos mediante las cuales se ha construido un grafo con las relaciones obtenidas, de forma que cada uno de los atributos que pertenecen a cada regla es un nodo del grafo y cada una de las asociaciones obtenidas es una arista del grafo. Para obtener el grafo resultante se ha realizado la intersección entre los grafos obtenidos en distintas ejecuciones con el objetivo de obtener las relaciones más relevantes.

Cabe destacar que MoQAR ha sido capaz de obtener todas las relaciones inferidas mediante el método basado en un árbol de decisión propuesto en [Soinov et al., 2003] además de siete relaciones nuevas. En cuanto a los resultados obtenidos en términos de medidas de calidad, los valores de soporte, confianza, *accuracy*, *lift* y *leverage* han sido bastante buenos mostrando la relevancia e importancia del grupo de genes encontrados a partir de las QAR obtenidas. Por tanto, podemos concluir que las reglas descubiertas han sido capaces de caracterizar correctamente los datos subyacentes y también de agrupar genes relevantes para el problema estudiado.

Los artículos asociados a esta parte son:

- (Sección 5.3) M. Martínez-Ballesteros, J.C. Riquelme. *Analisys of Measures of Quantitative Association Rules.* International Conference on Hybrid Artificial Intelligence Systems (HAIS 2011). Lecture Notes in Artificial Intelligence, Vol. 6679, No. 2, pp. 319-326, 2011. [Martínez-Ballesteros and Riquelme, 2011] [Ranking CORE C]

- M. Martínez-Ballesteros, I. Nepomuceno-Chamorro, J.C. Riquelme. *Inferring Gene-Gene Associations from Quantitative Association Rules.* IEEE International Conference on Intelligent Systems Design and Applications (ISDA'11), 2011. [Ranking CORE C]

Parte II

Publicaciones: Trabajos publicados, aceptados y sometidos

Capítulo 4

Trabajos de investigación seleccionados publicados en revistas indexadas

4.1. An evolutionary algorithm to discover quantitative association rules in multidimensional time series

Las publicaciones en revista asociadas a esta parte son:

- M. Martínez-Ballesteros, F. Martínez-Álvarez , A. Troncoso Lora, J.C. Riquelme Santos. *An Evolutionary Algorithm to Discover Quantitative Association Rules in Multidimensional Time Series.* Soft Computing (SOCO). Vol. 15, No. 10, pp. 2065-2084, October 2011 [Martínez-Ballesteros et al., 2011a].

    - Estado: Publicado
    - Índice de Impacto (JCR 2010): 1.512
    - Área de Conocimiento:
        - Computer Science, Artificial Intelligence. Ranking 47 / 108 - Q2
        - Computer Science, Interdisciplinary Applications. Ranking 41 / 97 - Q2

Soft Comput (2011) 15:2065–2084
DOI 10.1007/s00500-011-0705-4

ORIGINAL PAPER

An evolutionary algorithm to discover quantitative association rules in multidimensional time series

M. Martínez-Ballesteros · F. Martínez-Álvarez ·
A. Troncoso · J. C. Riquelme

Published online: 22 March 2011

Abstract An evolutionary approach for finding existing relationships among several variables of a multidimensional time series is presented in this work. The proposed model to discover these relationships is based on quantitative association rules. This algorithm, called QARGA (Quantitative Association Rules by Genetic Algorithm), uses a particular codification of the individuals that allows solving two basic problems. First, it does not perform a previous attribute discretization and, second, it is not necessary to set which variables belong to the antecedent or consequent. Therefore, it may discover all underlying dependencies among different variables. To evaluate the proposed algorithm three experiments have been carried out. As initial step, several public datasets have been analyzed with the purpose of comparing with other existing evolutionary approaches. Also, the algorithm has been applied to synthetic time series (where the relationships are known) to analyze its potential for discovering rules in time series. Finally, a real-world multidimensional time series composed by several climatological variables has been considered. All the results show a remarkable performance of QARGA.

M. Martínez-Ballesteros · J. C. Riquelme (✉)
Department of Computer Science,
University of Seville, Seville, Spain
e-mail: riquelme@us.es

M. Martínez-Ballesteros
e-mail: mariamartinez@us.es

F. Martínez-Álvarez · A. Troncoso
Department of Computer Science,
Pablo de Olavide University, Seville, Spain
e-mail: fmaralv@upo.es

A. Troncoso
e-mail: ali@upo.es

Keywords Time series · Quantitative association rules · Evolutionary algorithms · Data mining

1 Introduction

It is usual to find natural phenomena correlated to some other variables. Thus, real-world processes can be modeled by inferring knowledge from other associated variables that definitively have an effect on the original process. For instance, the existence of acid rain cannot be understood without the existence of other pollutant agents, such as monoxide carbon or sulfur dioxide. In other words, the knowledge of how some variables could affect other ones may be useful to obtain accurate behavior models.

Quantitative association rule (QAR) extraction in time series can be of the utmost usefulness for predictive purposes (Shidara et al. 2008; Wang et al. 2008). Thus, it could be interesting to find relationships among several time series to determine the range of values for a particular time series in a given time interval depending on the values of others for the same interval. For instance, rules such as $hour \in [10, 12] \wedge demand \in [12,000, 15,000] \Rightarrow price \in [3.2, 4.5]$ can provide useful knowledge for forecasting the electric energy price at peak hours (from 10 am to 12 pm) depending on the values of the energy demand during these hours. This information could help to obtain different models adjusted to different intervals or to develop a family of models for every rule. Hence, QAR are introduced in a new time series framework with the means of obtaining relationships among correlated time series that help to model their behavior.

Evolutionary algorithms (EA) have been extensively used for optimization and model adjustment in data mining tasks. In fact, the use metaheuristics in general, and of EA in particular, to deal with data mining-based problems is a hot topic of research nowadays (Alcalá-Fdez et al. 2009a, 2010; Chen et al. 2010; del Jesús et al. 2009; Yan et al. 2009). Also, EA have been used to build rule-based systems (Aguilar-Ruiz et al. 2007; Berlanga et al. 2010; Orriols-Puig and Bernadó-Mansilla 2009).

Real-coded genetic algorithms (RCGA) are very important within EA due to the increasing interest in solving real-world optimization problems. The main problem of RCGA, in which many researchers have focused their works, is the definition of adequate genetic operators (Herrera et al. 2004; Kalyanmoy et al. 2002). In particular, a new RCGA, henceforth called QARGA (Quantitative Association Rules by Genetic Algorithm) is proposed in this work. It is worth noting that QARGA does not perform previous variable discretization, that is, it handles numeric data during the whole rule extraction process, in contrast with many other approaches that perform data discretization to discover rules (Agrawal et al. 1993; Aumann and Lindell 2003; Vannucci and Colla 2004). Furthermore, the approach allows several degrees of freedom in specifying the user's preference regarding both of the number of attributes and structure of the rules. On the other hand, besides the well-known support and confidence measures, the accuracy of the rules is also obtained with a measure called lift due to its usefulness in the specific area of time series analysis (Ramaswamy et al. 1998).

First, QARGA has been applied to datasets from the Bilkent University Function Approximation (BUFA) repository (Guvenir and Uysal 2000). These datasets have been chosen because the literature offers multiple EA applied to them (Alatas and Akin 2006; Alatas et al. 2008; Mata et al. 2002). Later, time series have been synthetically generated to determine the suitability of applying QARGA to temporal data. Finally, multidimensional real-world time series have been used to extract QAR. In particular, climatological time series have been analyzed to discover the factors that cause high ozone concentration levels in atmosphere.

The remainder of the paper is divided as follows: Sect. 2 provides a formal description of QAR, as well as introduces the quality indices applied to QARGA. Section 3 presents the most relevant related works found in literature. Section 4 describes the main features of QARGA used in this work. The results of applying the proposed algorithm to different datasets are reported and discussed in Sect. 5. Finally, Sect. 6 summarizes the conclusions.

2 Preliminaries

This section is devoted to formally describe QAR and to introduce the quality measures used in this paper.

2.1 Quantitative association rules

Association rules (AR) were first defined by Agrawal et al. (1993) as follows. Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n items, and $D = \{tr_1, tr_2, \ldots, tr_N\}$ a set of N transactions, where each tr_j contains a subset of items. Thus, a rule can be defined as $X \Rightarrow Y$, where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. Finally, X and Y are called antecedent (or left side of the rule) and consequent (or right side of the rule), respectively.

When the domain is continuous, the association rules are known as QAR. In this context, let $F = \{F_1, \ldots, F_n\}$ be a set of features, with values in $\mathbb{R}$. Let A and C be two disjunct subsets of F, that is, $A \subset F$, $C \subset F$, and $A \cap C = \emptyset$. A QAR is a rule $X \Rightarrow Y$, in which features in A belong to the antecedent X, and features in C belong to the consequent Y, such that

$$X = \bigwedge_{F_i \in A} F_i \in [l_i, u_i] \tag{1}$$

$$Y = \bigwedge_{F_j \in C} F_j \in [l_j, u_j] \tag{2}$$

where l_i and l_j represent the lower limits of the intervals for F_i and F_j, respectively, and the couple u_i and u_j the upper ones. For instance, a QAR could be numerically expressed as

$$F_1 \in [12, 25] \wedge F_3 \in [5, 9] \Rightarrow F_2 \in [3, 7] \wedge F_5 \in [2, 8] \tag{3}$$

where F_1 and F_3 constitute the features appearing in the antecedent and F_2 and F_5 the ones in the consequent.

2.2 Quality parameters

This section provides a description of the support, confidence and lift indices (Brin et al. 1997) used to measure the interestingness of rules and of a new index, called *recovered*, to ensure that the full search space is explored.

The support of an itemset X is defined as the ratio of transactions in the dataset that contain X. Formally:

$$sup(X) = \frac{\#X}{N} = P(X) \tag{4}$$

where $\#X$ is the number of times that X appear in the dataset, and N the number of transactions forming such dataset. Other authors prefer naming the support of X simply as the probability of $X, P(X)$.

Let X and Y be the itemsets that identify the antecedent and consequent of a rule, respectively. The confidence of a rule is expressed as follows:

$$conf(X \Longrightarrow Y) = \frac{sup(X \Longrightarrow Y)}{sup(X)} \tag{5}$$

and it can be interpreted as the probability that transactions containing X, also contain Y. In other words, how certain is the rule subjected to analysis.

Finally, the interest or lift of a rule is defined as

$$lift(X \Longrightarrow Y) = \frac{sup(X \Longrightarrow Y)}{sup(X)sup(Y)} \qquad (6)$$

Lift means how many times more often X and Y are together in the dataset than expected, assuming that the presence of X and Y in transactions are occurrences statically independent. Lifts greater than one are desired because this fact would involve statistical dependence in simultaneous occurrence of X and Y and, therefore, the rule would provide valuable information about X and Y.

For a better understanding of such indices, a dataset comprising ten transactions and three features is shown in Table 1. Also consider an example rule

$$F_1 \in [180, 189] \wedge F_2 \in [85, 95] \Rightarrow F_3 \in [33, 36] \qquad (7)$$

In this case, the support of the antecedent is 20%, since two transactions, t_2 and t_9, simultaneously satisfy that F_1 and F_2 belong to the intervals [180, 189] and [85, 95], respectively (two transactions out of ten, $sup(X) = 0.2$). As for the support of the consequent, $sup(Y) = 0.2$ because only transactions t_6 and t_9 satisfy that $F_3 \in [33, 36]$. Regarding the confidence, only one transaction t_9 satisfies all the three features (F_1 and F_2 in the antecedent, and F_3 in the consequent) appearing in the rule; in other words, $sup(X \Rightarrow Y) = 0.1$. Therefore, $conf(X \Rightarrow Y) = 0.1/0.2 = 0.5$, that is, the rule has a confidence of 50%. Finally, the lift is $lift(X \Longrightarrow Y) = 0.1/(0.2 * 0.2) = 2.5$, since $sup(X \Rightarrow Y) = 0.1, sup(X) = 0.2$ and $sup(Y) = 0.2$, as discussed before.

Finally, the measure *recovered* is defined for finding rules covering different regions of the search space. An example e is covered by the rule r if the values of attributes of e belong to the intervals defined by the rule r. That is,

$$cov(e, r) = \begin{cases} 1 & \text{if } e \text{ is covered by r} \\ 0 & \text{otherwise} \end{cases} \qquad (8)$$

Given a set of rules $r_1, \ldots, r_n$ the measure *recovered* for the rule r_{n+1} is defined by

$$recov(r_{n+1}) = \frac{1}{N} \sum_{e \in D} \sum_{i=1}^{n} cov(e, r_i) \qquad (9)$$

Thus, this index provides a measure of the number of instances which have already been covered by a set of previous rules.

Table 1 Illustrative dataset

Transaction	F_1	F_2	F_3
t_1	178	75	24
t_2	186	93	37
t_3	167	60	22
t_4	199	112	30
t_5	154	47	42
t_6	173	83	33
t_7	177	91	63
t_8	159	53	48
t_9	183	88	35
t_{10}	178	93	58

3 Related work

A thorough review of recently published works reveals that the extraction of AR with numeric attributes is an emerging topic.

Mata et al. (2001) proposed a novel technique based on evolutionary techniques to find QAR that was improved in Mata et al. (2002). First, the approach found the sets of attributes which were frequently present in database and called *frequent itemsets*, and later AR were extracted from these sets.

Following with this topic, the optimization of the confidence—avoiding the initial threshold for the minimum support—was the main contribution of the work introduced in Yan et al. (2009). The authors used fitness function and non-threshold requirements for the minimum support. The fitness function is a key parameter in EA and the authors just used the relative confidence as fitness function.

Recently, Alcalá-Fdez et al. presented a study about three algorithms to analyze their effectiveness for mining QAR. In particular, EARMGA (Yan et al. 2009), GAR (Mata et al. 2002) and GENAR (Mata et al. 2001) were applied to two real-world datasets, showing their efficiency in terms of coverage and confidence.

On the other hand, data mining techniques for discovering AR in time series can be found in Bellazzi et al. (2005). The authors successfully mined temporal data retrieved from multiple hemodialysis sessions by applying preprocessing, data reduction and filtering as a previous step of the AR extraction process. Finally, AR were obtained by following the well-known Apriori itemset generation strategy (Venturini 1994).

An algorithm to discover frequent temporal patterns and temporal AR was introduced in Winarko and Roddick (2007). The algorithm extends the MEMISP algorithm (Lin and Lee 2002) which discovers sequential patterns by using a recursive find-then-index technique. Especially remarkable was the maximum gap time constraint included to remove insignificant patterns and consequently to reduce the number of temporal association rules.

Usually, the sliding window concept has been successfully applied to forecast time series (Martínez–Álvarez et al. 2011; Nikolaidou and Mitkas 2009). However, this concept has been recently used in (Khan et al. 2010) with the purpose of obtaining a low use of memory and low computational cost of the Apriori-based algorithm presented for discovering itemsets whose support increase over time.

In the non-supervised classification domain, the authors in Wan et al. (2007) made use of clustering processes to discretize the attributes of hydrological time series, as a first step of the rules extraction, which were eventually obtained by means of the Apriori algorithm. Following with clustering techniques, fuzzy clustering was used in Chen et al. (2010) to speed up the calculation for requirement satisfaction with multiple minimum supports, enhancing thus the results published in its initial work (Chen et al. 2009).

The work introduced in Huang et al. (2008) mined ocean data time series in order to discover relationship between salinity and temperature variations. Concretely, the authors discovered spatio-temporal patterns from the aforementioned variables and reported QAR using Prefix-Span and FITI algorithms (Pei et al. 2001; Tung et al. 2003).

Different models to forecast the ozone concentration levels have been recently proposed. Hence, the authors in Agirre-Basurko et al. (2006) developed two multilayer preceptron and a linear regression model for this purpose and prognosticated eight hours ahead for the Spanish city of Bilbao. They concluded that the insertion of extra seasonal variables may improve the general forecasting process. On the other hand, an artificial neural network model was presented in Elkamel et al. (2001). The authors also predicted the ozone concentrations by considering the analysis of additional climatological time series. Finally, temporal variations of the tropospheric ozone levels were analyzed in four sites of the Iberian Peninsula (Adame-Carnero et al. 2010) by means of statistical approaches.

Alternatively, the application of QAR can also be found in the data streams domain. In fact, the authors in Orriols-Puig et al. (2008) developed a model capable to classify on-line generated data for both continuous and discrete data streams.

MODENAR is a multi-objective pareto-based genetic algorithm that was presented in Alatas et al. (2008). In this approach, the fitness function aimed at optimizing four different variables: Support, confidence, comprehensibility of the rule and the amplitude of the intervals that constitutes the rule. A similar issue was addressed in Tong et al. (2005), in which the authors conducted research on the determination of existing conflicts when minimum support and minimum confidence are simultaneously required.

In Alatas and Akin (2008), the use of rough particle swarm techniques as an optimization metaheuristic was presented. In this work, the authors obtained the values for the intervals instead of frequent itemsets. Moreover, they proposed the use of some new operators such as rounding, repairing or filtrating.

Finally, QAR have also been used in the bioinformatics field. Thus, microarray data analysis by means of QAR was addressed in Georgii et al. (2005). The main novelty proposed by the authors was the definition of an AR as a linear combination of weighted variables, against a constant. Also in this context, the authors in Gupta et al. (2006) introduced a multi-step algorithm devoted to mine QAR for protein sequences. Once again, an Apriori-based methodology was used in Nam et al. (2009) to discover temporal associations from gene expression data.

4 Description of the search of rules

In a continuous domain, it is necessary to group certain sets of values that share same features and therefore it is required to express the membership of the values to each group. Adaptive intervals instead of fixed ranges have been chosen to represent the membership of such values in this work.

The search for the most appropriate intervals has been carried out by means of QARGA. Thus, the intervals are adjusted to find QAR with high values for support and confidence, together with other measures used in order to quantify the quality of the rule.

In the population, each individual constitutes a rule. These rules are then subjected to an evolutionary process, in which the mutation and crossover operators are applied and, at the end of the process, the individual that presents the best fitness is designated as the best rule. Moreover, the fitness function has been provided with a set of parameters so that the user can drive the process of search depending on the desired rules. The punishment of the covered instances allows the subsequent rules, found by QARGA, trying to cover those instances that were still uncovered, by means of an iterative rule learning (IRL) (Venturini 1993).

The following subsections detail the general scheme of the algorithm as well as the fitness function, the representation of the individuals, and the genetic operators.

4.1 Codification of the individuals

Each gene of an individual represents the upper and lower limit of the intervals of each attribute. The individuals are represented by an array of fixed length n, where n is the number of attributes belonging to the database. Furthermore, the elements are real-valued since the values of the attributes are continuous. Two structures are available for the representation of an individual:

- Upper structure. All the attributes included in the database are depicted in this structure. The limits of the intervals of each attribute are stored, where l_i is the inferior limit of the interval and u_i the superior one.
- Lower structure. Nevertheless, not all the attributes will be present in the rules that describe an individual. This structure indicates the type of each attribute, t_i, which can have three different values:
  - 0 when the attribute does not belong to the individual,
  - 1 when the attribute belongs to the antecedent, and
  - 2 when it belongs to the consequent.

Figure 1 shows the codification of a general individual of the population.

With the proposed codification, if an attribute is wanted to be retrieved for a specific rule, it can be done by modifying the value equal to 0 of the type by a value equal to 1 or 2. Analogously, an attribute that appears in a rule may stop belonging to such rule by changing the type of the attribute from values 1 or 2 to 0. An illustrative example is depicted in Fig. 2. In particular, the rule $X_1 \in [20, 34] \wedge X_3 \in [7, 18] \Rightarrow X_4 \in [12, 27]$ is represented. Note that attributes X_1 and X_3 appear in the antecedent, X_4 in the consequent, and X_2 is not involved in the rule. Therefore, $t_1 = t_3 = 1, t_2 = 0$ and $t_4 = 2$.

4.2 Generation of the initial population

The individuals of the initial population are randomly generated. In other words, the number of attributes appearing in the rule and the type and interval for each attribute are randomly generated. To assure that the individuals represent sound rules when the genes are generated, the following constraints are considered:

- Limits of the interval:
  - The lower limit of the interval has to be less than the upper limit of the interval. If the randomly generated values do not fulfill this requirement, the limits are swapped.
  - The lower and upper limits of the interval have to be greater and less than the lower and upper limits of the domain of the attribute, respectively. Otherwise, the corresponding limits of the domain of the attribute are assigned.
- Type of the attribute:
  - The number of attributes of the rule has to be greater than a minimum number of attributes defined by the user depending on the desired rule.
  - The number of attributes belonging to the antecedent of the rule has to be greater than 1.
  - The number of attributes belonging to the consequent of the rule has to be greater than 1, and less than a maximum number of allowed consequents, which is a parameter defined by the user depending for the desired rule.

l_1 u_1	l_2 u_2	...	l_n u_n
t_1	t_2	...	t_n

Fig. 1 Representation of an individual of the population

X_1	X_2	X_3	X_4
20 34	1 5	7 18	12 27
1	0	1	2

Fig. 2 Example of an individual

4.3 Genetic operators

This section describes the genetic operators used in the proposed algorithm, that is, selection, crossover and mutation operators.

1. *Selection*. An elitist strategy is used to replicate the individual with the best fitness. By contrast, a roulette selection method is used for the remaining individuals rewarding the best individuals according to the fitness. Note that the tournament selection was also used in preliminary studies, showing similar performance to that of roulette selection.
2. *Crossover*. Two parent individuals x and y, chosen by means of the roulette selection, are combined to generate a new individual z. Formally, let $[l_i^x, u_i^x], [l_i^y, u_i^y]$ and $[l_i^z, u_i^z]$ be the intervals in which the attribute a_i vary for the individuals x, y and z, respectively, and let t_i^x, t_i^y and t_i^z be the type of the attribute a_i for the individuals x, y and z, respectively. Then, for each attribute a_i two cases can occur:
   - $t_i^x = t_i^y$: The same type is assigned to the descendent and the interval is obtained by generating two random numbers among the limits of the intervals of both parents, as shown in Eqs. 10 and 11.

$$t_i^z = t_i^x \tag{10}$$

$$[l_i^z, u_i^z] = [random(l_i^x, l_i^y), random(u_i^x, u_i^y)] \tag{11}$$

- $t_i^x \neq t_i^y$: One of the two types is randomly chosen between both of the parents, without modifying the intervals of such attribute, as shown in Eqs. 12 and 13.

$$[l_i^z, u_i^z] = [l_i^x, u_i^x], \quad \text{if } t_i^z = t_i^x \tag{12}$$

$$[l_i^z, u_i^z] = [l_i^y, u_i^y], \quad \text{if } t_i^z = t_i^y \tag{13}$$

The limits and types of the attributes of the offspring are checked, as described in Sect. 4.2, to assure that it represents sound rules. If any attribute does not fulfill the required constraints regarding the type of attributes the individual is discarded and a new individual is obtained from the same parents. The crossover process is depicted in Fig. 3.

3. *Mutation*. The mutation process consists in modifying according to a probability the genes of randomly selected individuals. The mutation of a gene can be focused on

   - *Type of the attribute*. Two equally probable cases can be distinguished:
     – Null Mutation. The type t_i of the selected attribute is different to null, and eventually changed to null.
     – Not Null Mutation. The type t_i of the selected attribute is null and changed to antecedent or consequent.
   - *Intervals of the attribute*. Three equally probable cases are possible:
     – *Lower Limit*. A random value is added or subtracted to the lower limit of the interval.
     – *Upper Limit*. A random value is added or subtracted to the upper limit of the interval.
     – *Both Limits*. A random value is added or subtracted to both limits of the interval.
     For all the three cases, the random value is generated between 0 and a percentage (usually 10%) of the amplitude of the interval and it will be added or subtracted according to a certain probability.

The choice between the mutation of the type or the mutation of the interval depends on a given probability.

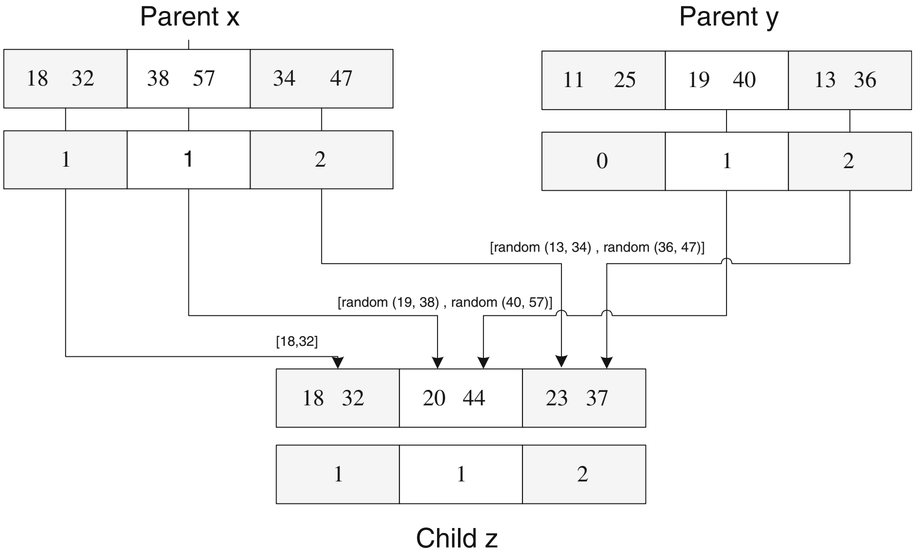
The limits and types of the attributes of the offspring are checked, as described in Sect. 4.2, to assure that it represents sound rules. If any attribute does not fulfill the required constraints regarding of the type of attributes, the individual is discarded and a new mutation is obtained from the same original individual.

Some examples of all the kind of mutations are illustrated in Figs. 4, 5, 6, 7 and 8.

4.4 The fitness function

The fitness of each individual allows deciding which are the best candidates to remain in subsequent generations. In order to make this decision, it is desirable that the support would be high, since this fact implies that more samples from the database are covered. Nevertheless, to take into consideration only the support is not enough to calculate the fitness because the algorithm would try to enlarge the amplitude of the intervals until the whole domain of each attribute would be completed. For this reason, it is

Fig. 3 Crossover for the individuals *x* and *y*



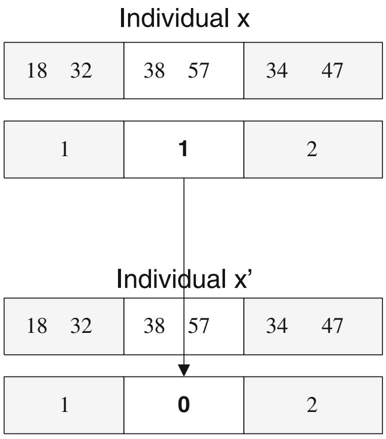


Fig. 4 Scheme of Null Mutation

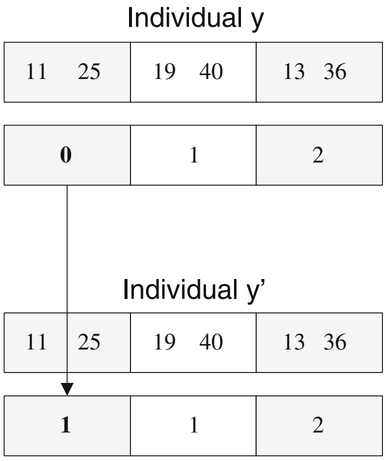


Fig. 5 Scheme of Not Null Mutation

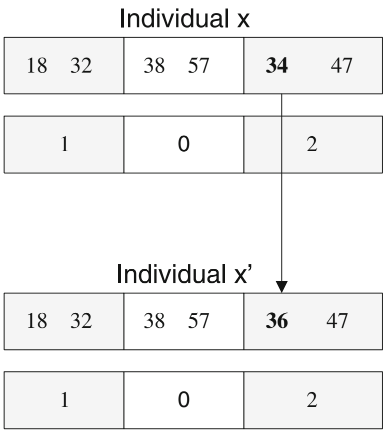


Fig. 6 Scheme of Lower Limit Mutation

necessary to include a measure to limit the growth of the intervals during the evolutionary process. The chosen fitness function to be maximized is

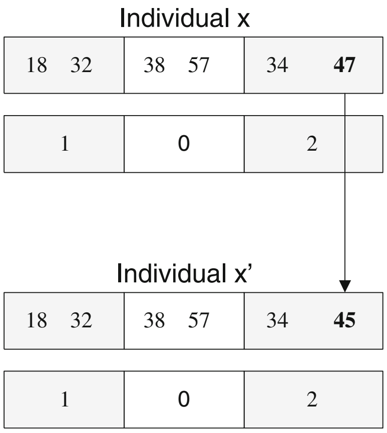


Fig. 7 Scheme of Upper Limit Mutation

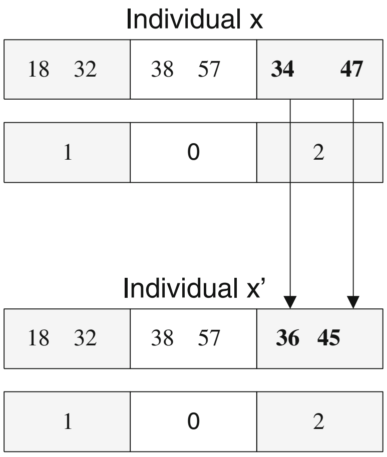


Fig. 8 Scheme of Lower and Upper Limits Mutation

$$f = w_s \cdot sup + w_c \cdot conf - w_r \cdot recov + w_n \cdot nAttrib - w_a \cdot ampl \tag{14}$$

where *sup*, *conf* and *recov* are defined in Sect. 2.2, *nAttrib* is the number of attributes appearing in the rule, *ampl* is the average size of intervals of the attributes that compose the rule, and w_s, w_c, w_r, w_n and w_a are the weights to drive the search, and will vary depending on the required rules.

The support rewards the rules fulfilled by many instances and the weight w_s can increase or decrease its effect.

The confidence together with the support are the most widely measures used to evaluate the quality of the QAR. The confidence is the grade of reliability of the rule. High values of w_c may be used when rules without error are desired, and viceversa.

The number of recovered instances is used to indicate that a sample has already been covered by a previous rule. Thus, rules covering different regions of the search space are preferred. The process of punishing the covered instances is now described. Every time the evolutionary process ends and the best individual is chosen as the best rule, the database is processed in order to find those instances already covered by the rule. Hence, each instance has a counter that increases its value by one every time a rule covers it.

The number of attributes of a rule can be adjusted by means of the weight w_n. Thus, when w_n is set to a value close to 0, few attributes are obtained and, on the other hand, when w_n is set to a value close to 1, many attributes appear in rules.

The amplitude controls the size of the intervals of the attributes that compose the rules and those individuals with large intervals are penalized by means of the factor w_a, which allows the rules be more or less permissive regarding the amplitude of the intervals.

Hence, the user can model the behavior of the rules that can be obtained by varying the weights in the fitness function. Therefore, the user can obtain rules according to their needs without a previous data discretization.

4.5 The IRL approach

The proposed algorithm is based on the iterative rule learning (IRL) process, whose general scheme is shown in Fig. 9.

The EA is applied in each iteration obtaining one rule per iteration, which is precisely the best individual discovered. While the number of desired rules is not reached, IRL allows penalization in already covered instances, with the aim of finding rules that cover those instances that have not been covered yet in subsequent iterations. The main advantage of the approach is that attempts at covering every region in the solutions domain, that is, the set of rules will cover all the consequent domain. The iterative process ends when it finds the desired number of rules.

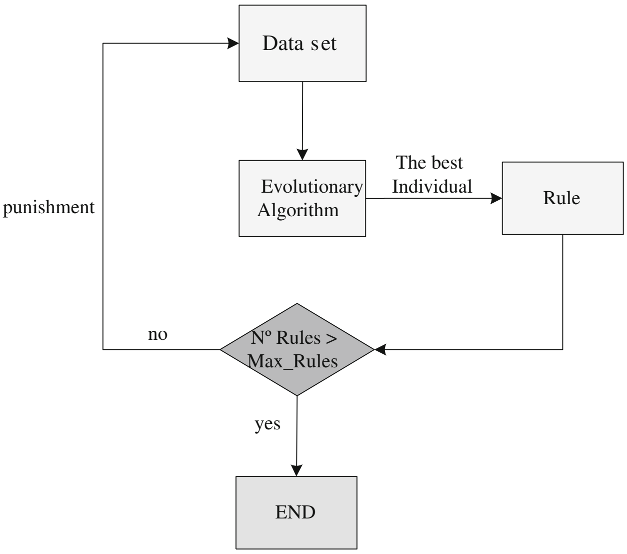


Fig. 9 Scheme of IRL

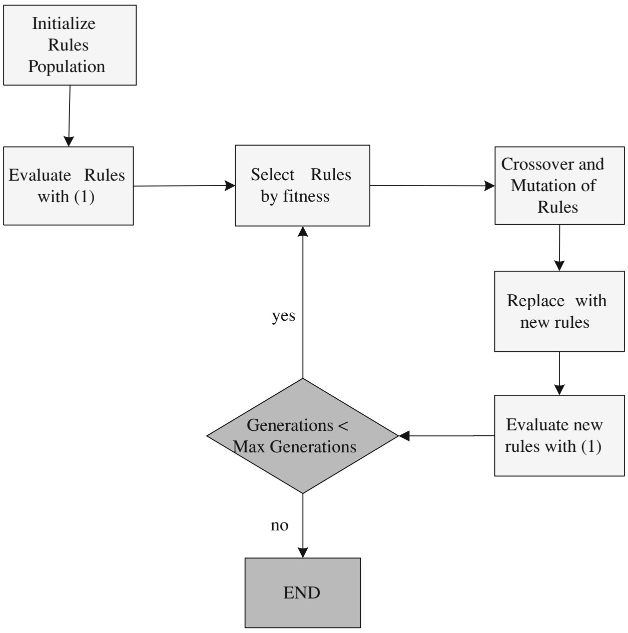

Initialize Rules Population
Evaluate Rules with (1)
Select Rules by fitness
Crossover and Mutation of Rules
Replace with new rules
yes
Generations < Max Generations
Evaluate new rules with (1)
no
END

Fig. 10 EA scheme

Figure 10 provides the scheme of the EA, which is the main step of the IRL process depicted in Fig. 9.

First, the rules population is initialized and evaluated. All rules are evaluated according to (14). Thus, in each iteration the selection operator is applied to select the best rules on the basis of the fitness function. Then, the crossover operator is applied to the selected rules while the population size is not completed. Individuals are randomly selected in order to apply the mutation operator. Finally, the new population is again evaluated by the fitness function and the evolutionary process restarts. Note that the process will be repeated as many times as the maximum number of preset generations indicates.

5 Results

In this section the results obtained from the application of the proposed approach to different datasets are presented. First, Sect. 5.1 provides a detailed description of all used datasets. A summary of the key parameters configuration used for all the algorithms can be found in Sect. 5.2. Finally, the results are gathered and discussed in Sect. 5.3.

The approach has been initially tested on several widely studied datasets from the public BUFA repository, and the accuracy of QARGA has been compared with that of the algorithms introduced in Yan et al. (2009) and Mata et al. (2001) in Sect. 5.3.1. On the other hand, two different kind of time series are analyzed: synthetically generated and real-world multidimensional temporal data. Sect. 5.3.2 is devoted to evaluate the accuracy of QARGA when it is applied to synthetically generated multidimensional time series. Likewise, the real-world case is reported in Sect. 5.3.3, where QAR are obtained to discover relationships between the tropospheric ozone and other climatological time series.

5.1 Dataset description

This section presents the number of records and number of attributes of the BUFA repository datasets as well as how synthetic time series were generated and what real-world time series consist of.

5.1.1 Public datasets

QARGA has been applied to 15 public datasets: Basketball, Bodyfat, Bolts, Kinematics, Longley, Normal Body Temperature, Plastic, Pollution, Pw Linear, Pyramidines, Quake, Schools, Sleep, Stock Price and Vineyard, which can be found at BUFA repository (Guvenir and Uysal 2000). Relevant information about these datasets is summarized in Table 2.

5.1.2 Synthetic multidimensional time series

In this section two different synthetically generated multidimensional time series are described: time series without and with disjunctions, respectively. In particular, multidimensional time series are generated, that is, time series characterized by more than one variable in each time stamp. Or, in other words, two or more time series simultaneously observed that characterize the same phenomenon. Formally, a multidimensional time series *MTS* can be expressed as $MTS = [X_1(t), \ldots, X_n(t)]^T$, where each $X_i(t)$ is a variable measured along with the time, t, and n is the number of inter-related time series that identifies the whole *MTS*. Thus, the goal of applying QAR to *MTS* is to discover existing relationships among those X_i forming the *MTS*, along with the time.

Table 2 Public datasets.

Dataset	Records	Attributes
Basketball	96	5
Bodyfat	252	18
Bolts	40	8
Kinematics	8192	9
Longley	16	7
Normal Body Temperature	130	3
Plastic	1650	3
Pw Linear	200	11
Pollution	60	16
Pyramidines	74	28
Quake	2178	4
School	62	20
Sleep	57	8
Stock price	950	10
Vineyard	52	4

Regarding the time series with no disjunctions, Table 3 defines a three-dimensional time series, $n = 3$, in which three variables X_1, X_2 and X_3 share static relationships in fixed intervals of time.

Thus, 100 values for each variable X_1, X_2 and X_3 were generated and uniformly distributed in four intervals. To obtain these series, values for variables X_1, X_2 and X_3 were randomly selected for every t_i according to constraints listed in Table 3, where t_i varies from 1 to 100. Finally, the resulting time series are depicted in Fig. 11.

With reference to time series with disjunctions, a bi-dimensional time series represented by two variables, X_1 and X_2, has been generated with, again, 100 values for each time series uniformly distributed in four intervals. However, the main difference regarding the previous situation lies in the fact that now the variables X_1 and X_2 can be defined by more than one possible set of values.

Table 4 shows the constraints considered to generate the time series with disjunctions. The series is generated then as follows: For every t_i and X_j one interval is randomly chosen and, then, a value is randomly chosen from the interval previously selected. For instance, X_1 can indistinctively belong to intervals $[10, 20]$ or $[15, 35]$ when $t \in [26, 50]$ in set #1.

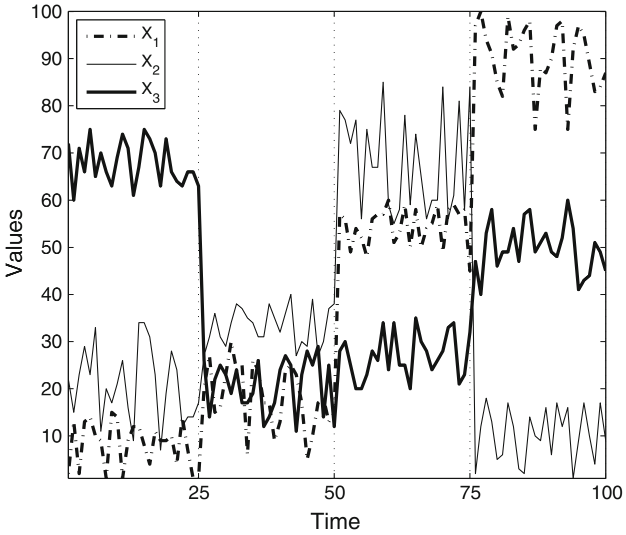


Fig. 11 Time series with no disjunctions

Table 3 Time series with no disjunctions

ID	Sets	Sup. (%)
#0	$t \in [1,25] \Longrightarrow X_1 \in [1, 15] \wedge X_2 \in [7, 35] \wedge X_3 \in [60, 75]$	25.0
#1	$t \in [26,50] \Longrightarrow X_1 \in [5, 30] \wedge X_2 \in [25, 40] \wedge X_3 \in [10, 30]$	25.0
#2	$t \in [51,75] \Longrightarrow X_1 \in [45, 60] \wedge X_2 \in [55, 85] \wedge X_3 \in [20, 35]$	25.0
#3	$t \in [76,100] \Longrightarrow X_1 \in [75, 100] \wedge X_2 \in [0, 20] \wedge X_3 \in [40, 60]$	25.0

Table 4 Time series with disjunctions

ID	Sets
#0	$t \in [1, 25] \Longrightarrow X_1 \in [20, 30] \wedge X_2 \in [50, 80]$
#1	$t \in [26, 50] \Longrightarrow (X_1 \in [10, 20] \vee X_1 \in [15, 35]) \wedge X_2 \in [40, 60]$
#2	$t \in [51, 75] \Longrightarrow X_1 \in [1, 15] \wedge (X_2 \in [40, 50] \vee X_2 \in [60, 70])$
#3	$t \in [76, 100] \Longrightarrow (X_1 \in [15, 25] \vee X_1 \in [30, 40]) \wedge (X_2 \in [30, 45] \vee X_2 \in [40,50])$

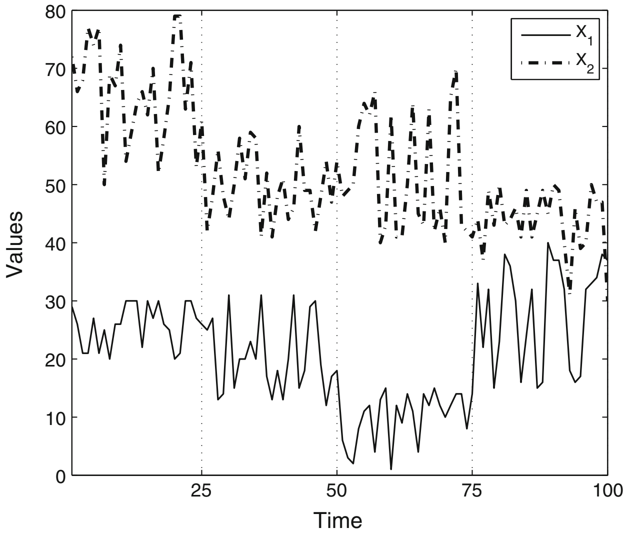


Fig. 12 Time series with disjunctions

Table 5 Expected rules from time series with disjunctions

ID	Sets	Sup. (%)
#0	$t \in [1, 25] \wedge X_1 \in [20, 30] \wedge X_2 \in [50, 80]$	25.0
$\#1_1$	$t \in [26, 50] \wedge X_1 \in [10, 20] \wedge X_2 \in [40, 60]$	12.5
$\#1_2$	$t \in [26, 50] \wedge X_1 \in [15, 35] \wedge X_2 \in [40, 60]$	12.5
$\#2_1$	$t \in [51, 75] \wedge X_1 \in [1, 15] \wedge X_2 \in [40, 50]$	12.5
$\#2_2$	$t \in [51, 75] \wedge X_1 \in [1, 15] \wedge X_2 \in [60, 70]$	12.5
$\#3_1$	$t \in [76, 100] \wedge X_1 \in [15, 25] \wedge X_2 \in [30, 45]$	6.25
$\#3_2$	$t \in [76, 100] \wedge X_1 \in [15, 25] \wedge X_2 \in [40, 50]$	6.25
$\#3_3$	$t \in [76, 100] \wedge X_1 \in [30, 40] \wedge X_2 \in [30, 45]$	6.25
$\#3_4$	$t \in [76, 100] \wedge X_1 \in [30, 40] \wedge X_2 \in [40, 50]$	6.25

The resulting *MTS* according with Table 4 is illustrated in Fig. 12. As it can be observed, relationships between time and variables are considerably more difficult to be mined. In fact, the expected rules are listed in Table 5, where every disjunction in the interval for each variable involve two possible conjunctions. For instance, the set #0 in Table 4 would solely generate rule #0 from Table 5. Nevertheless, set #1 would diverge into two different expect rules, $\#1_1$ and $\#1_2$ from Table 5, and so on. Note that the support in Table 5 for all possible rules is, actually, the expected support assuming that the portion of values of a variable for every disjunction were equal.

5.1.3 *Real-world time series application: ozone concentration*

The proposed algorithm has also been applied in order to discover QAR in real-world multidimensional time series. Specifically, QAR are intended to be found among climatological time series such as temperature, humidity, direction and speed of the wind, several temporal variables such as the hour of the day and the day of the week and, finally, the tropospheric ozone. These variables have influence on the ozone concentration in the atmosphere which is the target agent.

All variables have been retrieved from the meteorological station of the city of Seville in Spain for the months from July to August during years 2003 and 2004, generating a dataset with 1488 instances. The reason for selecting such periods is because during these periods the highest concentration of ozone was reported.

For predictive purposes, the climatological time series have been forced to belong to the antecedent and the ozone to the consequent. As a result, a prediction of the ozone is achieved on the basis of the rules extracted from these variables.

5.2 Parameters configuration

In this section, the values for the parameters of each method analyzed in Sect. 5.3 are described. This section is divided in subsections for all used datasets. It is noteworthy that the parameters of every method with which QARGA is compared were obtained from the original papers.

5.2.1 *Configuration for public datasets*

1. *EARMGA*. This algorithm (Yan et al. 2009) was executed five times and the average values of such executions were presented. The main parameters of EARMGA algorithm are 100 for the number of the rules, 100 for the size of the population and 100 for the number of generations. EARMGA use 0.0 for the minimum support and minimum confidence; 0.75 for the probability of selection; 0.7 for the probability of crossover, 0.1 for the probability of mutation; 0.01 for difference boundary and 4 for the number of partitions for numeric attributes.
2. *GENAR*. This algorithm was executed five times and the average values of such executions were presented. The main parameters of GENAR algorithm (Mata et al. 2001) are 100 for the number of the rules, 100 for the size of population and 100 for the number of generations. GENAR use 0.0 for the minimum support and minimum confidence; 0.25 for the probability of selection; 0.7 for the probability of crossover and 0.1 for the probability of mutation; 0.7 for the penalization factor and 2 for the amplitude factor.
3. *QARGA*. It has been executed five times, and the average results are also shown for this case. The main parameters of QARGA are 100 for the number of the rules, 100 for the size of the population, 100 for the number of generations, 0.0 for the minimum support and minimum confidence and 0.8 for the probability of mutation.

5.2.2 *Configuration for synthetic time series with no disjunctions*

1. *QARGA*. The proposed algorithm has been executed five times. The main parameters are as follows: 100 for the size of the population, 100 for the number of generations, 20 for the number of rules to be obtained, and 0.8 for the mutation probability. After an experimental study to assess the influence of the weights on the rules to be obtained, the weights chosen for the fitness function were 3 for w_s, 1 for w_c, 2 for w_r, 0.2 for w_n and 0.5 for w_a.

5.2.3 *Configuration for synthetic time series with disjunctions*

1. *QARGA*. The main parameters for these time series are exactly the same that those used to generate rules for synthetic time series with no disjunctions. That is, 100 for the size of the population, 100 for the number of generations, 20 for the number of rules to be obtained, and 0.8 for the mutation probability. In this case, the weights chosen for the fitness function were 1.5 for w_s, 0.5 for w_c, 0.2 for w_r, 0.2 for w_n and 0.3 for w_a.

5.2.4 *Configuration for ozone time series*

For this time series, QARGA has been compared with Apriori and due to its previous required discretization, two different kind of experimentation are distinguished. First, all the continuous variables have been discretized with three intervals. Thus, the obtained rules by Apriori present high amplitudes and therefore high supports. Second, all the real-valued attributes have been discretized in ten intervals, which involves rules with small amplitudes and low supports. For both of the experimentations, the selected rules are the ones that presented greater confidence.

For the first kind of experimentation, the main parameters of QARGA have been set as follows: 100 for the size of the population, 100 for the number of generations, 20 for the number of rules to be obtained and 0.8 for the mutation probability. After an empirical study to test the influence of the weights on the rules to be obtained, the weights of the fitness function, 3 for w_s, 0.2 for w_c, 0.2 for w_r, 0.3 for w_n and 0.2 for w_a have been chosen. This study consisted in determining the values of the weights for which the confidence of the rules was maximized. Note that w_s is high compared to the other ones because rules with high support are desired, making thus possible the comparison with the rules obtained by Apriori.

For the second kind of experimentation, the main parameters of QARGA have been set as follows: 100 for the size of the population, 100 for the number of generations, 20 for the number of rules to be obtained and 0.8 for the mutation probability. After an empirical study to test the influence of the weights on the rules to be obtained, the weights of the fitness function, 1 for w_s, 0.2 for w_c, 0.2 for w_r, 1 for w_n and 0.2 for w_a have been chosen. Analogous to the first experimentation, the weight associated with support is low to make possible the comparison with the Apriori algorithm.

5.3 Analysis of results

This section discusses all the results obtained from the application of QARGA to the selected datasets introduced in previous sections.

5.3.1 *Results in public datasets*

To carry out the experimentation and make a comparison with QARGA, the evolutionary algorithms EARMGA

(Yan et al. 2009) and GENAR (Mata et al. 2001), available in the KEEL tool (Alcalá-Fdez et al. 2009b), have been chosen.

Table 6 shows the results obtained by EARMGA, GENAR and QARGA for every dataset. The column *Number of rules* indicates the average of the number of the rules found by each algorithm after executions specified in Sect. 5.2. The percentage of the records covered by the rules for these datasets is shown in column *Records*.

It can be noticed that the average of the number of rules and the average of the percentage of the records covered by the rules founds by QARGA are greater than the rest of the algorithms.

Table 6 Number of rules and percentages of records covered by the mined rules obtained by QARGA and all other algorithms

Dataset	Number of rules			Records (%)		
	EARMGA	GENAR	QARGA	EARMGA	GENAR	QARGA
Basketball	100	100	100	74.16	91.04	99.48
Bodyfat	100	100	100	46.67	69.44	100
Bolts	100	100	100	57.50	51.00	100
Kinematics	100	100	100	42.33	38.84	89.17
Longley	100	11.70	100	50.00	100	100
Normal Body Temperature	100	100	100	100	97.08	98.00
Plastic	96.40	100	100	100	99.44	99.67
Pw Linear	100	100	100	53.00	21.40	98.00
Pollution	100	100	100	42.67	52.33	96.67
Pyramidines	100	82.15	100	43.78	100	100
Quake	100	100	100	97.41	82.12	91.85
School	100	100	100	55.08	85.57	100
Sleep	100	100	100	67.84	88.24	100
Stock price	100	100	100	59.37	87.98	100
Vineyar	100	100	100	93.85	94.62	100
	99.76	92.92	100	65.58	77.27	98.19

Table 7 Quality measurements of rules obtained by QARGA and all other algorithms.

Dataset	Support (%)			Confidence (%)		
	EARMGA	GENAR	QARGA	EARMGA	GENAR	QARGA
Basketball	2.70	30.82	33.52	100	96.52	97.43
Bodyfat	4.97	41.52	74.06	100	96.53	98.17
Bolts	11.43	14.43	11.17	100	100	99.76
Kinematics	2.09	0.53	22.02	100	96.50	83.89
Longley	12.69	24.37	36.68	100	100	100
Normal Body Temperature	22.29	64.27	7.70	100	72.89	97.55
Plastic	10.06	24.43	8.51	97.76	55.17	98.56
Pw Linear	3.68	1.09	16.47	100	100	98.50
Pollution	5.36	22.64	49.83	100	99.72	99.90
Pyramidines	7.84	2.19	11.34	38.49	100	99.82
Quake	3.40	35.17	7.73	100	64.40	94.66
School	7.21	8.13	41.73	100	100	99.23
Sleep	9.05	36.69	49.71	100	71.17	99.68
Stock price	4.05	30.71	32.71	100	91.49	98.93
Vineyar	7.80	43.75	39.35	100	98.60	99.31
	7.64	25.38	29.50	95.75	89.53	97.69

Table 7 shows some quality measurements of rules obtained by every algorithm. The first column, *support* (%), reports the average support obtained, that is, the percentage of covered instances. The next column, *confidence* (%), shows the average confidence obtained by every algorithm.

Concentrating on the results themselves, it can be appreciated that the average support found by QARGA is greater than that found by the other algorithms in almost all the datasets. The average confidence obtained by QARGA is greater than that of EARMGA and GENAR, that is, QARGA provides more reliable rules with smaller errors.

Table 8 and Table 9 show the average number of attributes and the average amplitude for both the antecedent and the consequent for the rules extracted by EARMGA, GENAR and QARGA. From its observation it can be

Table 8 Amplitudes for the antecedents, consequents and rules obtained by QARGA and all other algorithms

Dataset	Antecedent amplitude (%)			Consequent amplitude (%)			Rule amplitude (%)		
	EARMGA	GENAR	QARGA	EARMGA	GENAR	QARGA	EARMGA	GENAR	QARGA
Basketball	40.37	49.55	26.84	69.71	49.85	32.35	46.55	49.61	29.22
Bodyfat	39.19	47.20	28.80	81.57	49.99	26.86	48.65	47.35	27.80
Bolts	41.75	38.49	11.69	67.67	31.85	9.49	48.67	37.45	10.57
Kinematics	39.23	49.80	29.54	100	49.61	30.65	51.80	49.78	30.03
Longley	44.85	42.31	26.85	57.20	43.90	22.09	48.70	42.53	24.44
Normal Body Temperature	45.23	74.90	12.60	89.90	50	16.21	60.13	66.60	13.92
Plastic	32.06	49.17	15.89	81.96	46.90	25.78	48.70	48.41	19.43
Pw Linear	32.63	47.84	18.81	89.40	41	13.57	44.36	47.22	15.51
Pollution	40	45.73	14.30	53.38	49.82	13.90	43.18	45.98	14.15
Pyramidines	35.32	32.16	8.09	46.29	45.60	8	100	32.56	8.05
Quake	35.59	49.90	10.75	94.55	48.63	11.28	50.59	49.65	10.88
School	41.36	40.92	20.77	80.90	46.13	19.86	50.28	41.18	20.35
Sleep	43.84	41.92	10.51	79.13	49.89	10.33	51.74	42.91	10.41
Stock price	38.45	48.07	29.89	96.18	48.56	30.21	50.39	48.12	30.06
Vineyar	39.57	48.97	29.86	72.30	45.43	29.87	47.88	48.08	29.82
	39.30	47.13	19.68	77.34	46.48	20.03	52.77	46.50	19.64

Table 9 Size of the antecedents, consequents and rules obtained by QARGA and all other algorithms

Dataset	Antecedent size			Consequent size			Rule size		
	EARMGA	GENAR	QARGA	EARMGA	GENAR	QARGA	EARMGA	GENAR	QARGA
Basketball	3.96	4	1.37	1.04	1	1.01	5	5	2.38
Bodyfat	3.83	17	1.26	1.17	1	1.05	5	18	2.31
Bolts	3.60	7	4.31	1.40	1	2.15	5	8	6.46
Kinematics	3.97	8	1.96	1.03	1	1.03	5	9	2.99
Longley	3.58	6	1.08	1.42	1	1.08	5	7	2.15
Normal Body Temperature	2.00	2	1.87	1.00	1	1.04	3	3	2.92
Plastic	1.99	2	1.86	1.01	1	1.00	3	3	2.86
Pw Linear	3.97	10	1.65	1.03	1	1.01	5	11	2.65
Pollution	3.81	15	6.13	1.19	1	1.42	5	16	7.54
Pyramidines	3.65	27	12.51	1.35	1	2.62	5	28	15.13
Quake	2.97	3	2.79	1.03	1	1.02	4	4	3.81
School	3.90	19	1.37	1.10	1	1.13	5	20	2.50
Sleep	3.72	7	1.50	1.28	1	1.38	5	8	2.88
Stock price	3.96	9	1.48	1.04	1	1.01	5	10	2.49
Vineyar	2.98	3	1.09	1.02	1	1.03	4	4	2.11
	3.46	9.27	2.82	1.14	1.00	1.26	4.60	10.27	4.08

concluded that QARGA mined rules with short antecedent and short consequent, which helps to the comprehensiveness of the rules. The number of attributes per rule obtained by QARGA is similar to that of EARMGA and does not present relevant differences. For the case of the amplitude, QARGA obtained amplitudes smaller than EARMGA and GENAR in all datasets.

In short, QARGA presents greater average support, less number of attributes and smaller amplitudes than the other ones, which leads to the conclusion that QARGA obtained better rules in general terms.

From the reported results, it can be seen that rules with high support and confidence as well as moderate amplitude of intervals with small number of attributes have been found. In terms of support, confidence and amplitude, QARGA outperforms EARMGA and GENAR which leads to the obtention of more precise as well as comprehensible rules, since the number of attributes that appear in both antecedent and consequent is small, helping the user to easily understand them.

Last, a statistical analysis has been conducted to evaluate the significance of QARGA, following the non-parametric procedures discussed in García et al. (2009). For this purpose, the lift obtained from the application of QARGA, EARMGA and GENAR to the 15 datasets has been calculated, and it is shown in Table 10. From this table, it can be noticed that the algorithm QARGA reaches the highest rank in 15 datasets, EARMGA reaches the second and third positions in 10 and 5 datasets, respectively, and finally, GENAR obtains the second and third positions in 5 and 10 datasets. The average ranking for each algorithm is summarized in Table 11. It can be observed that the lowest value of average ranking is obtained by QARGA which is, therefore, the control algorithm.

Table 10 Lift of the mined rules by QARGA and all other algorithms

Dataset	Lift		
	EARMGA	GENAR	QARGA
Basketball	1.34	1.09	2.01
Bodyfat	1.77	1.11	4.49
Bolts	1.73	1.60	9.10
Kinematics	1.00	1.38	5.77
Longley	2.79	2.46	2.82
Normal Body Temperature	1.02	0.99	2.78
Plastic	1.26	1.10	3.44
Pw Linear	1.00	1.57	2.04
Pollution	3.00	1.23	6.86
Pyramidines	2.23	2.24	6.78
Quake	1.01	1.00	2.03
School	1.24	1.62	3.50
Sleep	1.93	1.15	9.48
Stock price	1.08	1.64	2.49
Vineyar	1.29	1.27	2.52
	1.58	1.43	4.41

Table 11 Average rankings of the algorithms.

Algorithm	Ranking
GENAR	2.66
EMARGA	2.33
QARGA	1.00

Friedman and Iman-Davenport (ID) tests have been applied to assess if there are global differences in the lifts obtained for three algorithms. The results obtained by both tests for the level of significance $\alpha = 0.05$ are summarized in Table 12. Note that the values in columns *Value in* χ^2 and *Value in* F_F have been retrieved from Tables A4 and A10 in Sheskin (2006), respectively. As the p values obtained from both of the tests are lower than the level of significance considered, it can be stated that there exist significant differences among the results obtained by three algorithms and a post-hoc statistical analysis is required.

The Holm and Hochberg tests have been applied to compare separately QARGA to GENAR and EMARGA. Table 13 shows the sorted p values obtained by GENAR and EMARGA for two levels of significance ($\alpha = 0.05$ and $\alpha = 0.10$). Both of the tests allow concluding that QARGA is better than EMARGA and GENAR for both levels of significance, as the two tests reject all hypotheses.

In addition, it is interesting to discover the precise p value for which each hypothesis can be rejected. These exact values are called adjusted p values and how to obtain them is thoroughly described in Wright (1992). Table 14

Table 12 Results of the Friedman and Iman-Davenport tests with $\alpha = 0.05$

Lift					
Friedman	Value in χ^2	p	ID	Value in F_F	p
23.33	5.99	8.57×10^{-6}	48.99	3.34	7.16×10^{-10}

Table 13 Holm and Hochberg tests results with QARGA as control algorithm

i	Algorithm	z	p	α/i ($\alpha = 0.05$)	α/i ($\alpha = 0.10$)
2	GENAR	4.56	5.01×10^{-6}	0.025	0.05
1	EMARGA	3.65	2.61×10^{-4}	0.05	0.10

Table 14 Adjusted p values when QARGA is compared to the remaining algorithms

Algorithm	Unadjusted p	p_{BD}	p_{Holm}	p_{Hoch}
GENAR	5.01×10^{-6}	1.00×10^{-5}	1.00×10^{-5}	1.00×10^{-5}
EMARGA	2.61×10^{-4}	5.21×10^{-4}	2.61×10^{-4}	2.61×10^{-4}

shows the adjusted p values for Bonferroni-Dunn (BD), Holm and Hochberg tests. It can be appreciated that the Holm and Hochberg tests show that QARGA is significantly better than the others with the lowest confidence level compared to the remaining tests ($\alpha = 2.61 \times 10^{-4}$). Again, the three tests coincide in rejecting all hypotheses for levels of significance $\alpha = 0.05$ and $\alpha = 0.10$, determining that QARGA is the best algorithm.

5.3.2 Results in synthetic time series

Once compared QARGA with other EA in public datasets—that were static and non-temporal-dependent—the algorithm is assessed when applied to time series. For this reason, two different types of synthetic time series were generated, as described in Sect. 5.1.2.

Table 15 shows the rules obtained by an execution of QARGA when multidimensional synthetic time series with no disjunctions (see Table 3 for detailed data description) were analyzed. Similar rules have been obtained by other executions of QARGA.

From the ten discovered rules, the four first ones (rules #0 to #3) highlight and are considered especially meaningful insofar as they represent, exactly, the intervals used in Table 3 to generate the time series itself. That is, QARGA was able to precisely discover the rules that model the synthetic time series generation.

It can also be observed that the support (*Sup*. column) in these rules is 25%, which coincides with the preset support when the time series were generated. Equally remarkable is that the confidence (*Conf*. column) is 100% for all the four rules. It is also noteworthy the precision of the intervals found since most of the limits discovered by QARGA coincide with those of the Table 3, which means a great level of reliability in the rules. Finally, note that the lift is much greater than one, in other words, such antecedents and consequents are likely to appear together.

As for the six remaining rules (rules #4 to #9), they correspond to rules with smaller support, confidence and lift. This fact can be justified by taking into consideration that when an IRL algorithm is applied, the search space is constantly being decreased and, therefore, the obtained rules cover less samples with less precision.

On the other hand, Table 16 shows the 12 most relevant rules obtained by QARGA when synthetic time series with disjunctions (see Table 4 for detailed data description) were analyzed. To facilitate the analysis, they are listed according to the interval to which the time variable –t belongs to, as listed in Table 5.

In general terms, each rule in Table 16 represents one of the expected rules listed in Table 5, except for some cases, which are discussed now. Thus, rule #0 in Table 16 represents the expected rule #0 in Table 5. The support is 19%, a value very close to the expected one. In addition, this rule has a 100% confidence.

When the time is in the interval $[26, 50]$, there were two possible expected conjunctions, $\#1_1$ and $\#1_2$. In this case, rule #1 approximately represents $\#1_1$ and rule #2 does $\#1_2$. Regarding the support, it is not significantly different from the expected one. Finally, the confidence is nearly 100% for both of the rules.

Rules #4 and #5 identify rules with $t \in [54, 75]$ and correspond to conjunctions $\#2_1$ and $\#2_2$, respectively. Again, the support is not very different from the expected one and the confidence is 100% for all of them.

Four different conjunctions were expected—$\#3_1$, $\#3_2$, $\#3_3$ and $\#3_4$—when $t \in [76, 100]$. Rule #8 identifies

Table 15 Rules found by QARGA for time series with no disjunctions

ID	Rules	Conf. (%)	Lift	Sup. (%)	Amp. (%)
#0	$X_1 \in [1, 15] \wedge X_3 \in [60, 75] \Longrightarrow t \in [1, 25] \wedge X_2 \in [7, 34]$	100	4.0	25.0	20.0
#1	$t \in [26, 50] \wedge X_1 \in [5, 30] \Longrightarrow X_2 \in [27, 40] \wedge X_3 \in [11, 29]$	100	4.0	25.0	20.0
#2	$X_1 \in [45, 60] \wedge X_2 \in [55, 85] \wedge X_3 \in [20, 35] \Longrightarrow t \in [51, 75]$	100	4.0	25.0	21.0
#3	$t \in [76, 100] \wedge X_1 \in [75, 100] \wedge X_3 \in [40, 60] \Longrightarrow X_2 \in [1, 18]$	100	2.9	25.0	21.5
#4	$t \in [1, 11] \wedge X_1 \in [1, 14] \wedge X_2 \in [20, 29] \Longrightarrow X_3 \in [66, 75]$	100	5.9	7.0	10.3
#5	$t \in [82, 92] \wedge X_2 \in [8, 17] \Longrightarrow X_1 \in [82, 92] \wedge X_3 \in [49, 55]$	50.0	7.1	3.0	8.7
#6	$t \in [48, 58] \wedge X_1 \in [47, 57] \Longrightarrow X_2 \in [72, 81] \wedge X_3 \in [20, 30]$	50.0	10.0	4.0	9.6
#7	$X_1 \in [1, 12] \wedge X_2 \in [24, 34] \wedge X_3 \in [64, 75] \Longrightarrow t \in [11, 21]$	85.7	7.8	6.0	10.5
#8	$X_2 \in [3, 17] \wedge X_3 \in [41, 49] \Longrightarrow t \in [82, 100] \wedge X_1 \in [84, 99]$	75.0	4.7	9.0	13.8
#9	$X_1 \in [53, 62] \Longrightarrow t \in [68, 78] \wedge X_2 \in [53, 61] \wedge X_3 \in [20, 26]$	17.6	5.9	3.0	8.7

Table 16 Rules found by QARGA for time series with disjunctions

ID	Rules	Conf. (%)	Lift	Sup. (%)	Amp. (%)
#0	$X_1 \in [20, 30] \wedge X_2 \in [61, 79] \Longrightarrow t \in [1, 25]$	100	4.00	19	17.33
#1	$t \in [28, 50] \wedge X_2 \in [47, 58] \Longrightarrow X_1 \in [12, 20]$	93.3	2.39	14	13.67
#2	$t \in [26, 46] \wedge X_1 \in [25, 31] \Longrightarrow X_2 \in [41, 49]$	100	2.38	7	11.33
#3	$t \in [32, 47] \wedge X_2 \in [46, 53] \Longrightarrow X_1 \in [16, 22]$	62.5	2.98	5	9.35
#4	$t \in [58, 75] \wedge X_2 \in [40, 50] \Longrightarrow X_1 \in [4, 15]$	100	3.23	13	12.67
#5	$X_1 \in [4, 14] \wedge X_2 \in [60, 70] \Longrightarrow t \in [54, 72]$	100	5.26	8	12.67
#6	$t \in [79, 95] \wedge X_1 \in [12, 18] \Longrightarrow X_2 \in [39, 49]$	85.7	1.82	6	10.62
#7	$t \in [76, 91] \wedge X_1 \in [15, 24] \Longrightarrow X_2 \in [41, 50]$	85.7	1.75	6	10.95
#8	$t \in [77, 95] \wedge X_1 \in [16, 22] \Longrightarrow X_2 \in [31, 39]$	100	33.33	3	10.62
#9	$t \in [76, 99] \wedge X_1 \in [18, 40] \Longrightarrow X_2 \in [31, 50]$	100	1.79	18	21.33
#10	$X_1 \in [28, 34] \wedge X_2 \in [36, 43] \Longrightarrow t \in [83, 98]$	60.0	4.29	3	9.35
#11	$t \in [76, 99] \wedge X_1 \in [30, 38] \Longrightarrow X_2 \in [40, 50]$	100	1.89	13	13.67

conjunction $\#3_1$ and rule #7 is approximately $\#3_2$. In the same fashion, rule #10 is related to $\#3_3$ and rule #11 to $\#3_4$.

The remaining rules discovered relationships varying among several conjunctions. For rule #6, note that the X_2 time series has values ranging in an interval formed from the union of rules $\#3_1$ and $\#3_2$. Last, rule #9 is a rule resulting from the four possible conjunctions in $t \in [76, 100]$, that is, it combines the intervals for both X_1 and X_2. Therefore, it would be a rule shared by $\#3_1, \#3_2, \#3_3$ and $\#3_4$. In general, rules in this interval of time share a support close to the expected one as well as a confidence verging on 100% for most cases.

Finally, it can be concluded that all the rules discovered by QARGA can be considered interesting since the lift is high for all of them. Moreover, the amplitude of intervals is moderate and intervals limits are very similar to those initially set in Table 5.

5.3.3 Results in ozone time series

Now QARGA is applied to ozone time series and other inter-dependant temporal variables. Table 17 shows the support, confidence, number of records, average amplitude and lift of the obtained rules by QARGA when the ozone is imposed to be in the consequent. The climatological variables that most frequently appear are temperature, humidity and hour of the day. Consequently, it can be concluded that the other variables are not as correlated with ozone as the aforementioned ones.

Some other interesting conclusions can be extracted from these rules. Hence, when the temperature reaches high values, the ozone concentration in the atmosphere presents high values, even reaching 203 μg/m^3. Nevertheless, when the temperature is relatively low, the concentration of ozone falls to values around 116 μg/m^3. That is, there exists a perfect correlation between the ranges of the temperature and the ozone. With reference to the humidity, there exists an inversely proportional relationship to the ozone. Thus, when examining the first rule, in contrast to the temperature, when the humidity falls, the ozone raises, and viceversa, as occurred in the fourth rule (rule #3).

From the remaining rules, it can also be observed that the time slot is present in two rules. This fact is due to the close association existing between the temperature and the hour of the day and, possibly, to the traffic, whose density varies along the day and typically generates high concentrations of ozone. Note that during the night and first hours

Table 17 Association rules found by QARGA with high confidence

ID	Rule	Sup. (%)	Conf. (%)	#r	Ampl.	Lift
#0	temp. $\in [32, 42]$ and hum. $\in [19, 41]$ and hour $\in [13, 19] \Longrightarrow O_3 \in [104, 203]$	14	84	213	34	2.27
#1	hour $\in [2, 11] \Longrightarrow O_3 \in [16, 97]$	37	90	557	45	1.64
#2	hour $\in [13, 22] \Longrightarrow O_3 \in [88, 189]$	35	84	522	55	1.67
#3	temp. $\in [16, 22]$ and hum. $\in [75, 90] \Longrightarrow O_3 \in [22, 110]$	16	100	234	36	1.43
#4	temp. $\in [18, 29]$ and speed $\in [0, 10] \Longrightarrow O_3 \in [23, 116]$	41	93	613	38	1.28
		28.6 (±12.6)	90.2 (±6.7)	427.80 (±189.5)	41.6 (±8.6)	1.7 (±0.4)

of the day, the ozone is relatively low, reaching values similar to that of low temperatures. However, from midday to the nightfall—the rushing hours—the amount of ozone increases considerably, reaching values near to 200 μg/m^3 as it happened with high temperatures.

Also note that in one rule the speed of the wind appears indicating that when it is low the ozone also is. However, this rule is not conclusive and the authors do not dare to state that the speed of the wind is directly proportional to the ozone.

With the aim of comparing the results and evaluating the quality, the Apriori algorithm has been applied to these time series. The most remarkable feature of this algorithm is that is based on a previous or *a priori* knowledge of the frequent itemsets in order to reduce the space of search and, consequently, increase the efficiency. Besides, the user has to establish the constraints for minimum support and confidence. It is also worth mentioning that Apriori does not work with real values directly and it performs a previous discretization of all continuous variables.

Hence, Table 18 collects the results provided by Apriori when discretizing the continuous variables with three intervals. In this case, the temperature and the humidity appear again but, by contrast, the hour of the day does not seem to be an important variable. The speed and the direction of the wind also appear in the antecedent.

It can also be observed that low temperatures also involve low ozone concentrations, and viceversa, as it happened with the rules shown in Table 17. With regard to the humidity the same situation is reported: it is inversely proportional to the ozone. However, when analyzing the direction of the wind in some rules, the results are not conclusive. Actually, for equal values of the direction, different ranges of ozone are mined, which means that this variable presents no proportional (neither direct nor inverse) relationship with the ozone and, therefore, it does not contribute with meaningful information. Finally, the speed of the wind presents the same behavior shown in Table 17, that is, low values involve low ozone concentrations.

The comparison between Tables 17 and 18 reveals that the support reached by QARGA is much greater in three rules whereas two rules present slightly lower supports. The confidence for the majority of the rules found by QARGA overcomes 90%, even reaching 100% in the fourth rule. This fact highlights the small errors committed by QARGA, providing exact rules in the majority of cases. Furthermore, the number of covered instances is higher than the ones by Apriori due to the direct relation existing with the support. The average amplitude for the rules provided by QARGA is much smaller, ranging from 38 to 55, while the intervals found by Apriori varies from 32 to 98. The lift is very similar in QARGA and Aprori, and for both algorithms it is greater than 1.

Last, Apriori has just found rules in which the values of the ozone varied only in two of the three possible intervals associated with the labels previously generated during the discretization process. Furthermore, it is unable to find rules with ozone concentrations higher than 183 μg/m^3. On the contrary, QARGA obtained rules for concentrations of ozone higher than 200 μg/m^3.

To sum up, from this kind of experimentation it can be concluded that QARGA obtains better results compared with Apriori, since support and confidence are higher and amplitude is smaller, which involves less errors in rules.

Table 19 shows the rules obtained by QARGA when the target is to find rules with the highest number of attributes, the highest confidence and the smallest amplitude possible, even if this fact may lead to lower supports. It can be observed that the majority of rules have a large number of attributes. The variables that most frequently appear, therefore the most meaningful ones, are the temperature, the humidity and the hour of day.

From the extracted rules, several conclusions can be drawn. First, note that the selected rules are those that present high concentrations of ozone in the consequent, since this is the situation that really involves environmental concerns. As with the first experimentation, a directly proportional relation between the temperature and the ozone has been discovered. In other words, when the

Table 18 Association rules found by Apriori (three intervals used for discretization)

ID	Rule	Sup. (%)	Conf. (%)	#r	Ampl.	Lift
#0	temp. $\in$ [16, 25] and hum. $\in$ [65, 91] and speed $\in$ [0, 9] $\Longrightarrow O_3 \in$ [14, 99]	20	90	297	32	1.59
#1	temp. $\in$ [16, 25] and dir. $\in$ [120, 240] and speed $\in$ [0, 9] $\Longrightarrow O_3 \in$ [14, 99]	16	85	239	56	1.49
#2	dir. $\in$ [240, 360] and speed $\in$ [0, 9] $\Longrightarrow O_3 \in$ [14, 99]	16	79	241	71	1.38
#3	hum. $\in$ [14, 40] and dir. $\in$ [120, 240] $\Longrightarrow O_3 \in$ [99, 183]	18	73	261	77	1.80
#4	temp. $\in$ [25, 35] and dir. $\in$ [120, 240] $\Longrightarrow O_3 \in$ [99, 183]	20	70	296	98	1.70
		18.0 (±2.0)	79.4 (±8.3)	266.8 (±28.5)	66.8 (±24.6)	1.6 (±0.2)

Table 19 Association rules found by QARGA with high confidence

ID	Rule	Sup. (%)	Conf. (%)	#r	Ampl.	Lift
#0	temp. ∈ [38, 42] and hum. ∈ [25, 33] and hour ∈ [15, 18] $\Longrightarrow O_3$ ∈ [140, 206]	3	90	47	20	6.61
#1	temp. ∈ [26, 33] and hum. ∈ [29, 46] and dir. ∈ [149, 231] and speed ∈ [12, 20] and hour ∈ [15, 19] $\Longrightarrow O_3$ ∈ [103, 160]	3	93	40	29	2.90
#2	temp. ∈ [29, 37] and hum. ∈ [21, 36] and dir. ∈ [161, 187] and hour ∈ [14, 19] $\Longrightarrow O_3$ ∈ [109, 180]	5	83	70	25	2.83
#3	temp. ∈ [34, 42] and hum. ∈ [22, 32] and dir. ∈ [152, 189] and speed ∈ [9, 16] and hour ∈ [15, 17] $\Longrightarrow O_3$ ∈ [134, 206]	2	87	33	23	4.99
#4	hum. ∈ [22, 32] and hour ∈ [13, 18] $\Longrightarrow O_3$ ∈ [144, 198]	7	48	97	23	4.27
		4.0 (±2.0)	80.2 (±18.4)	57.4 (±26.1)	24.0 (±3.3)	4.32 (±1.58)

temperature reaches values of almost 40 °C the ozone level raises up to values greater than 200 μg/m^3.

Moreover, the humidity also presents an inversely proportional relationship with the ozone. Thus, when it reaches high values, the concentration of ozone is low, as it can be determined from the observation of the second rule. Alternatively, when the humidity increases, the ozone level decreases, as listed in the remaining rules.

The hour of the day is also present in the majority of the rules. The time slot is similar for all the rules since, as discussed previously, ozone and the hour share a directly proportional relationship. The peaks of ozone are reached during the rushing hours (from midday to nightfall), that is, during the hours in which the temperature is high and the traffic is usually heavy.

In some rules, the speed of the wind appears as a crucial factor. However, there does not exist such a higher correlation with the ozone because greater speeds should have be found in the fourth rule (in which the ozone presents the highest concentration). By contrast, in the second rule, in which the ozone is lower, the speed of the wind is slightly superior.

The analysis of the direction of the wind reveals that it is not a variable that determines the amount of ozone in the atmosphere. However, when the direction is comprised in an interval from 150° and 200°, the concentration of ozone increases.

Table 20 gathers the results obtained by Apriori when data were discretized in ten intervals. The temperature is, again, the main variable. However, it is worth pointing out that no relevant rules were discovered in which the humidity or the hour of the day appear.

One of the most remarkable feature of the extracted rules by Apriori when discretizing with ten intervals is that they all have only one attribute in the antecedent. This situation highlights, once again, that rules provided by QARGA enhance that of Apriori, since they are more expressive and provide more information due to a greater number of attributes in antecedents.

With respect to the temperature, two rules with different antecedent but same consequent have been discovered. Note that they could have been fused into one rule as the consequent is the same. Besides, the obtained confidence is quite low which leads to rules with considerably high errors.

The case of the direction of the wind is similar to that of the temperature. The third and fifth rules share the same antecedent for the same direction and, however, the consequent for both rules is different even when they could have been fused into just one rule. The confidence hardly reaches 30%, which leads to an almost null reliability.

The speed of the wind appears in one rule in which its value is low and the ozone presents medium values. However, the confidence is quite low.

Table 20 Association rules found by Apriori (ten intervals used for discretization)

ID	Rule	Sup. (%)	Conf. (%)	#r	Ampl.	Lift
#0	temp. ∈ [27, 30] $\Longrightarrow O_3$ ∈ [90, 115]	5	37	79	14	1.43
#1	temp. ∈ [24, 27] $\Longrightarrow O_3$ ∈ [90, 115]	6	33	85	14	1.43
#2	dir. ∈ [144, 180] $\Longrightarrow O_3$ ∈ [90, 115]	10	29	156	31	1.18
#3	speed ∈ [6, 8] $\Longrightarrow O_3$ ∈ [90, 115]	5	24	75	14	1.03
#4	dir. ∈ [144, 180] $\Longrightarrow O_3$ ∈ [115, 141]	6	16	88	31	1.27
		6.4 (±2.1)	27.8 (±8.2)	96.6 (±33.6)	20.8 (±9.3)	1.27 (±0.17)

The comparison of the Tables 19 and 20 leads to several conclusions. As regards the attributes, QARGA always obtains rules with greater number of them and, consequently, the information brought by these rules is higher than that of Apriori.

When taking into consideration the support, QARGA presents low values since there are few instances in the dataset with high ozone values. However, if the confidence of the rules for both algorithms is compared, it can be observed that QARGA has values even greater than 90% while Apriori never overcomes 40%.

Unlike the lift values from the first kind of experimentation, where the interest of the rules in QARGA and Apriori was quite similar, it can be observed that, in this case, the results of the lift are very different. Rules found by QARGA present lift values between 3 and 6, while Apriori never exceeds 1.50. This is an important result that indicates that QARGA find more interesting rules than Apriori does.

Another relevant remark is that Apriori discovers rules with different intervals for the same variable in the antecedent but equal consequents and viceversa. This fact never occurs in QARGA.

The ozone levels obtained by Apriori never exceeds 140 μg/m^3, while QARGA reached values greater than 200 μg/m^3. This appreciation is of the utmost relevance, since environment is really concerned by high levels of ozone and, consequently, discovering rules with these values of ozone is useless.

6 Conclusions

An evolutionary algorithm has been proposed in this work to obtain QAR from time series. In order to evaluate its performance, the approach has been applied to several datasets and compared with the most recently published results. Thus, a bank of public datasets retrieved from the BUFA repository has been used to test the accuracy of the algorithm. The algorithm has shown to be efficient when mining synthetically generated multidimensional time series. Also, the proposed methodology has successfully obtained meaningful QAR from multidimensional real-world time series. In particular, relevant dependencies between the ozone concentration in the atmosphere and other climatological-related time series have been found.

Acknowledgments The financial support from the Spanish Ministry of Science and Technology, project TIN2007-68084-C02, and from the Junta de Andalucía, project P07-TIC-02611, is acknowledged. The authors also want to acknowledge the support by the Regional Ministry for the Environment (*Consejería de Medio Ambiente*) of Andalucía (Spain), that has provided all the pollutant agents time series.

References

Adame-Carnero JA, Bolfvar JP, de la Morena BA (2010) Surface ozone measurements in the southwest of the Iberian Peninsula. Environ Sci Pollut Res 17(2):355–368

Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD international conference on management of data, pp 207–216

Agirre-Basurko E, Ibarra-Berastegi G, Madariagac I (2006) Regression and multilayer perceptron-based models to forecast hourly o_3 and no_2 levels in the Bilbao area. Environ Model Softw 21:430–446

Aguilar-Ruiz JS, Giráldez R, Riquelme JC (2007) Natural encoding for evolutionary supervised learning. IEEE Trans Evol Comput 11(4):466–479

Alatas B, Akin E (2006) An efficient genetic algorithm for automated mining of both positive and negative quantitative association rules. Soft Comput 10(3):230–237

Alatas B, Akin E (2008) Rough particle swarm optimization and its applications in data mining. Soft Comput 12(12):1205–1218

Alatas B, Akin E, Karci A (2008) MODENAR: multi-objective differential evolution algorithm for mining numeric association rules. Appl Soft Comput 8(1):646–656

Alcalá-Fdez J, Alcalá R, Gacto MJ, Herrera F (2009a) Learning the membership function contexts forming fuzzy association rules by using genetic algorithms. Fuzzy Sets Syst 160(7):905–921

Alcalá-Fdez J, Sánchez L, García S, del Jesus MJ, Ventura S, Garrell JM, Otero J, Romero C, Bacardit J, Rivas VM, Fernández JC, Herrera F (2009b) Keel: a software tool to assess evolutionary algorithms for data mining problems. Soft Comput 13(3):307–318. http://sci2s.ugr.es/keel

Alcalá-Fdez J, Flugy-Pape N, Bonarini A, Herrera F (2010) Analysis of the effectiveness of the genetic algorithms based on extraction of association rules. Fundam Inform 98(1):1001–1014

Aumann Y, Lindell Y (2003) A statistical theory for quantitative association rules. J Intell Inf Syst 20(3):255–283

Bellazzi R, Larizza C, Magni P, Bellazzi R (2005) Temporal data mining for the quality assessment of hemodialysis services. Artif Intell Med 34:25–39

Berlanga FJ, Rivera AJ, del Jesus MJ, Herrera F (2010) GP-COACH: genetic programming-based learning of compact and accurate fuzzy rule-based classification systems for high-dimensional problems. Inf Sci 180(8):1183–1200

Brin S, Motwani R, Silverstein C (1997) Beyond market baskets: generalizing association rules to correlations. In: Proceedings of the 1997 ACM SIGMOD international conference on management of data, vol 26, pp 265–276

Chen CH, Hong TP, Tseng V (2009) Speeding up genetic-fuzzy mining by fuzzy clustering. In: Proceedings of the IEEE international conference on fuzzy systems, pp 1695–1699

Chen CH, Hong TP, Tseng V (2010) Genetic-fuzzy mining with multiple minimum supports based on fuzzy clustering. Soft Comput (in press)

del Jesús MJ, Gámez J, Puerta J (2009) Evolutionary and metaheuristics based data mining. Soft Comput Fusion Found Methodol Appl 13:209–212

Elkamel A, Abdul-Wahab S, Bouhamra W, Alper E (2001) Measurement and prediction of ozone levels around a heavily industrialized area: a neural network approach. Adv Environ Res 5:47–59

García S, Fernández A, Luengo J, Herrera F (2009) A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. Soft Comput 13(10):959–977

Georgii E, Richter L, Ruckert U, Kramer S (2005) Analyzing microarray data using quantitative association rules. BMC Bioinformatics 21(2):123–129

Gupta N, Mangal N, Tiwari K, Pabitra Mitra (2006) Mining quantitative association rules in protein sequences. Lect Notes Artif Intell 3755:273–281

Guvenir HA, Uysal I (2000) Bilkent university function approximation repository. http://funapp.cs.bilkent.edu.tr

Herrera F, Lozano M, Sánchez AM (2004) Hybrid crossover operators for real-coded genetic algorithms: an experimental study. Soft Comput 9(4):280–298

Huang YP, Kao LJ, Sandnes FE (2008) Efficient mining of salinity and temperature association rules from ARGO data. Expert Syst Appl 35:59–68

Kalyanmoy D, Ashish A, Dhiraj J (2002) A computationally efficient evolutionary algorithm for real-parameter optimization. Evol Comput 10(4):371–395

Khan MS, Coenen F, Reid D, Patel R, Archer L (2010) A sliding windows based dual support framework for discovering emerging trends from temporal data. Res Dev Intell Syst Part 2:35–48

Lin MY, Lee SY (2002) Fast discovery of sequential patterns by memory indexing. In: Proceedings of the 4th international conference on data warehousing and knowledge discovery, pp 150–160

Martínez–Álvarez F, Troncoso A, Riquelme JC, Aguilar JS (2011) Energy time series forecasting based on pattern sequence similarity. IEEE Trans Knowl Data Eng (in press)

Mata J, Álvarez J, Riquelme JC (2001) Mining numeric association rules with genetic algorithms. In: Proceedings of the international conference on adaptive and natural computing algorithms, pp 264–267

Mata J, Álvarez JL, Riquelme JC (2002) Discovering numeric association rules via evolutionary algorithm. Lect Notes Artif Intell 2336:40–51

Nam H, Lee K, Lee D (2009) Identification of temporal association rules from time-series microarray data sets. BMC Bioinformatics 10(3):1–9

Nikolaidou V, Mitkas PA (2009) A sequence mining method to predict the bidding strategy of trading agents. Lect Notes Comput Sci 5680:139–151

Orriols-Puig A, Bernadó-Mansilla E (2009) Evolutionary rule-based systems for imbalanced data sets. Soft Comput Fusion Found Methodol Appl 13:213–225

Orriols-Puig A, Casillas J, Bernadó-Mansilla E (2008) First approach toward on-line evolution of association rules with learning classifier systems. In: Proceedings of the 2008 GECCO genetic and evolutionary computation conference, pp 2031–2038

Pei J, Han JW, Mortazavi-Asl B, Pinto H, Chen Q, Dayal U, Hsu MC (2001) Prefixspan: mining sequential patterns efficiently by prefix-projected pattern growth. In: Proceedings of IEEE conference on data engineering, pp 215–224

Ramaswamy S, Mahajan S, Silberschatz A (1998) On the discovery of interesting patterns in association rules. In: Proceedings of the 24th international on very large data bases, pp 368–379

Sheskin D (2006) Handbook of parametric and nonparametric statistical procedures. Chapman and Hall/CRC

Shidara Y, Kudo M, Nakamura A (2008) Classification based on consistent itemset rules. Trans Mach Learn Data Min 1(1):17–30

Tong Q, Yan B, Zhou Y (2005) Mining quantitative association rules on overlapped intervals. Lect Notes Artif Intell 3584:43–50

Tung AKH, Han J, Lu H, Feng L (2003) Efficient mining of intertransaction association rules. IEEE Trans Knowl Data Eng 15(1):43–56

Vannucci M, Colla V (2004) Meaningful discretization of continuous features for association rules mining by means of a som. In: Proceedings of the European symposium on artificial neural networks, pp 489–494

Venturini G (1993) SIA: a supervised inductive algorithm with genetic search for learning attribute based concepts. In: Proceedings of the European conference on machine learning, pp 280–296

Venturini G (1994) Fast algorithms for mining association rules in large databases. In: Proceedings of the international conference on very large databases, pp 478–499

Wan D, Zhang Y, Li S (2007) Discovery association rules in time series of hydrology. In: Proceedings of the IEEE international conference on integration technology, pp 653–657

Wang YJ, Xin Q, Coenen F (2008) Hybrid rule ordering in classification association rule mining. Trans Mach Learn Data Min 1(1):17–30

Winarko E, Roddick JF (2007) ARMADA—an algorithm for discovering richer relative temporal association rules from interval-based data. Data Knowl Eng 63:76–90

Wright SP (1992) Adjusted *p*-values for simultaneous inference. Biometrics 48:1005–1013

Yan X, Zhang C, Zhang S (2009) Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. Expert Syst Appl Int J 36(2):3066–3076

4.2. Mining quantitative association rules based on evolutionary computation and its application to atmospheric pollution

Las publicaciones en revista asociadas a esta parte son:

- M. Martínez-Ballesteros, F.Martínez-Álvarez, A. Troncoso Lora, J.C. Riquelme Santos. *Mining quantitative association rules based on evoluationary computation and its application to atmospheric pollution.* Integrated Computer-Aided Engineering (ICAE), Vol. 17, No. 3, pages 227-242, 2010 [Martínez-Ballesteros et al., 2010].

    - Estado: Publicado
    - Índice de Impacto (JCR 2010): 2.122
    - Área de Conocimiento:
        - Computer Science, Artificial Intelligence. Ranking 26 / 108 - Q1
        - Computer Science, Interdisciplinary Applications. Ranking 17 / 97 - Q1
        - Engineering, Multidisciplinary. Ranking 5 / 87 - Q1

Integrated Computer-Aided Engineering 17 (2010) 227–242
DOI 10.3233/ICA-2010-0340
IOS Press

Mining quantitative association rules based on evolutionary computation and its application to atmospheric pollution

M. Martínez-Ballesteros[a], A. Troncoso[b,*], F. Martínez-Álvarez[b] and J. C. Riquelme[a]
[a]*Department of Computer Science, University of Seville, Seville, Spain*
[b]*Area of Computer Science, Pablo de Olavide University of Seville, Seville, Spain*

Abstract. This research presents the mining of quantitative association rules based on evolutionary computation techniques. First, a real-coded genetic algorithm that extends the well-known binary-coded CHC algorithm has been projected to determine the intervals that define the rules without needing to discretize the attributes. The proposed algorithm is evaluated in synthetic datasets under different levels of noise in order to test its performance and the reported results are then compared to that of a multi-objective differential evolution algorithm, recently published. Furthermore, rules from real-world time series such as temperature, humidity, wind speed and direction of the wind, ozone, nitrogen monoxide and sulfur dioxide have been discovered with the objective of finding all existing relations between atmospheric pollution and climatological conditions.

Keywords: Data mining, evolutionary algorithms, quantitative association rules

1. Introduction

Predicting a chronological sequence of observations on a variable, commonly known as *time series forecasting*, has been traditionally performed by the application of statistical methods [8]. The results obtained from such methods for synthetic data are usually satisfactory. Furthermore, the inherent simplicity shown by statistical-based methods makes their use popular and widespread. However, when dealing with real-world time series the accuracy of the predictions are not as expected since these datasets often present non-linear features that the classical Box-Jenkins approaches are unable to model.

The temporary evolution of most variables is usually influenced by the changes occurring in other time series. In other words, the correlation between different time series is a frequent phenomenon. For instance, when a rainfall forecast is required, the analysis of other variables such as temperature, humidity or atmospheric pressure is mandatory. Consequently, a diligent analysis of the correlated variables may lead to the discovery of how the variable in question may behave in the near future.

The goal of the association rules (AR) extraction process precisely consists of discovering the presence of pair conjunctions (attribute (A) – value (v)) that appear in a dataset with a certain frequency in order to formulate the rules that outline the existing relationship among attributes. Formally, an association rule is a relationship between attributes in a database such that $C_1 \Rightarrow C_2$, where C_1 and C_2 are pair conjunctions such as $A = v$ if $A \in \mathbb{Z}$ or $A \in [v_1, v_2]$ if $A \in \mathbb{R}$. Generally, the antecedent C_1 is formed by a conjunction of multiple pairs and the consequent C_2 is usually a single pair.

The main motivation of this research is to develop a genetic algorithm (GA) capable of finding quantitative association rules in databases with continuous attributes avoiding the discretization as a prior step of the

[*] Corresponding author: A. Troncoso, Pablo de Olavide University of Seville, Ctra. Utrera, Km.1, 41013, Sevilla, Spain. Tel.: +34 95 4977522; Fax: +34 95 4348377; E-mail: ali@upo.es.

ISSN 1069-2509/10/$27.50

process. Thus, a real-coded genetic algorithm (RCGA) that expands the general scheme of the CHC binary-coded evolutionary algorithm [15] is proposed in this work. The approach provides numeric association rules establishing relationships among all attributes of the datasets.

For evaluating the performance of the RCGA, two different kind of datasets are analyzed. On one hand, its application over synthetic datasets is reported. On the other hand, an attempt to forecast real-world time series is made by means of the extracted quantitative association rules.

With regard to the real-world time series, three environmental agents responsible for pollution are evaluated: ozone (O_3), sulfur dioxide (SO_2) and nitrogen monoxide (NO). The tropospheric ozone is an atmospheric particle typically identified as a pollutant when it overlaps some threshold. The variation in concentration of this agent in the air is continuously studied, as the noxious effects caused in all living beings is well known [30]. Both sulfur dioxide and nitrogen monoxide are usually formed in various industrial processes, and its concentration in the air has dramatically increased during the last decade. Higher concentrations may cause what experts usually call *acid rain,* which causes damage to living beings and infrastructures [17].

The search of AR in ozone time series must not be mistaken with the Subgroup Discovery (SD) issue [13]. The AR are a non-supervised learning tool, while the SD performs supervised learning. Both AR and SD search for rules but SD searches for conditions of a single attribute. Nevertheless, AR can deal with multiple attributes in the antecedent and in the consequent. Moreover, the AR do not preset the range to which the attributes of the consequent can vary.

The rest of the paper is divided as follows: Section 2 describes the state of the art. Section 3 provides the methodology used in this work. The results of the approach applied to synthetic data are discussed in Section 4. Section 5 refers to the results obtained for the atmospheric datasets. Finally, Section 6 discusses the resulting conclusions.

2. State of the art

There are many efficient algorithms that find AR. Genetic algorithms have been used profusely to generate rules in many learning problems [2,9,24]. Also, genetic algorithms are used as a tool in many real-world problems, such as scheduling [14], forecasting [35], design [26] or classification [10]. Finally, hybridization with fuzzy logic [31], neural networks [20] or simulation [11] are common strategies in evolutionary computation.

However, many researchers focus on databases with discrete attributes while most real-world databases essentially contain continuous attributes, as in the case with time series analysis [6]. Moreover, the majority of the tools said to work in the continuous domain just discretize the attributes using a specific strategy and later, handle these attributes as if they were discrete [1, 33].

A review of recently published literature reveals that the amount of works providing metaheuristics and search algorithms relating to AR with continuous attributes is scarce. Thus, the authors of [25] proposed an evolutionary algorithm to discover numeric association rules, dividing the process in two phases. The first one determined the frequent itemsets, that is, the set of features appearing with a certain frequency within a dataset. In the second phase, the rules were extracted from the itemsets previously calculated.

The work presented in [32] studied the conflict between minimum support and confidence problems. They proposed a method to find quantitative AR by clustering the transactions of a database. Afterwards, such groupings were projected into the domains of the attributes in order to create meaningful intervals which could be overlapped.

Hydrological time series were studied in [36]. First, the numeric attributes were transformed into intervals by means of clustering techniques. Then, the AR were generated making use of the well-known Apriori algorithm [1].

A classifier system was presented in [28] with the purpose of extracting quantitative AR over unlabeled (both numerical and categorical) data streams. The main novelty of this approach was the efficiency and adaptability to data gathered on-line.

A metaheuristic optimization based on rough particle swarm techniques was presented in [3]. In this case, the singularity was the obtention of the values that determine the intervals for the AR instead of frequent itemsets. In synthetic data, several new operators such as rounding, repairing and filtering were evaluated and tested.

MODENAR is a multi-objective pareto-based genetic algorithm that was presented in [4]. The fitness function was composed of four different objectives: Support, confidence, comprehensibility of the rule (to be maximized) and the amplitude of the intervals that constitutes the rule (to be minimized).

The work published in [38] exhibited a new approach based on three novel algorithms: value-interval clustering, interval-interval clustering and matrix-interval clustering. Their application was found especially useful when mining complex information.

Another GA was used in [37] in order to obtain numeric AR. However, the unique objective to be optimized in the fitness function was the confidence. To fulfill this goal, the authors avoided specifying the actual minimum support, which is the main contribution to this work.

The use of AR in bioinformatics is also widely spread. Hence, the work in [16] analyzed microarray data using quantitative AR. For this purpose, they chose a variant of the algorithm introduced in [29] based on half-spaces or linear combinations of bounded variables against a constant. Moreover, Gupta et al. mined quantitative AR for protein sequences [19] and for this reason they proposed a new algorithm with four steps to follow. They first equi-depth partitioned the attributes; second, the partitions were mapped on consecutive integers, thus representing the intervals; third, they found the support of all intervals; and, finally, they used the frequent itemsets to generate AR. On the other hand, the authors in [27] proposed a novel temporal association rule mining method based on the Apriori algorithm. Hence, they identified temporary dependencies from gene-related time series.

The AR had been applied in fuzzy sets by various authors. Thus, Kaya and Alhajj first proposed a GA-based framework for mining fuzzy AR in [21]. To be precise, they presented a clustering method for adjusting the centroids of the clusters and then, they provided a different approach based on the well known CURE [18] clustering algorithm to generate membership functions. Later, they introduced a GA to optimize membership functions for fuzzy weighted AR mining in [22]. Their proposal automatically adjusted these sets to provide maximum support and confidence. To fulfill this goal, the base values of the membership functions for each quantitative attribute were refined by maximizing two different evaluation functions: the number of large itemsets and the confidence interval average of the generated rules. Alternatively, Alcalá-Fdez et al. [5] presented a new algorithm for extracting fuzzy AR and membership functions by means of evolutionary learning based on the 2-tuples representation model.

Finally, Ayubi et al. [7] proposed an algorithm that mined general rules whose applicability ranged from discrete attributes to quantitative discretized ones. Thus, they stored general itemsets in a tree structure in order for it to be recursively computed. They equally addressed the association rules in tabular form allowing a set of different operators.

3. Description of the algorithm

In this work a real-coded [23] genetic algorithm (hereafter called RCGA) has been used to obtain AR from quantitative datasets. The proposed RCGA follows the general scheme of the CHC binary-coded evolutionary algorithm proposed by Eshelman in 1991 [15]. The original CHC presents an elitist strategy for selecting the population that will make up the next generation and includes strong diversity in the evolutionary process through mechanisms of incest prevention and a specific operator of crossover called Half Uniform (HUX). Furthermore, the population is reinitialized when its diversity is poor. Details of these main features of the CHC algorithm are outlined in the following points.

- Elitist selection: This kind of strategy guarantees the survival of the best individuals. Thus, the current population and its offspring are joined and the best individuals (according to the fitness function) are chosen to compose the population of the next generation.
- The HUX crossover operator: This operator swaps exactly half of the nonmatching genes of the parents. Therefore, the Hamming distance divided by two is the number of genes to be swaped. This crossover is highly destructive and introduces some diversity in the population preventing premature convergence.
- Incest prevention: In the CHC algorithm the crossover among siblings is forbidden. Therefore, in order to prevent this, the following function is applied: Two individuals are only crossed if their Hamming distance divided by two is greater than a certain threshold which is set to the length of the individual, i.e. the number of bits, divided by four. Consequently, only highly dissimilar parents are crossed. When there are no parents to be crossed due to their Hamming distance divided by two is less than the predetermined threshold, the threshold is decremented by one unit. As such, the key idea is to avoid the application of the crossover operator among similar individuals.

- Reinitialization: When the evolutionary process converges, the individuals are usually similar and if the iterated threshold becomes negative, the population is restarted in order to provide diversity to the population. Generally, the population is reinitialized with the best individual of the population and mutations of the best individual that usually implies flipping 35% of the genes with some probability.

The proposed RCGA approach for discovering AR from datasets with real values extends the CHC algorithm detailed below. However, it adopts a more conservative reinitialization strategy and a less disruptive crossover operator than the HUX crossover scheme. The pseudocode of the CHC algorithm is as follows:

```
Input: Maximum number of generations ($MaxNumGen$) and threshold for preventing incest ($MinDist$)
Output: Population of the last generation
 CHC()
   $numGen \leftarrow 0$
   Initialize $P(numGen)$
   Initialize $MinDist$
   while ($numGen <= MaxNumGen$)
       Evaluate $P(numGen)$
       $C(numGen) \leftarrow Crossover(P(numGen))$
       Evaluate $C(numGen)$
       $P(numGen+1) \leftarrow SelectBest(P$
       $(numGen) \cup C(numGen))$
       if($P(numGen+1)$ equals $P(numGen)$)
          $MinDist \leftarrow MinDist - 1$
          if ($MinDist < 0$)
             Initialize $P(numGen+1)$
             Initialize $MinDist$
          end if
       end if
       $numGen \leftarrow numGen + 1$
   end while
   return $P(numGen)$
```

In a continuous domain, it is necessary to group certain sets of values that share the same features and, as a consequence, it becomes necessary to be able to express the membership of the values in each group. Intervals have been chosen to represent the membership of such values in this work.

The search of the most appropriate intervals is carried out by means of the proposed RCGA. Thus, the intervals are adjusted to find the AR with high values for both support and confidence as well as other measures used in order to quantify the quality of the rule.

Within the population, each individual constitutes a rule. These rules are then subjected to an evolutionary process in which both crossover operator with incest prevention and reinitialization of the population are applied and, at the end of the process, the fittest individual is designated as the best rule. Moreover, the fitness function has been provided with a set of parameters so that the user can drive the search process depending on the desired rules. The punishment of the covered instances allows the subsequent rules found by the RCGA to try to cover those instances that were still not covered, by means of Iterative Rule Learning (IRL) [34].

The following subsections describe the representation of the individuals, the fitness function, the genetic operators and how the population is restarted.

Table 1
Representation of an individual of the population

i_1 s_1	i_2 s_2	...	i_n s_n
t_1	t_2	...	t_n

3.1. Codification of the individuals

Each gene of an individual represents the upper and lower limit of the intervals of each attribute. The individuals are represented by an array of fixed length n, where n is the number of attributes belonging to the database. Furthermore, the elements are real-valued since the values of the attributes are continuous.

Two structures are available for representing an individual, as is shown in Table 1. Note that all attributes included in the database are depicted in the first row. The limits of the intervals of each attribute are stored in this row, where i_j is the inferior limit of the interval and s_j the superior one.

Nevertheless, not all attributes will be present in the rules that describe an individual. A second row indicating the type of each attribute (shown in the second row of Table 1) has been developed to improve the efficiency. Note that t_i can have three different values: 0 when the attribute does not belong to any individual, 1 when the attribute belongs to the antecedent and 2 when it belongs to the consequent. Therefore, if an attribute is retrieved for a specific rule, it can be achieved by modifying the value equal to 0 of the type by a value equal to 1 or 2. Analogously, an attribute that appears in a rule may be removed by changing the type of the attribute from values 1 or 2 to 0.

3.2. Generation of the initial population

The number of attributes is randomly generated for each individual taking into consideration the desired structure for the rules, the maximum and minimum number of allowed antecedents and consequents and the maximum and minimum number of attributes forming an individual.

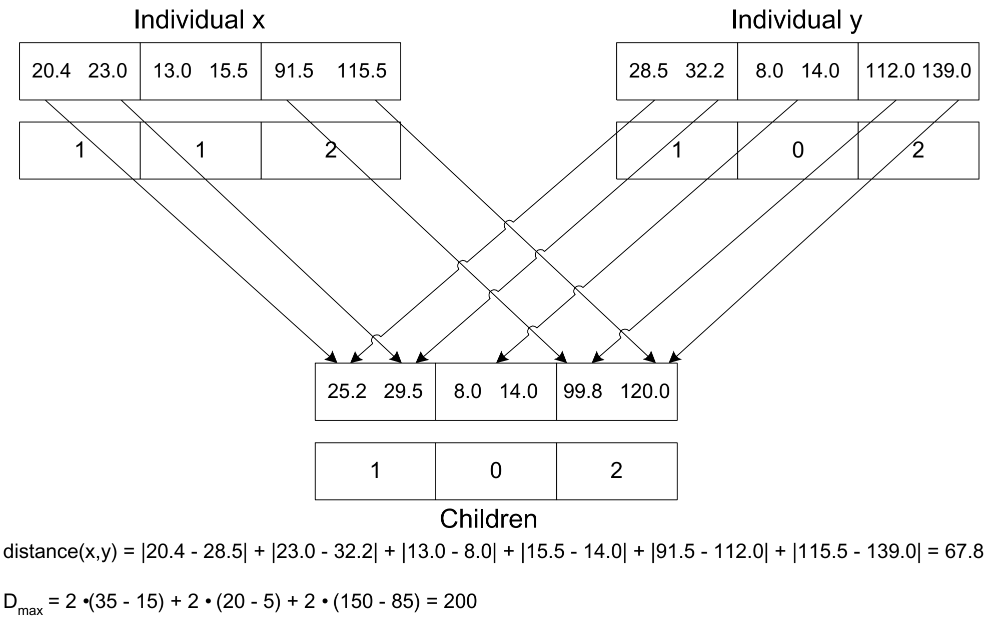

Individual x
20.4 23.0
13.0 15.5
91.5 115.5
1
1
2
Individual y
28.5 32.2
8.0 14.0
112.0 139.0
1
0
2
25.2 29.5
8.0 14.0
99.8 120.0
1
0
2
Children
distance(x,y) = |20.4 - 28.5| + |23.0 - 32.2| + |13.0 - 8.0| + |15.5 - 14.0| + |91.5 - 112.0| + |115.5 - 139.0| = 67.8
D_max = 2 •(35 - 15) + 2 • (20 - 5) + 2 • (150 - 85) = 200

Fig. 1. Crossover and distance for the individuals x and y (attribute ranges: $a_1 \in [15, 35]$, $a_2 \in [5, 20]$ and $a_3 \in [85, 150]$).

It is important to remark that the generation of the interval limits is not arbitrary. On the contrary, it is so that at least one sample of the dataset is covered and that the size of the intervals is less than a given maximum amplitude.

3.3. Crossover operator without incest

Two parent individuals, chosen by means of the roulette selection, are combined to generate a new individual. However, not all parents are crossed, only those parents who differ sufficiently. Thus, the distance between two possible parents is calculated and the parents are only crossed if the distance is greater than $D_{max}/4$ [12] where D_{max} is the maximum possible distance between two individuals and it is defined by:

$$D_{max} = 2 \cdot \sum_{i=1}^{n} (MAX_i - MIN_i) \tag{1}$$

where MAX_i and MIN_i are the maximum and minimum of the range in which the attribute i varies and n is the number of attributes in the dataset. When there are no individuals or potential parents to be crossed due to a distance less than $D_{max}/4$, this threshold is decremented by a percentage of its initial value [5] (10% in this work).

Therefore, the parents to be crossed have to be at least 25% different in order to prevent incest. The distance between two individuals x and y is defined as follows:

$$\textit{distance}(x, y) = \sum_{j \epsilon S} |i_j^x - i_j^y| + |s_j^x - s_j^y| \tag{2}$$

where i_j^x, s_j^x, i_j^y and s_j^y are the inferior and superior limits of the interval of the attribute j which is associated to the individual x and y, respectively. S is the set of attributes and the type for both parents, t_j^x and t_j^i, may or may not coincide as one of them is 1 and the other is 0, or viceversa.

When the type of the attribute t_j, is zero for an individual, the attribute does not form part of the rule represented by the individual and the interval considered to calculate the distance is the range in which the attribute varies. This process is depicted in Fig. 1. However, when the same attribute is an antecedent for one individual and consequent for another one, these two individuals are considered different enough to be crossed and it is not necessary to calculate the distance between them.

Once the distance between the parents has been calculated and it is greater than $D_{max}/4$ the parents are crossed as follows. First, all the attributes associated to each parent are analyzed in order to discover their type. Then, if the same attribute in both parents belonged to the same type of attribute, this type of attribute would be assigned to the descendent and the interval would be obtained generating two random numbers among the limits of the intervals of both parents. Thus, the lower interval would be generated by a random number that belonged to the interval formed by both lower intervals of the parents; the upper interval is analogously calculated. Otherwise, one of the two types would be randomly chosen between both parents, without modifying the intervals of such attribute. Formally, x and y are the two individuals to be crossed and $[l_i^x, s_i^x]$ and $[l_i^y, s_i^y]$ are the intervals in which the attributes vary, respectively. t_i^x is the type of the attribute a_i that belongs to the individual x and finally z is the offspring obtained by the crossover between x and y. Then,

$$\begin{aligned}[l_i^z, s_i^z] &= [random(l_i^x, l_i^y), random(s_i^x, s_i^y)] \\ &\text{if } t_i^x = t_i^y\end{aligned} \tag{3}$$

If $t_i^x \neq t_i^y$, then t_i^z is randomly selected to be equal to t_i^x or t_i^y and the intervals will be equal to that of any parent:

$$[l_i^z, s_i^z] = [l_i^x, s_i^x]\,, \text{if } t_i^z = t_i^x \tag{4}$$

$$[l_i^z, s_i^z] = [l_i^y, s_i^y]\,, \text{if } t_i^z = t_i^y \tag{5}$$

3.4. Reinitialization of the population

The population is restarted when the threshold is set to a negative value in order to introduce diversity in the population and avoid the well-known premature convergence of genetic algorithms. In this work, the population is reinitialized with 35% of the best individuals of the population and mutations of the best individual.

The mutation consists in varying one gene of the individual. The mutation is focused on the intervals, in which three different cases are possible: equiprobable of the upper limit or of the lower limit or of both limits of the interval. To this regard, a random value between 0 and a percentage (10% usually) of the amplitude in which the attribute varies is generated and is added or subtracted to the limit of the interval randomly selected.

3.5. The fitness function

The fitness of each individual allows for determining which are the best candidates to remain in subsequent generations. In order to make this decision, it is preferable to have high support since this fact implies that more samples from the database are covered. Nevertheless, to only take into consideration the support is not enough to calculate the fitness because the algorithm would attempt to enlarge the amplitude of the intervals until the whole domain of each attribute was completed. For this reason, it is necessary to include a measure to limit the growth of the intervals during the evolutionary process. The chosen fitness function to be maximized is:

$$\begin{aligned}f(i) = w_s \cdot sup + w_c \cdot conf - w_r \cdot recov \\ +w_n \cdot nAttrib - w_a \cdot ampl\end{aligned} \tag{6}$$

where sup is the support, $conf$ is the confidence, $recov$ is the number of recovered instances, $nAttrib$ is the number of attributes appearing in the rule, $ampl$ is the average size of attribute intervals that compose the rule and w_s, w_c, w_r, w_n and w_a are weights in order to drive the search depending on the required rules.

The support prefers the rules with a high value of support, that is, rules fulfilled by many instances and the weight w_s can increase or decrease its effect.

The confidence together with the support are the most widely used measures to evaluate the quality of the AR. The confidence is the reliability grade of the rule. High values of w_c may be used when rules without error are desired, and viceversa.

The number of recovered instances is used to indicate that a sample has already been covered by a previous rule. Thus, rules covering different regions of the search space are preferred. The process of penalizing the covered instances is now described. Every time the evolutive process ends and the best individual is chosen as the best rule, the database is processed in order to find those instances already covered by the rule. Hence, each instance has a counter that increases its value by one every time a rule covers it.

Rules with a high number of attributes provide more information but also, in many cases, it is difficult to find rules where a high number of attributes appear. The number of attributes of a rule can be adjusted by means of the weight w_n.

Finally, the amplitude controls the size of the intervals of the attributes that compose the rules and those individuals with large intervals are penalized by means of the factor w_a, which allow the rules to be more or less permissive regarding the amplitude of the intervals.

3.6. The IRL approach

The proposed algorithm is based on the Iterative Rule Learning (IRL) process, whose general scheme is illustrated in Fig. 2.

In each iteration, the CHC takes place. Thus, in each evolutionary process, the individual that represents the best rule is chosen. If the maximum number of rules

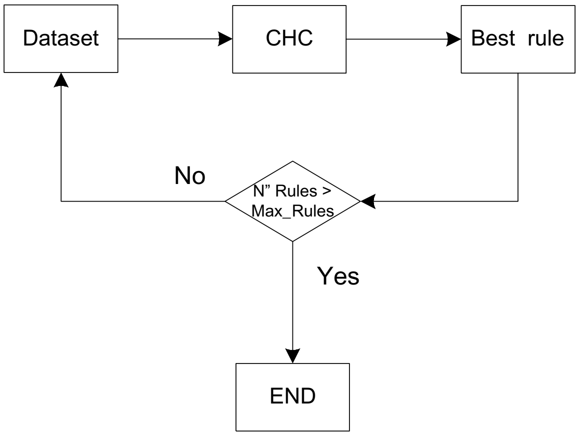


Fig. 2. Scheme of IRL.

to be found is not reached, the samples that have been covered are checked.

The goal of this process is to penalize the instances covered by the best rule in order to cover the remaining instances in subsequent iterations. Subsequently, coverage of search space regions is attempted and the set of rules covers all the domain of the consequent. The iterative process ends when the maximum number of desired rules are found.

Figure 3 illustrates how the proposed algorithm works with the CHC inserted as a crucial step of the IRL process.

First, the population is initialized and the crossover threshold, *MinDist*, is set to prevent incest. In each iteration of the CHC, the population is evaluated and the fitness of each individual is calculated according to (6). Then, the crossover operator without incest is applied to a maximum number of parents (equal to half the population), as described in Section 3.3. Those parents that overlap *MinDist* are crossed to prevent the incest and generate new offspring. Thus, a maximum distance between offspring and parents is guaranteed. Later, the elitist selection takes place. The N best individuals are chosen from the current generation and from the offspring. If no new individuals are created in the current generation, *MinDist* is decremented. In case the threshold was less than zero, the population and the threshold are reinitialized. Finally, the process has to be carried out as many times as the maximum number of generations indicates.

4. Application to synthetic datasets

The proposed algorithm has been applied to the same synthetic dataset used in [4] with the aim of determining if it is possible to find AR with the precise values for the numeric intervals to which each attribute of the rule belongs to.

Table 2
Synthetic sets

A1 ∈ [1, 10] ∧ A2 ∈ [15, 30]
A1 ∈ [15, 45] ∧ A3 ∈ [60, 75]
A2 ∈ [65, 90] ∧ A4 ∈ [15, 45]
A3 ∈ [80, 100] ∧ A4 ∈ [80, 100]

The synthetic dataset is composed of 1000 instances with four numeric attributes each. The selected interval is $[0, 100]$ and all values are uniformly distributed according to Table 2. Note that the amplitude of the intervals is different for each attribute. Furthermore, the datasets have been generated so that the support is 25% and the confidence is 100%. The generation of values out of such datasets are carried out so that no rules better than the ones provided by themselves can exist.

The main parameters of the proposed RCGA are as follows: 100 for the size of the population, 100 for the number of generations and 20 for the number of rules to be obtained. After an experimental study to test the influence of the weights on the rules to be obtained, the weights of the fitness function, 3 for w_s, 1 for w_c, 1.2 for w_r, 0.2 for w_n and 1 for w_a have been chosen.

Table 3 shows the best AR found by the proposed RCGA for the sythentic datasets described previously. The values for support and confidence are also provided, as well as the percentage of covered instances by all rules. It can be noted that the rules have a support of 25% and a confidence of 100%, according to the real values for both measures on the synthetic datasets considered.

These rules have been compared to those shown in Table 4, which have been obtained through a multi-objective differential evolution algorithm (MODENAR) that was recently published in [4]. It can be appreciated that rules obtained by the RCGA share the same support and confidence to those found by MODENAR. Nevertheless, the intervals, to which the numeric attributes belong, determined that RCGA is more precise than MODENAR, since such intervals present the same range and amplitude as those intervals shown in Table 2. In conclusion, it can be stated that the rules found by RCGA are more precise to those found by MODENAR even if the support and confidence are the same.

Different levels of noise have been added to the synthetic datasets in order to validate the efficiency of the RCGA. Thus, values that are not comprised of the in-

Table 3
Association rules found by RCGA

Rule	Support (%)	Confidence (%)	Records (%)
A1 $\in$ [1, 10] $\Longrightarrow$ A2 $\in$ [15, 30]	25	100	100
A1 $\in$ [15, 45] $\Longrightarrow$ A3 $\in$ [60, 75]	25	100	
A3 $\in$ [80, 100] $\Longrightarrow$ A4 $\in$ [80, 100]	25	100	
A2 $\in$ [65, 90] $\Longrightarrow$ A4 $\in$ [15, 45]	25	100	
A2 $\in$ [15, 30] $\Longrightarrow$ A1 $\in$ [1, 10]	25	100	
A3 $\in$ [60, 75] $\Longrightarrow$ A1 $\in$ [15, 45]	25	100	
A4 $\in$ [80, 100] $\Longrightarrow$ A3 $\in$ [80, 100]	25	100	
A4 $\in$ [15, 45] $\Longrightarrow$ A2 $\in$ [65, 90]	25	100	

Table 4
Association rules found by MODENAR

Rule	Support (%)	Confidence (%)	Records (%)
A1 $\in$ [1, 10] $\Longrightarrow$ A2 $\in$ [15, 30]	25	100	100
A1 $\in$ [15, 45] $\Longrightarrow$ A3 $\in$ [60, 75]	25	100	
A3 $\in$ [80, 100] $\Longrightarrow$ A4 $\in$ [80, 98]	25	100	
A2 $\in$ [65, 90] $\Longrightarrow$ A4 $\in$ [15, 43]	25	100	
A2 $\in$ [15, 30] $\Longrightarrow$ A1 $\in$ [1, 10]	25	100	
A3 $\in$ [60, 75] $\Longrightarrow$ A1 $\in$ [15, 45]	25	100	
A4 $\in$ [80, 98] $\Longrightarrow$ A3 $\in$ [80, 100]	25	100	
A4 $\in$ [15, 44] $\Longrightarrow$ A2 $\in$ [65, 89]	25	100	

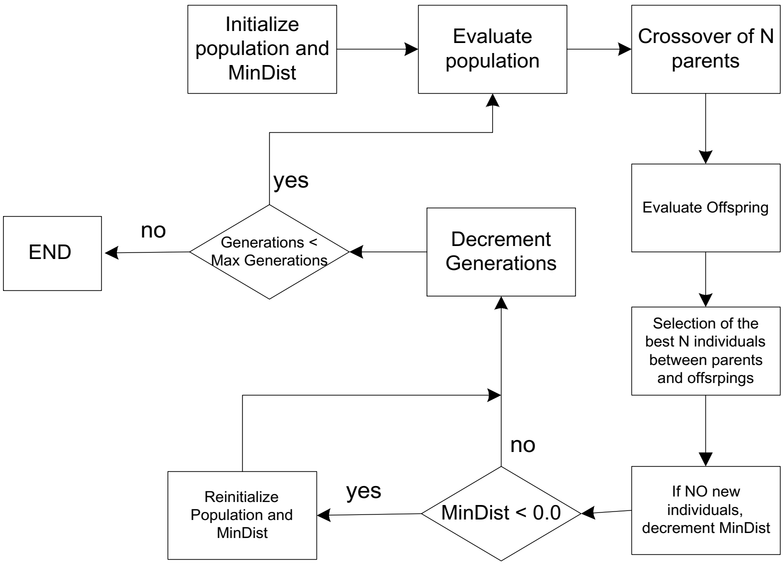

Initialize population and MinDist
Evaluate population
Crossover of N parents
Evaluate Offspring
Selection of the best N individuals between parents and offsrpings
If NO new individuals, decrement MinDist
MinDist < 0.0
yes
Reinitialize Population and MinDist
no
Decrement Generations
Generations < Max Generations
yes
no
END

Fig. 3. Scheme of algorithm CHC.

terval of the second item (A_2) of the dataset have been inserted, that is, a percentage r of instances exist whose second item does not belong to the preset interval. The RCGA has been tested with three different levels of noise (4%, 6% and 8% for the value r).

Table 5 shows the rules obtained by applying RCGA to the different synthetic datasets after the noise addition. The support and confidence values are also provided, as well as the percentage of covered instances by all rules. For the three noise levels, all the extracted rules (but one) have exact intervals. Equally remarkable is that for all noise levels the support in most rules coincides with the real support values which are 24%, 23.5% y 23%, for noise levels of 4%, 6% and 8% respectively.

Table 6 shows the AR, the support values, the confidence and the percentage of covered instances obtained by the MODENAR algorithm for different levels of noise in synthetic datasets. Note that for the case where noise level is 4% the range of the intervals are close

Table 5
Rules mined under different noise level (RCGA)

Mined rules	Support (%)	Confidence (%)	Records (%)
r = 4%			
$A1 \in [1, 10] \Longrightarrow A2 \in [15, 30]$	24.0	96.0	96.0
$A1 \in [15, 45] \Longrightarrow A3 \in [60, 75]$	24.0	96.0	
$A3 \in [80, 100] \Longrightarrow A4 \in [80, 100]$	24.0	96.0	
$A2 \in [65, 90] \Longrightarrow A4 \in [15, 46]$	24.2	95.0	
$A2 \in [15, 30] \Longrightarrow A1 \in [1, 10]$	24.0	100	
$A3 \in [60, 75] \Longrightarrow A1 \in [15, 45]$	24.0	100	
$A4 \in [80, 100] \Longrightarrow A3 \in [80, 100]$	24.0	98.8	
$A4 \in [15, 45] \Longrightarrow A2 \in [65, 90]$	24.0	99.0	
r = 6%			
$A1 \in [1, 10] \Longrightarrow A2 \in [12, 30]$	23.7	94.0	94.0
$A1 \in [15, 45] \Longrightarrow A3 \in [60, 75]$	23.5	94.0	
$A3 \in [80, 100] \Longrightarrow A4 \in [79, 100]$	23.6	92.9	
$A2 \in [65, 90] \Longrightarrow A4 \in [15, 45]$	23.5	92.5	
$A2 \in [15, 30] \Longrightarrow A1 \in [1, 10]$	23.5	100	
$A3 \in [60, 75] \Longrightarrow A1 \in [15, 45]$	23.5	100.0	
$A4 \in [80, 100] \Longrightarrow A3 \in [80, 100]$	23.5	97.5	
$A4 \in [15, 45] \Longrightarrow A2 \in [65, 90]$	23.5	97.5	
r = 8%			
$A1 \in [1, 10] \Longrightarrow A2 \in [15, 30]$	23.0	92.0	92.0
$A1 \in [15, 45] \Longrightarrow A3 \in [60, 75]$	23.0	92.0	
$A3 \in [80, 100] \Longrightarrow A4 \in [80, 100]$	23.0	89.8	
$A2 \in [65, 90] \Longrightarrow A4 \in [15, 45]$	23.0	92.0	
$A2 \in [15, 30] \Longrightarrow A1 \in [1, 10]$	23.0	100.0	
$A3 \in [60, 75] \Longrightarrow A1 \in [15, 45]$	23.0	100.0	
$A4 \in [80, 100] \Longrightarrow A3 \in [80, 100]$	23.0	97.8	
$A4 \in [15, 45] \Longrightarrow A2 \in [65, 90]$	23.0	95.4	

Table 6
Rules mined under different noise level (MODENAR)

Mined rules	Support (%)	Confidence (%)	Records (%)
r = 4%			
$A1 \in [1, 10] \Longrightarrow A2 \in [15, 29]$	24.1	100.0	96.0
$A1 \in [15, 45] \Longrightarrow A3 \in [60, 73]$	24.0	100.0	
$A3 \in [80, 100] \Longrightarrow A4 \in [80, 96]$	23.7	96.7	
$A2 \in [65, 90] \Longrightarrow A4 \in [15, 46]$	24.2	98.3	
$A2 \in [15, 29] \Longrightarrow A1 \in [1, 10]$	24.1	100.0	
$A3 \in [60, 73] \Longrightarrow A1 \in [15, 45]$	24.0	100.0	
$A4 \in [80, 96] \Longrightarrow A3 \in [80, 100]$	23.7	96.7	
$A4 \in [15, 46] \Longrightarrow A2 \in [65, 89]$	24.2	98.3	
r = 6%			
$A1 \in [1, 11] \Longrightarrow A2 \in [14, 31]$	23.3	98.9	94.0
$A1 \in [15, 45] \Longrightarrow A3 \in [56, 73]$	23.6	99.0	
$A3 \in [80, 100] \Longrightarrow A4 \in [84, 95]$	23.3	94.5	
$A2 \in [65, 89] \Longrightarrow A4 \in [14, 49]$	23.8	97.8	
$A2 \in [14, 31] \Longrightarrow A1 \in [1, 11]$	23.3	98.9	
$A3 \in [56, 73] \Longrightarrow A1 \in [15, 45]$	23.6	99.0	
$A4 \in [84, 95] \Longrightarrow A3 \in [80, 100]$	23.3	94.5	
$A4 \in [14, 49] \Longrightarrow A2 \in [65, 89]$	23.8	97.8	
r = 8%			
$A1 \in [1, 11] \Longrightarrow A2 \in [14, 29]$	22.4	97.6	91.8
$A1 \in [15, 45] \Longrightarrow A3 \in [62, 76]$	22.9	98.0	
$A3 \in [79, 100] \Longrightarrow A4 \in [82, 98]$	22.8	93.4	
$A2 \in [65, 90] \Longrightarrow A4 \in [15, 48]$	23.7	95.8	
$A2 \in [14, 29] \Longrightarrow A1 \in [1, 11]$	22.4	97.6	
$A3 \in [62, 76] \Longrightarrow A1 \in [15, 45]$	22.9	98.0	
$A4 \in [82, 98] \Longrightarrow A3 \in [79, 100]$	22.8	93.4	
$A4 \in [15, 48] \Longrightarrow A2 \in [65, 90]$	23.7	95.8	

to the real intervals synthetically generated but are not exact. The support in all cases is close to the real value being equal in just two rules. For this level the proposed algorithm provided better rules with more exact intervals than those provided by MODENAR, which implies better support for such rules. However, the confidence values for rules found by the RCGA are slightly lower than those found by MODENAR.

For a noise of 6%, it can be observed that none of the obtained rules by MODENAR has exactly the same intervals as those used to generate the synthetic dataset. Therefore, the support differs from the real value –equal to 23.5%– for this noise level. Likewise, none of the cases reach a confidence of 100%. Nevertheless, it can be observed in Table 5 that in most cases where the RCGA is applied, exact intervals are obtained. This fact entails confidence values of 100% for some rules and a support of 23.5% for most cases, as opposed to MODENAR.

Analogously for a noise level of 8%, if the rules shown in Tables 5 and 6 are compared, it can be concluded that the behavior of the proposed algorithm with noise is similar to that of previous levels. Consequently, the rules obtained for this level of noise have more precise intervals than those obtained by MODENAR. This improvement entails reaching the real value of the support in the majority of cases. Also, the confidence achieved with the RCGA is 100% for two rules, whereas value has never been fully achieved with the rules obtained from MODENAR.

In conclusion, it can be stated that the RCGA satisfactorily extracted rules for synthetic datasets containing noise, since it showed its ability to overcome different levels of noise, even providing an improvement to the rules provided by MODENAR.

5. Application to atmospheric pollution

The proposed algorithm has been applied in order to discover AR between climatological variables –temperature, humidity, wind direction, wind speed–, the hour of the day and day of the week, and three pollutant agents (ozone, nitrogen monoxide and sulfur dioxide). Therefore, these variables are forced to belong to the consequent. However, the intervals are not previously fixed which differentiates from Apriori and the SD issue.

All variables have been retrieved from a meteorological station placed in the outskirts of Seville city (Spain), providing hourly records of them. It is worth mentioning that Seville is a very hot city that frequently reaches temperatures greater than 40 °C during the summer. The following sections detail the rules obtained for each variable.

5.1. Extracting rules for the ozone

AR have been extracted for ozone (O_3) in two different time periods: from July to August in both 2003 and 2004, which leads to a dataset composed of 1688 instances. The selection of such periods is due to the high concentration of ozone present in the aforementioned summers. For prediction purposes, the climatological time series have been forced to belong to the antecedent and the ozone to the consequent. As a result, a prediction of ozone is achieved on the basis of rules extracted from these variables.

Several experiments have been carried out, in which the main parameters of the GA were as follows: 100 for the size of the population and 100 for the number of generations; 20 for the number of rules to be obtained. After an experimental study to test the influence of the weights on the rules to be obtained, the weights of the fitness function, 0.5 for w_s, 2 for w_c, 6 for w_r, 0.2 for w_n and 7 for w_a have been selected.

The most relevant rules are the ones that identify high concentration of ozone. However, this situation is just under 6.5% of the whole dataset and for this reason, the value w_s has been set low and w_a high, since rules with small amplitudes are desirable. Also, w_r has been set with a high value in order to promote rules that cover instances with high ozone concentration.

The experimentation carried out is detailed in following Tables, in which only the most significative rules are represented. Also, it must be noted that the confidence is the percentage of instances covered by the rule in which only the antecedent is covered.

Table 7 outlines the rules obtained when temperature was the antecedent and ozone the consequent, taking into consideration only those rules whose consequent possesses values of high ozone concentration –typically 170 microgrammes per cubic meter, $[\mu g/m^3]$– to which citizens must be informed of such situations. It can be easily concluded that temperature and ozone are directly related, since an increase in temperature involves an increase in the ozone. Another remarkable feature is the perfect division of the temperature ranges regarding ozone as no overlapping is detected. For temperatures ranging from 35°C to 37°C, ozone values were from 157 $\mu g/m^3$ to 175 $\mu g/m^3$ approximately. Likewise, a temperature in the range [38, 40]°C entails ozone

Table 7
Association rules found by RCGA for temperature (°C) and O_3 ($\mu g/m^3$)

Rule	Support (%)	Confidence (%)
temperature ∈ [34.9, 37.0] $\Longrightarrow O_3$ ∈ [157.7, 175.8]	9.7	19.4
temperature ∈ [38.6, 40.6] $\Longrightarrow O_3$ ∈ [180.0, 202.3]	8.3	22.6
temperature ∈ [42.8, 44.9] $\Longrightarrow O_3$ ∈ [205.8, 223.5]	1.4	66.6

Table 8
Association rules found by RCGA for humidity (%) and O_3 ($\mu g/m^3$)

Rule	Support (%)	Confidence (%)
humidity ∈ [14.0, 20.0] $\Longrightarrow O_3$ ∈ [124.2, 163.7]	4.8	77.7
humidity ∈ [38.6, 40.6] $\Longrightarrow O_3$ ∈ [180.0, 202,3]	5.5	19.5

Table 9
Association rules found by RCGA for wind direction (°) and O_3 ($\mu g/m^3$)

Rule	Support (%)	Confidence (%)
direction ∈ [91.8, 117.1] $\Longrightarrow O_3$ ∈ [144.0, 161.7]	2.1	33.3
direction ∈ [208.6, 233.8] $\Longrightarrow O_3$ ∈ [127.8, 145.5]	13.8	20.0

Table 10
Association rules found by RCGA for wind speed (m/s) and O_3 ($\mu g/m^3$)

Rule	Support (%)	Confidence (%)
speed ∈ [18.1, 20.0] $\Longrightarrow O_3$ ∈ [91.2, 160.8]	29.6	89.5

Table 11
Association rules found by RCGA for hour of the day and O_3 ($\mu g/m^3$)

Rule	Support (%)	Confidence (%)
hour ∈ [11 am, 1:30 pm] $\Longrightarrow O_3$ ∈ [123.5, 141.2]	14.4	17.8
hour ∈ [2 pm, 3:30 pm] $\Longrightarrow O_3$ ∈ [137.3, 157.7]	25.5	30.8
hour ∈ [3 pm, 4:30 pm] $\Longrightarrow O_3$ ∈ [160.3, 178.0]	8.9	21.3
hour ∈ [4 pm, 5:30 pm] $\Longrightarrow O_3$ ∈ [130.7, 166.3]	32.4	38.5
hour ∈ [8 pm, 9:30 pm] $\Longrightarrow O_3$ ∈ [135.9, 153.6]	8.2	19.6

levels of 180 $\mu g/m^3$ and 200 $\mu g/m^3$. Finally, when the temperature reaches 42°C, the ozone has values greater than 200 $\mu g/m^3$. The last rule is of the utmost importance since the confidence obtained is 66%.

Table 8 shows the rules in which ozone reaches its highest values when the humidity is the antecedent. As it can be noted, the humidity triggers considerably high values of ozone when it reaches values between 14% and 20%. Equally remarkable is the second rule in which the ozone exceeds levels of 180 $\mu g/m^3$ when the humidity lies between 38.6% and 40.6%.

Table 9 describes the rules in which ozone had high values when analyzing the wind direction. Ozone levels start to rise when wind direction varies between 210° and 230°. However, the highest ozone concentration found in the atmosphere is when the wind direction is in the range from 90° to 120°, reaching values around 160 $\mu g/m^3$. The precision of both rules is similar, since confidence verges on 25% for both situations.

The rules that relate wind speed and ozone are found in Table 10. With high accuracy, a confidence of 89.5%, ozone reaches moderate values when wind speed is between 18 m/s and 20 m/s.

Table 11 presents hours of the day (in the antecedent) when higher values of ozone (in the consequent) are detected in the atmosphere. According to the obtained rules, it can be concluded that these hours coincide with hours of heavy traffic, that is, the highest concentrations are found from 2 pm to 4:30 pm and from 8 pm to 9:30 pm. These intervals of time are typically associated with the end of schooltime and the working day in Spain. On the contrary, the lowest levels are detected from 11 am to 1:30 pm, the time in which most people are working or studying. All the rules share values of similar confidence, comprising between 20% and 40%.

Table 12 makes reference to the highest concentrations of ozone distributed throughout the days of the week. It can be appreciated that on the first (Monday) and third day of the week, ozone may reach levels greater than 180 $\mu g/m^3$. In addition, Fridays also produce elevated concentrations of ozone. Applying a

Table 12
Association rules found by RCGA for day of the week and O_3 ($\mu g/m^3$)

Rule	Support (%)	Confidence (%)
day ∈ [1, 2] $\Longrightarrow O_3$ ∈ [168.4, 186.1]	8.2	5.6
day ∈ [2, 3] $\Longrightarrow O_3$ ∈ [130.7, 166.3]	9.6	6.9
day ∈ [3, 4] $\Longrightarrow O_3$ ∈ [171.6, 189.3]	6.9	5.0
day ∈ [4, 5] $\Longrightarrow O_3$ ∈ [136.6, 154.2]	13.8	9.9
day ∈ [5, 6] $\Longrightarrow O_3$ ∈ [154.1, 171.8]	7.6	5.1
day ∈ [6, 7] $\Longrightarrow O_3$ ∈ [132.2, 149.9]	13.8	9.3

Table 13
Association rules found by RCGA for temperature (°C) and NO ($\mu g/m^3$)

Rule	Support (%)	Confidence (%)
temperature ∈ [35.7, 37.7] $\Longrightarrow$ NO ∈ [3.0, 8.0]	4.3	88.8
temperature ∈ [38.3, 40,3] $\Longrightarrow$ NO ∈ [3.0, 6.9]	3.9	96.6
temperature ∈ [40.5, 42.6] $\Longrightarrow$ NO ∈ [3.0, 8.2]	1.1	94.1
temperature ∈ [42.9, 45.0] $\Longrightarrow$ NO ∈ [3.0, 6.9]	0.3	100

Table 14
Association rules found by RCGA for humidity (%) and NO ($\mu g/m^3$)

Rule	Support (%)	Confidence (%)
humidity ∈ [14.0, 19.4] $\Longrightarrow$ NO ∈ [3.0, 6.9]	0.5	100
humidity ∈ [36.1, 41.5] $\Longrightarrow$ NO ∈ [3.0, 7.0]	6.7	73.0

Table 15
Association rules found by RCGA for wind direction (°) and NO ($\mu g/m^3$)

Rule	Support (%)	Confidence (%)
direction ∈ [88.1, 114.1] $\Longrightarrow$ NO ∈ [3.0, 6.9]	0.6	81.8
direction ∈ [208.3, 233.5] $\Longrightarrow$ NO ∈ [3.0, 7.0]	6.4	93.2

Table 16
Association rules found by RCGA for wind speed (m/s) and NO ($\mu g/m^3$)

Rule	Support (%)	Confidence (%)
speed ∈ [18.1, 20.0] $\Longrightarrow$ NO ∈ [3.0, 6.9]	3.6	100

similar rationale to that of Table 11, it can be concluded that the highest values are associated with days with heavy traffic, that is, the first and last working days of the week. A slight decrease of the ozone is detected in the middle of the week as well as over the weekend. All rules present similar levels of confidence, within 5% and 10%.

5.2. Extracting rules for nitrogen monoxide

AR have also been extracted for nitrogen monoxide (NO). This pollutant agent is typically generated by the direct combination of nitrogen and oxygen. The analysis of NO levels in the atmosphere is relevant since it directly contributes to the generation of nitrogen dioxide NO_2, which is an extremely oxidant agent resulting from the oxidation of NO. NO_2 is one of the precursors of photochemical smog and it can easily be recognized in big cities due to the reddish coloration of the air.

To carry out the experimentation, the climatological variables used in the previous section (temperature, humidity, wind direction, wind speed, hour of the day and day of the week) have been considered to belong to the antecedent and the nitrogen monoxide to the consequent. It also needs to be mentioned that the parameters as well as the associated weights to each attribute in the fitness function are the same to the ones used for the ozone experimentation.

Furthermore, in order to perform comparisons with results from ozone, rules with antecedents similar to those of ozone have been chosen, that is, rules in which ozone presented high levels of concentration.

Tables 13, 14, 15, 16, 17 and 18 show the rules discovered for the NO and related with temperature, humidity, wind direction, wind speed, hour of the day and day of the week, respectively.

Table 17
Association rules found by RCGA for hour of the day and NO ($\mu g/m^3$)

Rule	Support (%)	Confidence (%)
hour ∈ [12 pm, 1:30 pm] $\Longrightarrow$ NO ∈ [3.0, 6,9]	6.9	83.1
hour ∈ [2 pm, 3:30 pm] $\Longrightarrow$ NO ∈ [3.0, 6.9]	4.1	100
hour ∈ [3 pm, 4:30 pm] $\Longrightarrow$ NO ∈ [3.0, 6.9]	8.3	100
hour ∈ [4 pm, 5:30 pm] $\Longrightarrow$ NO ∈ [3.0, 6.9]	8.2	99
hour ∈ [8 pm, 9:30 pm] $\Longrightarrow$ NO ∈ [3.0, 6.9]	4.1	100

Table 18
Association rules found by RCGA for day of the week and NO ($\mu g/m^3$)

Rule	Support (%)	Confidence (%)
day ∈ [1,2] $\Longrightarrow$ NO ∈ [3.0,7.0]	12.8	88.4
day ∈ [2,3] $\Longrightarrow$ NO ∈ [3.0, 6.9]	12.8	88.4
day ∈ [3,4] $\Longrightarrow$ NO ∈ [3.0, 6.9]	12.6	87.0
day ∈ [4,5] $\Longrightarrow$ NO ∈ [3.0, 6.9]	12.7	87.5
day ∈ [5,6] $\Longrightarrow$ NO ∈ [3.0, 6.9]	13.8	95.3
day ∈ [6,7] $\Longrightarrow$ NO ∈ [3.0, 6.9]	14.2	98.1

Table 19
Association rules found by RCGA for temperature (°C) and SO_2 ($\mu g/m^3$)

Rule	Support (%)	Confidence (%)
temperature ∈ [35.9, 38.0] $\Longrightarrow$ SO_2 ∈ [10.8, 13.0]	1.4	27.5
temperature ∈ [37.9, 40.0] $\Longrightarrow$ SO_2 ∈ [7.0, 10.3]	2.2	53.2
temperature ∈ [42.9, 45.0] $\Longrightarrow$ SO_2 ∈ [3.7, 7.5]	0.3	100

The analysis of the aforementioned Tables reveals that the values for nitrogen monoxide in each case always varies in the interval comprising of 3 $\mu g/m^3$ and 6 $\mu g/m^3$ with a confidence verging on 100% in all cases. These values are typically considered to be very low and, moreover, it remains invariable with independence of the values of the intervals appearing in the antecedent. This feature allows for concluding that nitrogen monoxide cannot be predicted by means of any of the attributes existing in the dataset. That is, these time series are not correlated enough in regard to NO (coefficient of correlation equals 0.1233 in comparison to ozone which equals 0.3777) and, consequently, no useful information can be extracted from their analysis.

Fortunately, these results are logical because, on one side, NO oxidizes and creates NO_2 and, on the other, NO_2 is dissociated in particles of NO and atomic oxygen (O) in presence of solar light. Besides, O reacts with molecular environmental oxygen (O_2) and produces ozone (O_3). Therefore, low values of nitrogen monoxide in the atmosphere are strongly ligated to high values of ozone in the intervals of interest.

5.3. *Extracting rules for sulfur dioxide*

The study of sulfur dioxide in the air is a concerning subject since, apart from being responsible for the generation of sulfuric acid(H_2SO_4), it deeply affects people's health, causing respiratory diseases. The atmospheric SO_2 may oxidize and generate SO_3 and react with humidity (H_2O) by absorption, thus generating thus the molecules of sulfuric acid. These molecules can be dispersed in the air, contributing to the acidification process of the earth and water particles.

Hence, this section describes the experimentation carried out to predict sulfur dioxide (SO_2) from the same climatological time series used in the previous sections. Moreover, all parameters and weights that take part in the fitness function have been set with the same values. The time series are only allowed to appear in the antecedent and sulfur dioxide only in the consequent. Thus, the forecasting is performed on the same basis used for rules discovered in previous sections. Also note that, in order to perform comparisons with results from ozone and nitrogen monoxide, rules with antecedents similar to those of ozone have been chosen, that is, rules in which ozone presented high concentration levels.

Table 19 shows rules relating to the temperature (in the antecedent) and sulfur dioxide (in the consequent), in which no overlapped intervals exist. From its findings, it can be stated that higher the temperature, the less sulfur dioxide there is in the air. This statement may be contradictory since it is reasonable to think that sulfur dioxide increases along with temperature. However, the obtained data are what experts expect since

Table 20
Association rules found by RCGA for humidity (%) and SO_2 ($\mu g/m^3$)

Rule	Support (%)	Confidence (%)
humidity ∈ [14.1, 19.5] $\Longrightarrow SO_2$ ∈ [9.5, 11.6]	0.2	50.0
humidity ∈ [34.2, 39.5] $\Longrightarrow SO_2$ ∈ [3.0, 9.7]	6.2	62.1

Table 21
Association rules found by RCGA for wind direction (°) and SO_2 ($\mu g/m^3$)

Rule	Support (%)	Confidence (%)
direction ∈ [44.4, 69.6] $\Longrightarrow SO_2$ ∈ [3.0, 10.7]	1.3	80.0
direction ∈ [125.4, 150.6] $\Longrightarrow SO_2$ ∈ [3.0, 10.0]	8.4	82.3
direction ∈ [185.2, 210.4] $\Longrightarrow SO_2$ ∈ [3.0, 9.2]	6.1	71.8
direction ∈ [245.7, 270.9] $\Longrightarrow SO_2$ ∈ [3.0, 10.2]	2.8	82.3

Table 22
Association rules found by RCGA for wind speed (m/s) and SO_2 ($\mu g/m^3$)

Rule	Support (%)	Confidence (%)
speed ∈ [0.0, 1.9] $\Longrightarrow SO_2$ ∈ [3.0, 6.9]	8.4	83.4
speed ∈ [17.5, 19.4] $\Longrightarrow SO_2$ ∈ [3.0, 6.9]	2.7	74.5
speed ∈ [25.7, 27.7] $\Longrightarrow SO_2$ ∈ [3.0, 6.9]	0.5	63.6

Table 23
Association rules found by RCGA for hour of the day and SO_2 ($\mu g/m^3$)

Rule	Support (%)	Confidence (%)
hour ∈ [3 am, 4:30 am] $\Longrightarrow SO_2$ ∈ [3.0, 6.3]	2.3	54.8
hour ∈ [11 am, 12:30 pm] $\Longrightarrow SO_2$ ∈ [8.6, 10.8]	1.9	23.4
hour ∈ [12 pm, 1:30 pm] $\Longrightarrow SO_2$ ∈ [11.6, 13.8]	1.4	17.7
hour ∈ [1 pm, 2:30 pm] $\Longrightarrow SO_2$ ∈ [13.6, 15.8]	0.6	14.5
hour ∈ [4 pm, 5:30 pm] $\Longrightarrow SO_2$ ∈ [6.7, 11.8]	2.2	51.6
hour ∈ [8 pm, 9:30 pm] $\Longrightarrow SO_2$ ∈ [11.3, 13.7]	1.3	16.1

the presence of this particle in the air is inversely related to the solar radiation, that is, to the temperature. Therefore, when the temperature increases, the dioxide reacts quicker, generating sulfuric acid and reducing sulfur dioxide concentration. Specifically, when temperature ranges from 35 °C to 38 °C, sulfur dioxide falls in the interval 10-13 μ/m^3 and when temperature reaches 40°C, sulfur dioxide reduces its concentration from 3 $\mu g/m^3$ to 7 $\mu g/m^3$. The last rule is especially reliable due to its high confidence.

Table 20 shows the rules relating to the humidity (in the antecedent) and sulfur dioxide (in the consequent), in which no overlapped intervals exist either. As with temperature, sulfur dioxide is inversely related to humidity. Thus, for humidity between 14% and 19%, sulfur dioxide levels are in the range of 9–11 μ/m^3. Furthermore, when the temperature nears 40 °C, gas concentration is reduced to a level of 3 μ/m^3. The explanation for this phenomenon is similar to that of temperature since the reaction of sulfur dioxide is accelerated by means of humidity absorption, that is, the more humidity, the less sulfur dioxide.

Table 21 presents the rules extracted when using the wind direction as antecedent. In this case, the intervals obtained for sulfur dioxide remain invariable even if the wind direction varies. Consequently, it can be concluded that wind direction does not influence levels of sulfur dioxide in the atmosphere.

Table 22 is devoted to presenting rules extracted when using wind speed as antecedent. As with wind direction, the intervals obtained for the consequent do not vary, independently of the values in the antecedent. Consequently, it can be concluded that the wind speed does not influence levels of sulfur dioxide in the atmosphere.

Table 23 shows the rules for the different hours of a day. As it can be appreciated, high concentrations of sulfur dioxide are concentrated in Spanish rush hours. For instance, when considering the interval from 1 pm to 2:30 pm, the gas reaches values close to 15 $\mu g/m^3$. In comparison, the concentration from 3 am to 4:30 am is no greater than 6 $\mu g/m^3$.

Table 24 describes the rules associated with the day of the week that help sulfur dioxide forecasting. The highest concentrations are on Mondays and Fridays and

Table 24
Association rules found by RCGA for day of the week and SO_2 ($\mu g/m^3$)

Rule	Support (%)	Confidence (%)
day ∈ [1,2] $\Longrightarrow SO_2 \in$ [14.2, 16.3]	0.9	6.7
day ∈ [2,3] $\Longrightarrow SO_2 \in$ [9.8, 12.1]	1.6	11.5
day ∈ [3,4] $\Longrightarrow SO_2 \in$ [12.4, 14.5]	1	15.2
day ∈ [4,5] $\Longrightarrow SO_2 \in$ [8.9, 11.0]	2.2	15.2
day ∈ [5,6] $\Longrightarrow SO_2 \in$ [15.7, 17.8]	0.8	5.5
day ∈ [6,7] $\Longrightarrow SO_2 \in$ [9.4, 11.5]	2.6	18.0

the explanation of this situation is similar to the one provided for the hour of the day. Heavy traffic at the beginning and end of the week causes an increase in the combustion levels, which leads to a higher concentration of this gas in the atmosphere.

6. Conclusions

A new algorithm has been proposed in this work in order to discover quantitative AR. The approach is based on the well-known CHC and works diametrally different as most algorithms do, since it does not discretize the attributes as a first step of the process. Moreover, the algorithm has been evaluated over different datasets. On one hand, synthetic data have been mined and the results were compared with those provided by the MODENAR algorithm, reporting better rules in terms of confidence and support. Additionally, the algorithm has been applied to pollutant agents time series and shown to be effective for forecasting purposes. The use of these kind of tools with such data is, to the best of the authors knowledge, unique. Furthermore, the mined rules agreed with chemical processes associated with these agents.

Acknowledgments

The financial support from the Spanish Ministry of Science and Technology, project TIN2007-68084-C-00, and from the Junta de Andalucía, project P07-TIC-02611, is acknowledged. The authors also want to acknowledge the support of the Regional Ministry for the Environment (*Consejería de Medio Ambiente*) of Andalucía (Spain), that has provided all the pollutant agents time series.

References

[1] R. Agrawal, T. Imielinski and A. Swami, *Mining Association Rules Between Sets of Items in Large Databases*, In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pages 207–216, 1993.

[2] J.S. Aguilar-Ruiz, R. Giráldez and J.C. Riquelme, Natural encoding for evolutionary supervised learning, *IEEE Transactions on Evolutionary Computation* **11**(4) (2007), 466–479.

[3] B. Alatas and E. Akin, Rough particle swarm optimization and its applications in data mining, *Soft Computing* **12**(12) (2008), 1205–1218.

[4] B. Alatas, E. Akin and A. Karci, MODENAR: Multi-objective differential evolution algorithm for mining numeric association rules, *Applied Soft Computing* **8**(1) (2008), 646–656.

[5] J. Alcalá-Fdez, R. Alcalá, M.J. Gacto and F. Herrera, Learning the membership function contexts forming fuzzy association rules by using genetic algorithms, *Fuzzy Sets and Systems* **160**(7) (2009), 905–921.

[6] Y. Aumann and Y. Lindell, A statistical theory for quantitative association rules, *Journal of Intelligent Information Systems* **20**(3) (2003), 255–283.

[7] S. Ayubi, M.K. Muyeba, A. Baraani and J. Keane, An algorithm to mine general association rules from tabular data, *Information Sciences* **179** (2009), 3520–3539.

[8] G. Box and G. Jenkins, *Time Series Analysis: Forecasting and Control*, John Wiley and Sons, 2008.

[9] L. Carro-Calvo, S. Salcedo-Sanz, R. Gil-Pita, A. Portilla-Figueras and M. Rosa-Zurera, An evolutive multiclass algorithm for automatic classifiøcation of high range resolutio radar targets, *Integrated Computer-Aided Engineering* **16**(1) (2009), 51–60.

[10] L. Carro-Calvo, S. Salcedo-Sanz, R. Gil-Pita, A. Portilla-Figueras and M. Rosa-Zurera, An evolutive multiclass algorithm for automatic classifiøcation of high range resoluti radar targets, *Integrated Computer-Aided Engineering* **16**(1) (2009), 51–60.

[11] T.M. Cheng and R.Z. Yan, Integrating messy genetic algorithms and simulation to optimize resource utilization, *Computer-Aided Civil and Infrastructure Engineering* **24**(6) (2009), 401–415.

[12] O. Cordón, S. Dama and J. Santamara, Feature-based image registration by means of the CHC evolutionary algorithm, *Image and Vision Computing* **24** (2006), 525–533.

[13] M. J. del Jesús, P. González, F. Herrera and M. Mesonero, Evolutionary fuzzy rule induction process for subgroup discovery: A case study in marketing, *IEEE Transactions on Fuzzy Systems* **15**(4) (2007), 578–592.

[14] L. Dridi, M. Parizeau, A. Mailhot and J.P. Villeneuve, Using evolutionary optimisation techniques for scheduling water pipe renewal considering a short planning horizon, *Computer-Aided Civil and Infrastructure Engineering* **28**(8) (2008), 625–635.

[15] L. Eshelman, *The CHC Adaptative search algorithm: How to have Safe Search when Engaging in Nontraditional Genetic Recombination*, Morgan Kaufmann, 1991.

[16] E. Georgii, L. Richter, U. Rckert and S. Kramer, Analyzing microarray data using quantitative association rules, *BMC Bioinformatics* **21**(2) (2005), 123–129.

[17] D.L. Godbold and A. Huttermann, *Effects of Acid Rain on Forest Processes*, John Wiley and Sons, 1994.

[18] S. Guha, R. Rastogiand and K. Shim, *CURE: An Efficient Clustering Algorithm for Large Databases*, In Proceedings of ACM SIGMOD International Conference on Management of Data, pages 73–84, 1998.

[19] N. Gupta, N. Mangal, K. Tiwari and Pabitra Mitra, Mining quantitative association rules in protein sequences, *Lecture Notes in Artificial Intelligen* **3755** (2006), 273–281.

[20] X. Jiang and H. Adeli, Neuro-genetic algorithm for nonlinear active control of highrise buildings, *International Journal for Numerical Methods in Engineering* **75**(8) (2008), 770–786.

[21] M. Kaya and R. Alhajj, Genetic algorithm based framework for mining fuzzy association rules, *Fuzzy Sets and Systems* **152**(3) (2005), 587–601.

[22] M. Kaya and R. Alhajj, Utilizing genetic algorithms to optimize membership functions for fuzzy weighted association rules mining, *Applied Intelligence* **24**(1) (2006), 7–152.

[23] H. Kim and H. Adeli, Discrete cost optimization of composite Æoors using a Æoating point genetic algorithm, *Engineering Optimization* **33**(4) (2001), 485–501.

[24] H. Lee, E. Kim and M. Park, A genetic feature weighting scheme for pattern recognition, *Integrated Computer-Aided Engineering* **12**(2) (2007), 161–171.

[25] J. Mata, J.L. Álvarez and J.C. Riquelme, Discovering numeric association rules via evolutionary algorithm, *Lecture Notes in Artificial Intelligence* **2336** (2002), 40–51.

[26] S. Mathakari, P.P. Gardoni, P.P. Agarwal, A. Raich and T. Haukaas, Reliability-based optimal design of electrical transmission towers using multi-objective genetic algorithms, *Computer-Aided Civil and Infrastructure Engineering* **22**(4) (2007), 282–292.

[27] H. Nam, K. Lee and D. Lee, Identiøcation of temporal association rules from time-series microarray data sets, *BMC Bioinformatics* **10**(3) (2009), 1–9.

[28] A. Orriols-Puig, J. Casillas and E. Bernadó-Mansilla, *First Approach Toward on-line Evolution of Association Rules with Learning Classifier Systems*, In Proceedings of the 2008 GECCO Genetic and Evolutionary Computation Conference, pages 2031–2038, 2008.

[29] U. Ruckert, L. Richter and S. Kramer, *Quantitative Association Rules based on Half-Spaces: An Optimization Approach*, In Proceedings of the IEEE International Conference on Data Mining, pages 507–510, 2004.

[30] S.K. Sahua, S. Yipc and D.M. Hollandb, Improved space-time forecasting of next day ozone concentrations in the eastern US, *Atmospheric Environment* **43**(3) (2009), 494–501.

[31] K. Sarma and H. Adeli, Fuzzy genetic algorithm for optimization of steel structures, *Journal of Structural Engineering* **126**(5) (2000), 596–604.

[32] Q. Tong, B. Yan and Y. Zhou, Mining quantitative association rules on overlapped intervals, *Lecture Notes in Artificial Intelligenc* **3584** (2005), 43–50.

[33] M. Vannucci and V. Colla, *Meaningful Discretization of Continuous Features for Association Rules Mining by Means of a Som*, In Proceedings of the European Symposium on Artiøcial Neural Networks, pages 489–494, 2004.

[34] G. Venturini, *SIA: a Supervised Inductive Algorithm with Genetic Search for Learning Attribute Based Concepts*, In Proceedings of the European Conference on Machine Learning, pages 280–296, 1993.

[35] E.I. Vlahogianni, M.G. Karlaftis and J.C. Golias, Spatio-temporal short-term urban traffic flow forecasting using genetically-optimized modular network, *Computer-Aided Civil and Infrastructure Engineering* **22**(5) (2007), 317–325.

[36] D. Wan, Y. Zhang and S. Li, *Discovery Association Rules in Time Series of Hydrology*, In Proceedings of the IEEE International Conference on Integration Technology, pages 653–657, 2007.

[37] X. Yan, C. Zhang and S. Zhang, Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support, *Expert Systems with Applications: An International Journal* **36**(2) (2009), 3066–3076.

[38] Y. Yin, Z. Zhong and Y. Wang, Mining quantitative association rules by interval clustering, *Journal of Computational Information Systems* **4**(2) (2008), 609–616.

4.3. Evolutionary Association Rules for Total Ozone Content Modeling from Satellite Observations

- M. Martínez-Ballesteros, S. Salcedo-Sanz, J. C. Riquelme, C. Casanova-Mateo, J. L. Camacho. *Evolutionary Association Rules for Total Ozone Content Modeling from Satellite Observations*. Chemometrics and Intelligent Laboratory Systems. Vol 109, No. 2, pp 217-227, December 2011 [Martínez-Ballesteros et al., 2011]

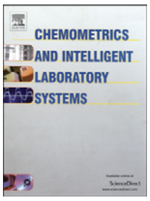
    - Estado: Publicado
    - Índice de Impacto (JCR 2010): 2.222
    - Área de Conocimiento:
        - Automation & Control Systems. Ranking 5 / 60 - Q1
        - Chemistry, Analytical. Ranking 29 / 73 - Q2
        - Computer Science, Artificial Intelligence. Ranking 23 / 108 - Q1
        - Instruments & Instrumentation. Ranking 9 / 61 - Q1
        - Mathematics, Interdisciplinary Applications. Ranking 7 / 93 - Q1
        - Statistics & Probability. Ranking 9 / 110 - Q1

Chemometrics and Intelligent Laboratory Systems 109 (2011) 217–227

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab



Evolutionary association rules for total ozone content modeling from satellite observations

M. Martínez-Ballesteros [a], S. Salcedo-Sanz [b,*], J.C. Riquelme [a], C. Casanova-Mateo [c], J.L. Camacho [d]

[a] Department Languages and Information Systems, Universidad de Sevilla, Seville, Spain
[b] Department of Signal Theory and Communications, Universidad de Alcalá, Madrid, Spain
[c] Department of Applied Physics, Universidad de Valladolid, Valladolid, Spain
[d] Meteorological State Agency of Spain (AEMET), Madrid, Spain

ARTICLE INFO

Article history:
Received 30 March 2011
Received in revised form 5 September 2011
Accepted 23 September 2011
Available online 1 October 2011

Keywords:
Association rules
Evolutionary algorithms
Total Ozone Content (TOC)
Satellite data

ABSTRACT

In this paper we propose an evolutionary method of association rules discovery (EQAR, Evolutionary Quantitative Association Rules) that extends a recently published algorithm by the authors and we describe its application to a problem of Total Ozone Content (TOC) modeling in the Iberian Peninsula. We use TOC data from the Total Ozone Mapping Spectrometer (TOMS) on board the NASA Nimbus-7 satellite measured at three locations (Lisbon, Madrid and Murcia) of the Iberian Peninsula. As prediction variables for the association rules we consider several meteorological variables, such as Outgoing Long-wave Radiation (OLR), Temperature at 50 hPa level, Tropopause height, and wind vertical velocity component at 200 hPa. We show that the best association rules obtained by EQAR are able to accurate modeling the TOC data in the three locations considered, providing results which agree to previous works in the literature.

1. Introduction

Modeling ozone series from satellite observations, past data and its relationship with meteorological variables is an important topic quite often tackled in the literature [1–11]. The interest in modeling ozone series started on the early 70's, when changes into the stratospheric ozone were claimed to be caused by catalytic reactions in the stratosphere that originated losses in the total amount of ozone [12,13]. More specifically, other studies on this topic focused on the role of chlorine [14] and the CFCs [15] in ozone losses at the stratosphere. Those theories were confirmed by the observation of a sharp decrease in the stratospheric ozone levels over Antarctica at the start of the southern spring season in the middle 80's over several polar bases of this continent [16]. A wide review on concepts and history of ozone depletion can be found in [17,18].

In recent years, ozone variation has been related to climate change, so ozone modeling has become an important indicator of deep changes in the atmosphere. That is why very different approaches can be found in the modeling of ozone series in recent bibliography. Specifically, a large amount of works dealing with Total Ozone Content (TOC) of the atmosphere have been published in the last few years, since it seems that variations in these TOC series are a more complete indicator of climate change than only stratospheric ozone series. Thus, there are important works devoted to comparison of different satellite and terrestrial measurements of TOC over different sites [19–21]. The influence of aerosols in total ozone measures is analyzed in [22], where ground and satellite measures are considered. Studies on rare events related to ozone content are studied as well in the literature [23], this includes cases located at the Iberian Peninsula [24]. Also, the modeling of TOC variability has been previously studied, treating different aspects such as its relationship with atmospheric circulation and dynamics or with greenhouse gasses [1,8,25,26]

In this paper, we present an analysis of TOC series modeling in the Iberian Peninsula using Association Rules (ARs) obtained by an evolutionary algorithm. The discovery of ARs is a non-supervised learning and descriptive tool, which explains or summarizes the data, i.e., ARs are used to explore the properties of the data, instead of predicting the class of new data [27]. The aim of ARs mining is discover the presence of pairs (attribute–value), which appear in a dataset with certain frequency, in order to obtain rules that show the existing relationships among the attributes. There exist many algorithms for obtaining ARs from a dataset, such as AIS [28], Apriori [29], and SETM [30]. However, many of these tools that work in continuous domains just discretize the attributes by using a specific strategy and deal with these attributes as if they were discrete, which may lead to poor results in real continuous datasets. Another important class of techniques for ARs discovery is based on evolutionary algorithms (EAs), which have been extensively used for the optimization and adjustment of models in data mining tasks. EAs are search algorithms which generate solutions for optimization problems using

[*] Corresponding author at: Department of Signal Theory and Communications, Universidad de Alcalá, 28871 Alcalá de Henares, Madrid, Spain. Tel.: +34 91 885 6731; fax: +34 91 885 6699.
E-mail address: sancho.salcedo@uah.es (S. Salcedo-Sanz).

0169-7439/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2011.09.011

techniques inspired by natural evolution [31]. They are implemented as a computer simulation in which a population of abstract representations (chromosomes) of candidate solutions (individuals) for an optimization problem evolves toward better solutions. EAs can be used to discover ARs, since they offer a set of advantages for knowledge extraction and specifically for rule induction processes. In this work, the evolutionary algorithm proposed in [32] has been extended and called EQAR (Evolutionary Quantitative Association Rules). EQAR is applied to the ARs extraction to explain TOC data. The new features added improve the AR mining task and result in the TOC modeling in the Iberian Peninsula. We show that the best rules obtained by the EQAR approach are able to accurately model the TOC data in the three locations considered, providing results which agree to previous works in the literature.

The structure of the rest of the paper is as follows: next section presents the available TOC data and input meteorological variables collected for this study, the description of measurement location and their characteristics. In this section we also detail the prediction variables used in the paper. Section 3 describes the main characteristics of the evolutionary algorithm used to obtain the associated rules. Section 4 presents the main results obtained using the associative rules obtained in the explanation of the TOC series in the three locations considered within the Iberian Peninsula. Section 6 closes the paper giving some final conclusions.

2. TOC data over the Iberian Peninsula and prediction variables

Monthly mean satellite measurements of TOMS (Total Ozone Mapping Spectrometer, on board the NASA Nimbus-7 satellite [33,34]) data for the period 1979–1993 have been used in this study. In addition, a group of several meteorological variables has been selected as input (prediction) variables. Specifically: tropopause height (hPa), *TP*, outgoing longwave radiation (Wm^{-2}), OLR, temperature at 50 hPa (K), t_{50} and air vertical velocity at 200 hPa (hPa/s), a_{200}. All these variables have been obtained with a spatial resolution of 2.5 degree latitude × 2.5 degree longitude from NCEP/NCAR reanalysis [35,36]. These four meteorological variables have been selected because all of them have a close relation to TOC concentration:

1. Temperature at 50 hPa (t_{50}): Many studies have shown that maps of total ozone and 50 hPa temperature look very similar, reflecting a very close coupling between them [8,37]. These studies highlight the fact that, as a rule of the thumb, a 10 Dobson Units (DU) change in total ozone corresponds to a 1 K change of 50 hPa temperature. Consequently, this meteorological variable should be correlated with TOC values.
2. Tropopause height (*TP*): The tropopause is a transition layer between the troposphere and the stratosphere. It is not uniformly thick, and it is not continuous from the equator to the poles. As well, tropopause separates the well-mixed ozone poor troposphere and the stratified ozone rich and well mixed stratosphere. This fact gives the key to use the tropopause as a proxy to analyze TOC values. According to [8], in a tropospheric high pressure system, sinking air in the troposphere leads to an adiabatic warming, causing tropopause and low stratosphere air to rise. As a consequence of these vertical movements, the lower stratosphere cools adiabatically and ozone-poor air moves up, decreasing total ozone. The opposite occurs in tropospheric low pressure systems. Thus, it can be said that high tropopause values are correlated with low total ozone and a low tropopause values with high total ozone [7].

   As has been shown in [38], the selected definition of the tropopause, thermal or dynamical, is not critical. Therefore we have decided to use the thermal one, following the standard criterion of the World Meteorological Organization (WMO) to define thermal tropopause: the lowest level at which the lapse rate decreases to $2\,K\cdot km^{-1}$ or less, provided also the average lapse rate between this level and all higher levels within 2 km does not exceed $2\,K\cdot km^{-1}$ [39]. To determine each thermal tropopause from NCAR reanalysis data, we have used the methodology proposed by [40] using European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis (ERA) data.
3. Vertical wind velocity (ω_{200}): As stated before, vertical movements through the tropopause bring ozone-poor air into the stratosphere, attenuating ozone-layer. Conversely, descending air from the upper layers of the stratosphere bring ozone-rich air into the ozone-layer, increasing the density of this layer. In [41,42] the authors have proposed a phenomenological model to explain this idea (another discussion about this model can be found in [43]). Thus, in order to deepen in the correlation between vertical movements and variations in total ozone, ω (total time derivative of the pressure -isobaric coordinate system-) can be used for this purpose. ω negative values will indicate ascending movements, whereas positive omega values will indicate descending movements.
4. Outgoing longwave radiation (OLR): Among other gasses, ozone is one of the most important absorbers in the atmosphere. The ozone molecule has a relatively strong rotation spectrum. The three fundamental ozone vibrational bands occur at wavelengths of 9.066, 14.27, and 9.597 μm, respectively. The very strong 9.597 μm and moderately strong 9.066 μm fundamentals combine to make the well-known 9.6 μm band of ozone [44]. Because this 9.6 μm band is a portion of the infrared region of the electromagnetic spectrum, a direct relationship exists between ozone and the OLR [45] and can be used to characterize TOC.

Our study is focused in three locations of the Iberian Peninsula: Lisbon (38.70 N, 9.10 W), Madrid (40.40 N, 3.70 W) and Murcia (38.00 N, 1.10 W). The four meteorological variables have been calculated using a spatial grid covering the three locations. In addition, having into account the strong correlation between tropopause height and TOC, we have decided to calculate this meteorological variable with a customized grid for each of the three locations, i.e., *TP* variable is divided into four different variables depending on each location (the global *TP* for the Iberian Peninsula (TP_G) and the *TP* variable calculated with a grid centered at each location (TP_W, TP_C and TP_E for Lisbon, Madrid and Murcia, respectively)). Table 1 summarizes the grids used in this study.

3. Methods

In this section we introduce the main AR concepts necessary to follow the rest of the paper, and also the evolutionary algorithm proposed in this work in order to look for ARs.

3.1. Association rules

The massive use of computational processing techniques has revolutionized the scientific research due to the high volume of data

Table 1
Meteorological input variables and associated grid size. "IP" stands for Iberian Peninsula.

Area	Met. variable	Variable name	Grid coordinates
IP	50 hPa temperature	t_{50}	35–42.5 N, 12.5 W–5E
IP	Outgoing Longwave Radiation	OLR	35–42.5 N, 12.5 W–5E
IP	Omega at 200 hPa	ω_{200}	35–42.5 N, 12.5 W–0E
IP	Tropopause height	TP_G	35–42.5 N, 12.5 W–5E
Lisbon	Tropopause height	TP_W	35–42.5 N, 12.5 W–5 W
Madrid	Tropopause height	TP_C	35–42.5 N, 7.5 W–0 W
Murcia	Tropopause height	TP_E	35–42.5 N, 2.5 W–5E

which can be obtained. Data mining is one of the most used instrumental tool for discovering knowledge from transactions. In the field of data mining, the learning of ARs is a popular and well-known research method for discovering interesting relations among variables in large databases [29,46].

Formally, ARs were first defined by Agrawal et al. in [28] as follows. Let $I=\{i_1,i_2,\ldots,i_n\}$ be a set of n items and $D=\{t_1,t_2,\ldots,t_N\}$ a set of N transactions, where each t_j contains a subset of items. Thus, a rule can be defined as $X\Rightarrow Y$, where X,Y I and $X\cap Y=\varnothing$. Finally, X and Y are called antecedent (or left side of the rule) and consequent (or right side of the rule), respectively.

When the domain is continuous, the ARs are known as Quantitative Association Rules (QAR). In this context, let $F=\{F_1,\ldots,F_n\}$ be a set of features, with values in $\mathbb{R}$. Let A and C be two disjunct subsets of F, that is, $A\subset F$, $C\subset F$, and $A\cap C=\varnothing$. A QAR is a rule $X\Rightarrow Y$, in which features in A belong to the antecedent X, and features in C belong to the consequent Y, such that X and Y are formed by a conjunction of multiple boolean expressions of the form $F_i\in[v_1,v_2]$. The consequent Y is usually a single expression. In this proposal, QAR is used, since the domain variable (TOC) is a continuous one.

3.2. Quality measures for association rules

The following paragraphs detail the most popular quality measures used to evaluate an AR. Note that it is very important to have a measure of the quality of a given rule in order to select the best set of rules. In the ARs mining process, probability-based measures that evaluate the generality and reliability of ARs have been selected. In particular, the *support* measure is used to represent the generality of the rule and the *confidence*, the *lift* and the *leverage* are normally used to represent the reliability of the rule [47,48]. The formal definitions of these variables are the following:

- *Support(X)* [48]: The support of an itemset X is defined as the ratio of instances in the dataset that satisfy X. Usually, the support of X is named as the probability of X.

$$sup(X) = P(X) = \frac{n(X)}{N}. \tag{1}$$

where *n(X)* is the number of occurrences of the itemset X in the dataset, and N is the number of instances forming such dataset.

- *Support*$(X\Rightarrow Y)$ [48]: The support of the rule $X\Rightarrow Y$ is the percentage of instances in the dataset that satisfy X and Y simultaneously.

$$sup(X\Rightarrow Y) = P(Y\cap X) = \frac{n(XY)}{N}. \tag{2}$$

where *n(XY)* is the number of instances that satisfy the conditions for the antecedent X and consequent Y simultaneously.

- *Confidence*$(X\Rightarrow Y)$ [48]: The confidence is the probability that instances satisfying X, also satisfy Y. In other words, it is the support of the rule divided by the support of the antecedent.

$$conf(X\Rightarrow Y) = P(X|Y) = \frac{sup(X\Rightarrow Y)}{sup(X)} \tag{3}$$

- *Lift*$(X\Rightarrow Y)$ [49]: Lift or interest is defined as how many times more often X and Y are together in the dataset than expected, assuming that the presence of X and Y in instances is statistically independent. Lifts greater than one involve statistical dependence in simultaneous occurrence of X and Y, in other words, the rule provides successful information about X and Y occurring together in the dataset.

$$lift(X\Rightarrow Y) = \frac{sup(X\Rightarrow Y)}{sup(X)sup(Y)} = \frac{conf(X\Rightarrow Y)}{sup(Y)} \tag{4}$$

- *Leverage*$(X\Rightarrow Y)$ [50]: Leverage measures the proportion of additional cases covered by both X and Y above those expected if X and Y were independent of each other. Leverage takes values inside [−1, 1]. Values equal or under value 0, indicate a strong independence between antecedent and consequent. On the other hand values near 1 are expected for an important association rule. Values above 0 are desirable. In addition, leverage is a lower bound for support, and therefore, optimizing only the leverage guarantees a certain minimum support (contrary to optimizing only the confidence or only the lift).

$$lev(X\Rightarrow Y) = sup(X\Rightarrow Y) - sup(X)sup(Y) \tag{5}$$

- *Accuracy*$(X\Rightarrow Y)$ [48]: Accuracy measures the degree of veracity thus, the degree of fit (matching) between the obtained values and the actual data. An accuracy of 100% means that the measured values are exactly the same as the given values. In the field of mining association rules, accuracy measures the sum of the percentage of instances in the dataset that satisfy the antecedent and the consequent and the percentage of instances in the dataset that do not satisfy neither the antecedent nor the consequent. Accuracy takes values inside [0, 1] and values near 1 are expected for a rule with high quality and veracity.

$$Acc(X\Rightarrow Y) = sup(X\Rightarrow Y) + sup\left(\ddot{A}X\overset{\ddot{A}}{\Rightarrow}Y\right) \tag{6}$$

where ¬ means negation, therefore $sup(\neg X\Rightarrow\neg Y)$ is the percentage of instances in the dataset that do not satisfy X and Y simultaneously.

In most cases, it is enough to focus on a combination of support, confidence, and either lift or leverage to obtain a good measure of the rule "quality". However, how good a rule is for modeling a dataset in terms of usefulness and actionability is a subjective concept, and depends on the particular domain and the business objectives.

For a better understanding of these quality measures, we give a small example, by using a dataset comprising eight instances and three features are shown in Table 2. Consider then two example rules, henceforth called Rule (7) and Rule (8), respectively:

$$F_1\in[32,35]\wedge F_2\in[179,188]\Rightarrow F_3\in[84,94] \tag{7}$$

$$F_1\in[32,35]\wedge F_2\in[179,188]\Rightarrow F_3\in[46,94] \tag{8}$$

In Rule (7), the support of the antecedent is 12.5%, since one instance, t_1, simultaneously satisfy that F_1 and F_2 belong to the intervals [32,35] and [179,188], respectively (one instance out of eight, $sup(X)=0.125$). As for the support of the consequent, $sup(Y)=0.375$ because instances t_1, t_3 and t_7 satisfy that $F_3\in[84,94]$. Regarding the confidence, only one instance t_1 satisfies all the three features (F_1 and F_2 in the antecedent, and F_3 in the consequent) appearing in the rule; in other words, $sup(X\Rightarrow Y)=0.125$. Therefore, $conf(X\Rightarrow Y)=0.125/0.125=1$, that is, the rule has a confidence of 100%. The lift

Table 2
Illustrative dataset.

Instance	F_1	F_2	F_3
t_1	35	183	88
t_2	42	154	47
t_3	37	186	93
t_4	30	199	112
t_5	33	173	83
t_6	24	178	75
t_7	63	177	91
t_8	22	167	60

is $lift(X \Rightarrow Y) = 0.125/(0.125 \cdot 0.375) = 2.66$, the leverage is $lev(X \Rightarrow Y) = 0.125 - (0.125 \cdot 0.375) = 0.078$ and the accuracy is $acc(X \Rightarrow Y) = 0.125 + 0.625 = 0.75$, since $sup(X \Rightarrow Y) = 0.125$, $sup(\neg X \Rightarrow \neg Y) = 0.625$, $sup(X) = 0.125$ and $sup(Y) = 0.375$, as discussed before.

In Rule (8), the support of the antecedent is the same as in Rule (7), i.e. 12.5%, since one instance, t_1, simultaneously satisfy that F_1 and F_2 belong to the intervals [32,35] and [179,188]. However, the support of the consequent is $sup(Y) = 0.875$ because all instances except t_4 satisfy that $F_3 \in [46,94]$. The confidence in this rule is the same that in Rule (7), only one instance satisfies all the three features appearing in the rule, i.e., $sup(X \Rightarrow Y) = 0.125$. Therefore, the confidence value of this rule is also 100%. Regarding the lift or interest, $lift(X \Rightarrow Y) = 0.125/(0.125 \cdot 0.875) = 1.14$, the leverage is $lev(X \Rightarrow Y) = 0.125 - (0.125 \cdot 0.875) = 0.016$ and the accuracy is $acc(X \Rightarrow Y) = 0.125 + 0.125 = 0.25$, since $sup(X \Rightarrow Y) = 0.125$, $sup(\neg X \Rightarrow \neg Y) = 0.125$, $sup(X) = 0.125$ and $sup(Y) = 0.875$, as discussed before.

Note that confidence does not take into account the support of the rule consequent, because the confidence is the same in the two Rules (7) and (8). The lift of the rule should be considered to solve this drawback. Lift or interest measures the degree of dependence between the antecedent and the consequent. The lift of Rule (7) and Rule (8) is 2.66 and 1.14 respectively. Here, lift of Rule (7) is larger than the lift of Rule (8), which corresponds to our intuition that the first rule is more interesting than the second one. Regarding the values of accuracy and leverage are also higher for the Rule (7). Therefore, we can conclude that first rule has better quality, accuracy, interest and strong dependency between the antecedent and consequent than the second one even if they have the same confidence.

3.3. EQAR: an effective evolutionary algorithm for AR searching

As has been previously mentioned, EAs have been quite used to discover ARs, since they offer several advantages for knowledge extraction and specifically for rule induction processes. In [51] the authors proposed an EA to obtain numeric ARs, dividing the process in two phases. Another EA was used in [52] to obtain QAR where the confidence was optimized in the fitness function. In [53] a multi-objective pareto-based EA was presented in which the fitness function was composed by four different objectives. A study of three evolutionary ARs extraction methods was presented in [54] to show their effectiveness for mining ARs in quantitative datasets. Other EAs that use a weighted scheme for the fitness function which involved several evaluation measures of rules were presented in [55] and [32]. The main motivation of these works was to develop an algorithm able to find QAR over datasets with continuous attributes without a previous discretization in the process. In fact, in this paper, we use the basic scheme algorithm proposed in [32] and we extend this approach, henceforth called EQAR (Evolutionary Quantitative Association Rules), with new features in order to improve the ARs mining task. The results were obtained by EQAR in our problem of TOC modeling in the Iberian Peninsula.

EQAR follows the general scheme of the CHC binary-coded evolutionary algorithm proposed by Eshelman in 1991 [56]. The original CHC presents an elitist strategy for selecting the population that will make up the next generation and includes strong diversity in the evolutionary process through mechanisms of incest prevention and a specific operator of crossover called Half Uniform (HUX). Furthermore, the population is re-initialized when its diversity is poor. However EQAR adopts a more conservative re-initialization strategy and a less disruptive crossover operator than the HUX crossover procedure.

The search of the most appropriate intervals is carried out by means of EQAR and the intervals are adjusted to find ARs with high quality. Each individual constitutes a rule in the population. Each gene of an individual represents the limits of the intervals and the type of each attribute to indicate whether it belongs to the antecedent, consequent or not belonging to the rule. Thus, the representation of an individual consists in two data structures as shown in Fig. 1. The upper structure includes all the attributes of the database, where l_j is the lower limit of the range and u_j is the upper limit. The bottom structure indicates the membership of an attribute to the rule represented by an individual. The type of each attribute t_j, can have three values: 0 when the attribute does not belong to the rule, 1 if it belongs to the antecedent of the rule and 2 when it belongs to the consequent part.

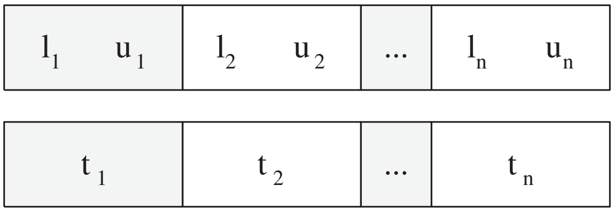


Fig. 1. Representation of an individual of the evolutionary algorithm's population.

An illustrative example of Rule (7) is depicted in Fig. 2. In particular, the rule $F_1 \in [32,35] \wedge F_2 \in [179,188] \Rightarrow F_3 \in [84,94]$ is represented. Note that attributes F_1 and F_2 appear in the antecedent and F_3 in the consequent. Therefore $t_1 = t_2 = 1$ and $t_3 = 2$.

The individuals of the population are subjected to an evolutionary process in which both crossover operator with incest prevention and re-initialization of the population are applied. At the end of this process, the fittest individual is designated as the best rule. Moreover, the fitness function has been provided with a set of parameters so that the user can drive the search process depending on the desired rules. The proposed algorithm is based on the Iterative Rule Learning (IRL) [57]. The punishment of the covered instances allows the subsequent rules found by EQAR to try to cover those instances that were still not covered. General scheme of the IRL is shown in Fig. 4.

In addition to the features of the algorithm described above, new ones have been added in order to improve the performance and the quality of the rules obtained in this specific problem of TOC analysis. The generation of the initial population for each evolutionary process has been modified to help the examples that are covered by a few rules, and also the fitness function has been expanded. These new functionalities of EQAR are detailed in the following subsections.

3.3.1. Generation of the initial population

The generation of the initial population is carried out at the beginning of each evolutionary process. It must be noted that the generation of the rules in EQAR is different to the algorithm proposed in [32], in which the process for generating the initial population was carried out in such a way that at least one randomly chosen sample or instance of the dataset was covered. However, in EQAR the samples of the dataset are not randomly selected but they are selected based on their level of hierarchy. The hierarchy is organized according to the number of rules which cover a sample. Thus, the records are

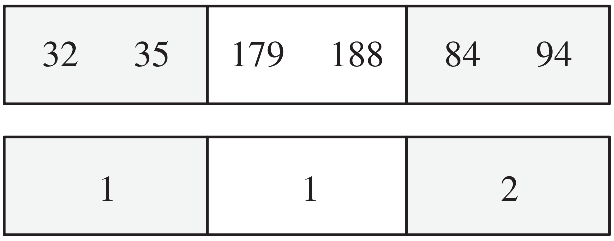


Fig. 2. Representation of Rule (6) example.

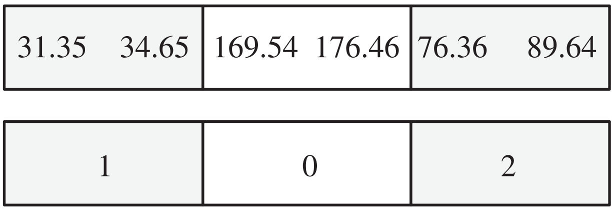


Fig. 3. Representation of an individual example of generation of initial population.

sorted by the number of rules that are covered and the samples covered by few rules have a higher priority.

A sample is selected according to the inverse of the number of rules which cover such sample. Intuitively, the process is similar to roulette selection method where the parents are selected depending on their fitness. In the roulette selection method, a sample is represented by a portion of roulette inversely proportional to the number of rules that cover such sample. Thus, the samples covered by a few rules have a greater portion of the roulette and, therefore, they will be more likely selected. In the first evolutionary process, all samples have the same probability to be selected.

This process for generating the initial population can be described by means of a pseudo-code, as follows.

(1) For all instances of the database the cumulative sum, *totalSum*, of the inverse of the number of rules that cover every instance is calculated.
(2) A random number R between 0 and *totalSum* is generated.
(3) For each instance of database, if the *totalSum* is greater or equal than R, then the current example is selected.

Constraints to generate individuals are given by the following settings:

- number of attributes that belong to rule represented by an individual.
- number of attributes in the antecedents and consequents.
- structure of the rule (attributes fixed or not fixed in consequent).

For a better understanding of the generation of initial population, we describe one example of generation of an individual following the Table 2. For each iteration one instance is selected based on their level of hierarchy. In this case, we have assumed that the algorithm is starting, that is, the first evolutionary iteration of the process, and all instances have the same probability to be selected.

Assuming that the instance t_5 of Table 2 is randomly selected, the values of each attribute are: $F_1=33$, $F_2=173$ and $F_3=83$. In order to generate an individual (one rule), we have to randomly select the number of attributes appearing in the rule and the type and interval of each attribute. We have supposed that the number of attributes chosen is 2 and the type for each attribute is 1 for F_1, 0 for F_2 and 2 for F_3. Then, we have to select a random number between 0 and a maximum amplitude (10%) for generating the intervals for each attribute. The value obtained is added and subtracted to the value corresponding for each attribute to the instance selected (t_5) in Table 2. For example: 5% for F_1, 2% for F_2 and 8% for F_3. Therefore, the intervals of each attribute are $[33\pm(0.05\cdot 33)]$ for F_1, $[173\pm(0.02\cdot 173)]$ for F_2 and $[83m(0.08\cdot 83)]$.

The individual generated is shown in Fig. 3 and the rule obtained is represented as follows:

$$F_1 \in [31.35, 34.65] \Rightarrow F_3 \in [76.36, 89.64] \tag{9}$$

3.3.2. Fitness functions proposed

The fitness of each individual in the evolutionary algorithm allows determining which are the best candidates to remain in subsequent generations. In order to make this decision, its calculation involves several measures that provide information about the rules. In this work, two fitness functions have been designed to maximize different objectives depending on the desired rules. Both are formed by the combination of different measures of association rules but their goals are different.

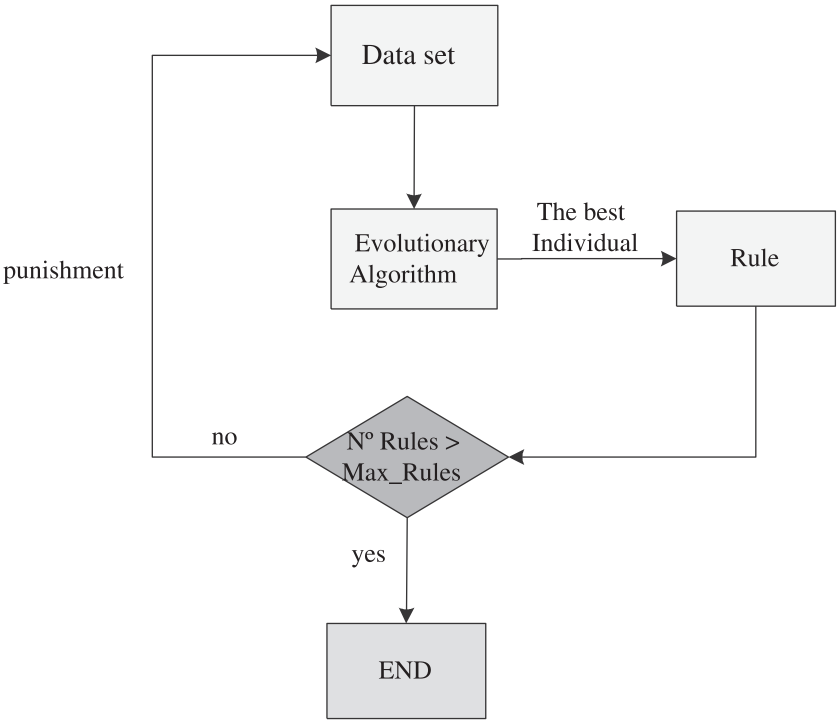


Fig. 4. Scheme of the Iterative Rule Learning algorithm.

Table 3
Ozone quartiles (DU) for each location.

Quartile	Lisbon	Madrid	Murcia
1°	[255.7, 291.2]	[253.9, 291.1]	[259.92, 293.03]
2°	[291.2, 326.7]	[291.1, 328.35]	[293.03, 326.15]
3°	[326.7, 362.2]	[328.35, 365.6]	[326.15, 359.26]
4°	[362.2, 397.7]	[365.6, 402.8]	[359.26, 392.38]

As first fitness function of guide the evolutionary search, we propose the following:

$$f(i) = w_s \cdot sup + w_c \cdot conf - w_r \cdot recov - w_a \cdot ampl \qquad (10)$$

where *sup* is the support of the rule, *conf* is the confidence of the rule, *recov* is the number of recovered instances (it is used to indicate when a sample has already been covered by a previous rule, thus, rules covering different regions of search of space are preferred), *ampl* is the average size of intervals of the attributes belong to the rule and w_s, w_c, w_r and w_a are weights in order to drive the process of rules searching. Note that this function takes into consideration the support and the confidence of the rule. This function is used when QAR with high support and confidence is desired. High values of w_s imply that more samples from the database are covered and high values of w_c imply rules with greater reliability, that is, rules with fewer errors.

Nevertheless, only the support is usually not enough to calculate the fitness, because the algorithm would try to enlarge the amplitude of the intervals until the whole domain of each attribute would be completed to get a great support. For this reason, this fitness function includes a measure to limit the growth of the intervals during the evolutionary process. In addition, this function is able to find rules that cover different regions of the search space because it also includes a measure to negatively affect an instance that has already been covered by a previous rule.

However, this function is not entirely appropriate in some situations because the confidence has some drawbacks. Specifically, confidence does not take into account the support of the rule consequent hence it is not able to detect negative dependencies between items.

For this reason, a new fitness function has been proposed as alternative to the support and confidence measures. The second fitness function to be maximized used by EQAR is given by the following expression:

$$f(i) = w_i \cdot lift + w_l \cdot lev - w_r \cdot recov \qquad (11)$$

where *lift* is the lift or interest of the rule and *lev* is the leverage of the rule and w_i, w_l and w_r are weights in order to drive the process of search of rules.

This function considers lift and leverage measures instead of support and confidence measures. This function is used when QARs with a high lift and high leverage are desired. High values of w_i ensure a degree of dependence between antecedent and consequent. The higher this value, the more likely that the existence of antecedent and consequent together in an instance is not just a random occurrence, but because there is some relationship or dependency between them. High values of w_l guarantee a certain minimum support. Thus, for leverage, values above 0 are desirable, whereas for lift, we want to see values greater than 1. Note that leverage and lift measure similar things, except that leverage measures the proportion of additional cases covered by both antecedent and consequent above those expected if antecedent and consequent were independent of each

Table 4
Association rules for TOC concentration at Lisbon. The "Code" of the rules describes the location and TOC concentration, i.e., L_{m1} stands for Lisbon, medium TOC rule 1, L_{m2} stands for Lisbon, medium TOC rule 2, and L_{h1} stands for Lisbon, high TOC rule 1, and so on.

Code	Rules	TOC (DU)	Scores						Fitness
			Sup(%)	Conf(%)	Ampl(%)	Lift	Lev	Acc(%)	
L_{m1}	$TP_G[12.5, 12.1]$ & $t_{50}[215.9, 216.9]$	[332.4350.5]	5.2	100	10.0	6.0	0.04	88.5	Eq. (11)
L_{m2}	$TP_W[11.5, 11.2]$ & $t_{50}[214.6, 215.7]$ & $OLR[237.5, 256.6]$	[353.2367.4]	3.5	100	9.2	14.4	0.03	96.6	Eq. (11)
L_{h1}	$TP_G[11.4, 10.6]$ & $t_{50}[212.7, 216.8]$ & $OLR[224.4, 239.8]$	[349.0387.8]	10.4	81.8	21.3	4.3	0.08	89.1	Eq. (10)
L_{h2}	$t_{50}[215.1, 217.3]$ & $OLR[231.5, 257.1]$ & $\omega_{200}[-5, 8]$	[349.6,387.8]	10.4	62.1	22.0	3.4	0.07	85.8	Eq. (10)
L_{v1}	$t_{50}[215.6, 217.7]$ & $OLR[231.5, 245.4]$	[369.5,397.7]	4.0	87.5	14.8	12.6	0.04	96.5	Eq. (11)

Table 5
Association rules for TOC concentration at Madrid. The "Code" of the rules describes the location and TOC concentration, i.e., M_{m1} stands for Madrid, medium TOC rule 1, M_{m2} stands for Madrid, medium TOC rule 2, M_{h1} stands for Madrid, high TOC rule 1, and so on.

Code	Rules	TOC (DU)	Scores						Fitness
			Sup(%)	Conf(%)	Ampl(%)	Lift	Lev	Acc(%)	
M_{m1}	$TP_G[14.1, 12.5]$	[285.6,327.8]	23.1	97.6	37.1	1.9	0.11	71.1	Eq. (10)
M_{m2}	$TP_G[12.4, 11.9]$ & $OLR[262.9, 270.3]$ & $t_{50}[215.4, 217.2]$	[329,347.5]	5.8	100	11.2	5.8	0.05	88.4	Eq. (11)
M_{h1}	$TP_C[12.0, 11.3]$ & $t_{50}[215.1, 216.4]$	[334.1, 361]	5.8	100	16.1	4.6	0.05	83.8	Eq. (11)
M_{h2}	$TP_C[11.5, 10.8]$ & $t_{50}[214.4, 216.1]$	[356.3,392.5]	6.9	100	20.6	6.2	0.06	90.8	Eq. (11)
M_{v1}	$TP_G[11.2, 10.6]$ & $OLR[224.4, 245.4]$ & $t_{50}[214.5, 216.8]$	[368.1,402.8]	6.4	91.7	16.9	11.3	0.06	97.7	Eq. (11)

Table 6
Association rules for TOC concentration at Murcia. The "Code" of the rules describes the location and TOC concentration, i.e., U_{m1} stands for Murcia, medium TOC rule 1, U_{m2} stands for Murcia, medium TOC rule 2, U_{h1} stands for Murcia, high TOC rule 1, and so on.

Code	Rules	TOC (DU)	Scores						Fitness
			Sup(%)	Conf(%)	Ampl(%)	Lift	Lev	Acc(%)	
U_{m1}	$TP_E[11.5, 10.7]$ and $t_{50}[212.5, 213.4]$ and $OLR[211.5, 243.6]$	[341.9,359.7]	4.6	100	15.7	6.9	0.04	90.1	Eq. (11)
U_{m2}	$TP_G[11.8, 11.4]$ and $TP_E[12.1, 10.6]$ and $OLR[250.6, 256.6]$	[343.0,354.5]	3.5	100	11.0	9.6	0.03	93.1	Eq. (11)
U_{h1}	$TP_G[11.2, 10.6]$ and $t_{50}[214.6, 216.5]$ and $OLR[214.9, 217.7]$	[343.7,387.5]	8.7	100	22.6	4.7	0.07	87.4	Eq. (11)
U_{v1}	$TP_E[11.2, 10.8]$ and $t_{50}[214.5, 217.7]$	[359.3,392.4]	8.1	100	19.7	7.2	0.07	94.2	Eq. (11)

Table 7
TOC variation obtained with different association rules for different meteorological variables at Lisbon.

Met. variable	Ratio	Rule code				
		L_{m1}	L_{m2}	L_{h1}	L_{h2}	L_{v1}
With TP_G	DU/km	22.5	5.9	19.6	13	17.1
With TP_W	DU/km	15.3	13.9	14.2	8.5	18
With t_{50}	DU/K	6.4	6.6	6.2	6.4	4.3

Table 8
TOC variation obtained with different association rules for different meteorological variables at Madrid.

Met. variable	Ratio	Rule code				
		M_{m1}	M_{m2}	M_{h1}	M_{h2}	M_{v1}
With TP_G	DU/km	13.7	13.8	23.0	21.4	10.8
With TP_C	DU/km	13.4	9.3	22.2	21.4	11.1
With t_{50}	DU/K	9.0	9.1	5.7	6.6	6.0

Table 9
TOC variation obtained with different association rules for different meteorological variables at Murcia.

Met. variable	Ratio	Rule code			
		U_{m1}	U_{m2}	U_{h1}	U_{v1}
With TP_G	DU/km	9.6	8.7	14	14.7
With TP_E	DU/km	9.3	25.1	13.7	18.7
With t_{50}	DU/K	9.4	6.6	6.9	6.3

other. Leverage is also included because lift is susceptible to noise in small databases. Rare itemsets with low probability that per chance occur a few times (or only once) together can produce enormous lift values. In this function, the amplitude of intervals is not included because leverage is inversely proportional to the size of the intervals. If leverage is maximized, we ensure that the intervals of attributes do not extend to the whole domain. This function also includes a measure to negatively affect an instance that has already been covered by a previous rule in order to find rules that cover different regions of the search space.

In conclusion, the first fitness function corresponding to Eq. (10) should be used when rules covering many examples with a high degree of reliability are desired without interesting the degree of dependence between antecedent and consequent of the rule. High confidence and high support could imply interdependence between antecedent and consequent. While the second fitness function corresponding to Eq. (11) should be used when a rule with a high degree of dependence between antecedent and consequent is desired regardless of the number of instances covered by the rule. High lift and high leverage could imply low support.

4. Experimental results

In order to apply the methodology describe above, we have divided TOC values of each location (Lisbon, Madrid and Murcia) into four equal groups or quartiles, each representing a fourth of each TOC data set, i.e., first quartile [0%, 25%], second [26%,50%], third [51%, 75%] and fourth [76%, 100%]. Following this idea, ozone quartiles for each location can be calculated for the period of study considered in this work (1979–1993), as shown in Table 3. Once TOC values have been divided into the four quartiles, the following criteria to set different Ozone concentrations have been used:

- Medium ozone concentration: Rules which ozone values belong to third quartile or a lower quartile.
- High ozone concentration: Rules which ozone values belong to third and fourth quartiles.
- Very high ozone concentration: Rules which ozone values belong only to fourth quartile.

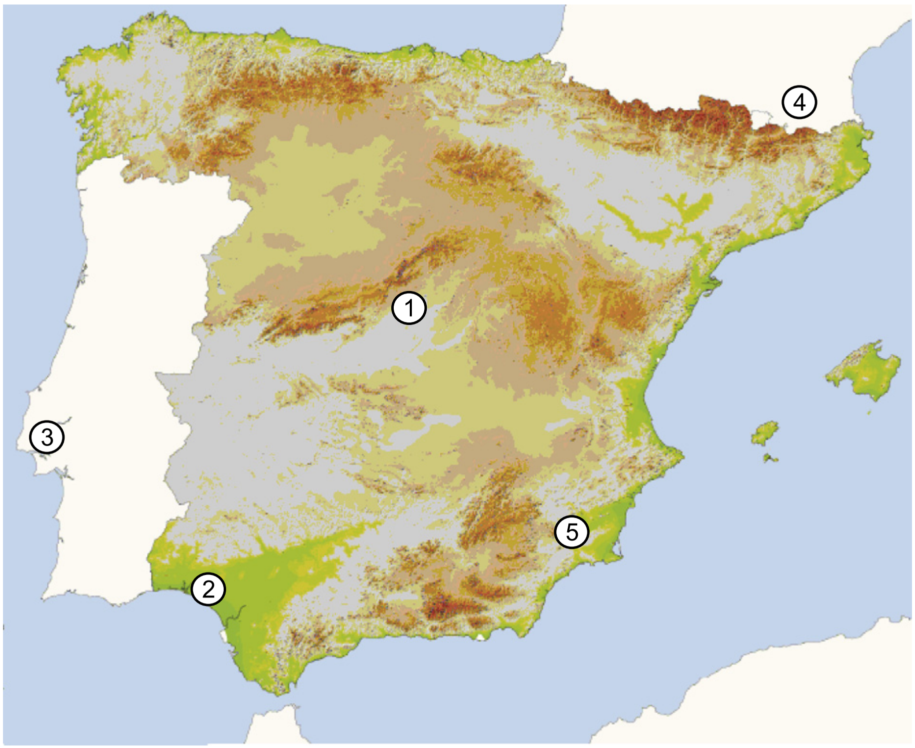


Fig. 5. Location of the different O_3 observing stations considered in the validation process of the results: 1. Madrid, 2. Arenosillo, 3. Lisbon, 4. Montlouis and 5. Murcia.

Table 10
Accuracy of association rules for TOC concentration trained in Lisbon data and tested in Lisbon, Murcia, Madrid, Arenosillo and Montlouis data separately.

Rule code Lisbon	Accuracy (%)				
	Lisbon	Murcia	Madrid	Arenosillo	Montlouis
L_{m1}	88.4	86.7	87.9	87.3	83.8
L_{m2}	96.5	90.2	90.8	91.3	90.2
L_{h1}	89.0	88.4	86.1	88.4	80.9
L_{h2}	82.1	81.5	78.0	81.5	76.3
L_{v1}	95.4	96.5	94.8	95.4	89.0

Table 11
Accuracy of association rules for TOC concentration trained in Murcia and tested in Lisbon, Murcia, Madrid, Arenosillo and Montlouis data separately.

Rule code Murcia	Accuracy (%)				
	Lisbon	Murcia	Madrid	Arenosillo	Montlouis
U_{m1}	85.5	90.2	86.7	87.3	83.2
U_{m2}	88.4	93.1	91.9	91.3	89.0
U_{h1}	83.8	87.3	83.2	89.6	73.4
U_{v1}	89.0	94.2	89.6	92.5	85.0

Thus it is possible to calculate association rules for each location and TOC range (medium, high and very high) and the considered input meteorological variables (given in Table 1).

Applying the EQAR described in Section 3.3, association rules for the TOC[1] and the considered input meteorological variables have been calculated. EQAR has been executed five times for the two fitness function represented by the Eqs. (10) and (11) considering different TOC concentration (medium, high and very high) in each dataset. The best QAR obtained, thus, the rules with support greater than 3% and accuracy greater than 70% have been examined by the group of expert authors in meteorological data. Tables 4 to 6 show the results obtained for the three locations, where we have displayed the rules selected by the expert authors from the best ones according to their meteorological relevance. The column *Scores* describes the values obtained for the different interestingness measures used to qualify the QAR (support, confidence, amplitude, lift, leverage and accuracy). The column *Fitness* indicates the number of equation used as fitness function that has been optimized to obtain each QAR respectively.

It can be shown that most of the QARs provide in these tables were obtained by the second fitness function (Eq. (11)) which shows that the enhancement carried out in EQAR adding a new fitness function to evaluate the individuals in the population provides better and more relevant rules in terms of TOC concentration. Therefore the results obtained by the second fitness function improve the results obtained by the first fitness function (Eq. (10)). This enhancement is due to the first fitness function that only optimizes confidence and support, while the second fitness function optimizes the interest of the rules and the degree of dependence among the attributes belonging to the antecedent and the consequent (TOC concentration in this paper).

It can be appreciated that the scores of the quality measures of the QARs are very good in terms of the confidence and accuracy. Most of them reaches values very close to 100% also the lift and leverage values are greater than 1 and 0 respectively, therefore, the rules obtained present have high accuracy, reliability, and strong dependence among the attributes belonging to the antecedent and the consequent.

It is interesting to observe that all the considered variables form part of the association rules obtained, with some interesting peculiarities: variable ω_{200} is used to explain high and very high TOC concentration in Lisbon, but it does not appear in Murcia nor Madrid. Also in the Murcia case the confidence score is always 100. Note also that, in general, the confidence score is better for Madrid and Murcia than for Lisbon.

In order to discuss the physical correctness of the obtained association rules, we will do a comparison of these rules with results in previous studies. Several previous works have shown the quasi-linear relation that exists between TOC and the meteorological variables tropopause height and temperature at 50 hPa [8,37]. Note that this quasi-linear relationship cannot be found for OLR and ω200. Thus, in order to analyze how good association rules obtained are, ratios DU/km and DU/K have been calculated to be compared against the results obtained by other authors using similar data sources. In Tables 7–9, ratios (in absolute value) for the different tropopause heights (DU/km) and the temperature at 50 hPa (t_{50}) (DU/K) are showed. In each table we show values of TOC variation for two *TP* variables (the global (TP_G) and the *TP* variable calculated with a grid centered in the point (TP_W, TP_C and TP_E) for Lisbon, Madrid and Murcia, respectively).

In general, values for the four tropopause heights considered and temperature at 50 hPa (t_{50}) agree with results in different previous studies. In the case of the tropopause height in [7,37] it is shown that TOC values change approximately between 8 and 25 DU per 1 km increase in tropopause height. Note that the results obtained with the proposed association rules also follow these range of TOC

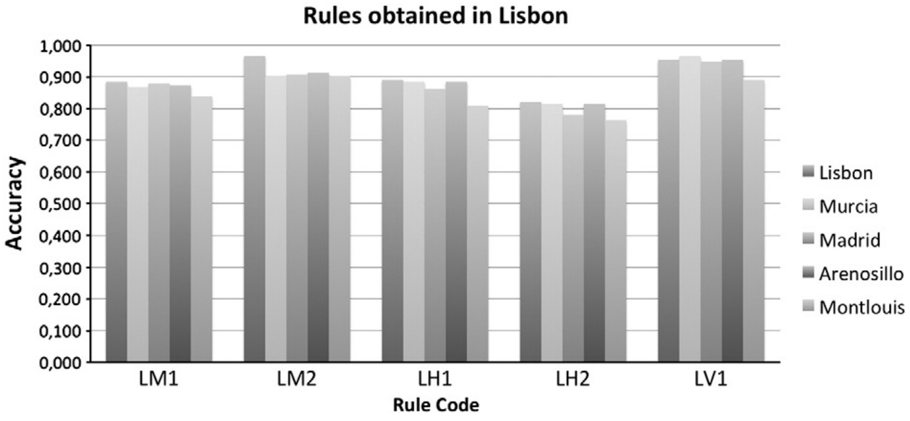


Fig. 6. Accuracy of association rules for TOC concentration trained in Lisbon data and tested in Lisbon, Murcia, Madrid, Arenosillo and Montlouis data separately.

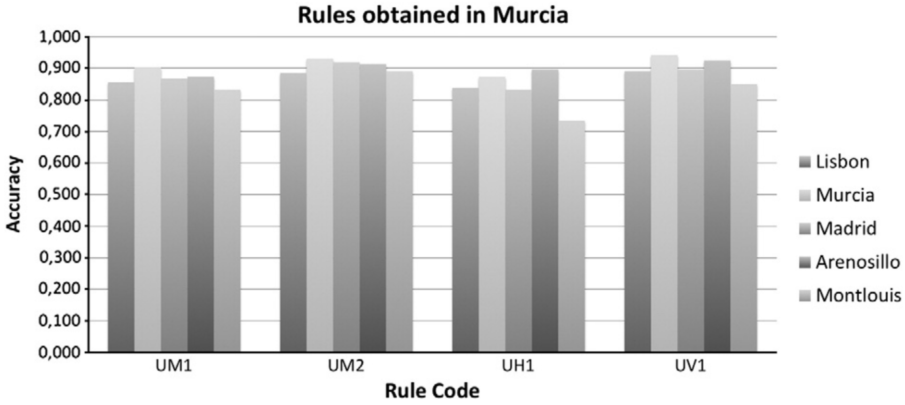


Fig. 7. Accuracy of association rules for TOC concentration trained in Murcia data and tested in Lisbon, Murcia, Madrid, Arenosillo and Montlouis data separately.

Table 12
Accuracy of association rules for TOC concentration trained in Madrid data and tested in Lisbon, Murcia, Madrid, Arenosillo and Montlouis data separately.

Rule code Madrid	Accuracy (%)				
	Lisbon	Murcia	Madrid	Arenosillo	Montlouis
M_{m1}	65.9	67.6	71.1	64.2	72.3
M_{m2}	80.9	81.5	83.8	85.0	81.5
M_{h1}	86.7	85.0	88.4	86.7	86.7
M_{h2}	90.2	91.9	90.8	90.2	87.9
M_{v1}	94.2	96.0	97.1	95.4	93.6

variation against all the TP variables considered at each location, and for all the rules (medium, high and very high TOC) considered. The only result out of this range in our rules is for Lisbon, medium TOC concentration (rule L_{m2}), in which a value of 5.9 DU/km is found for TOC variation against variable TP_G. For the case of TOC variation with variable t_{50} the variation ratio obtained in other works such as [8,37] is that 10 DU change in TOC corresponds to a roughly 1 K change of t_{50}. However, in other studies it is shown that these values can be quite affected by atmospheric variability, El Niño Southern Oscillation (ENSO), Quasi-Bienial Oscillation (QBO) [58], and values of TOC variation with t_{50} of 6, 12 or even 16 DU/K can be found at mid latitudes. Our results show that these thumb rule of 10 DU/K is very well fulfilled in the TOC variation against t_{50} in Madrid (medium TOC concentration) and in Murcia, mainly in the rule U_{m1}. The rest of the cases are not far away from these rule, showing a TOC variation with t_{50} between 6 and 7 DU/K which also agrees with values obtained in other works.

5. Validation of the obtained results

This section describes the tests carried out to validate the results obtained by EQAR in the previous section. In order to confirm that our model has no risk of over-fitting, the rules obtained by EQAR have been tested with six different datasets evaluating the accuracy of the rules. First, the rules obtained for each considered location (Lisbon, Madrid and Murcia) have been tested in the datasets corresponding to five locations (Lisbon, Madrid, Murcia, Arenosillo and Montlouis) separately (Fig. 5 shows these locations, the 3 previously considered and Montlouis and Arenosillo, newly added for this validation study). In addition, the rules have been tested in a dataset containing the TOC values of Arenosillo and Montlouis which consist of 346 instances in total.

Table 10 and Fig. 6 describe the accuracy values corresponding to the rules obtained in the dataset of Lisbon as training data for each level of TOC concentration (medium, high and very high) in the five locations as test data separately. Similarly, Table 11 and Fig. 7 show the accuracy values obtained for the rules belonging to the dataset of Murcia. Finally, Table 12 and Fig. 8 indicate the accuracy values obtained for the rules corresponding to the dataset of Madrid. It can be observed that the accuracy obtained for each rule discovered with the training datasets (Lisbon, Murcia and Madrid) are quite similar when they are applied to other dataset used as test data (Lisbon, Murcia, Madrid, Arenosillo and Montlouis). As can be seen, even in some cases there is greater value of accuracy on the test

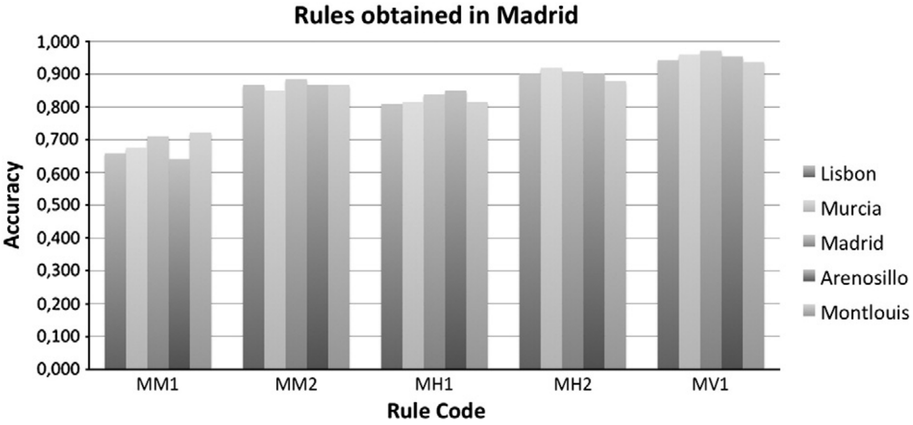


Fig. 8. Accuracy of association rules for TOC concentration trained in Madrid and tested in Lisbon, Murcia, Madrid, Arenosillo and Montlouis data separately.

Table 13
Accuracy of association rules trained in Lisbon, Murcia and Madrid and tested in a dataset containing all TOC concentration of Arenosillo and Montlouis.

Rule code	Accuracy (%)	Rule code	Accuracy (%)	Rule code	Accuracy (%)
Medium		High		Very high	
L_{m1}	85.5	L_{h1}	84.7	L_{v1}	92.2
L_{m2}	90.8	L_{h2}	78.9		
U_{m1}	85.3	U_{h1}	81.5	U_{v1}	88.7
U_{m1}	90.2				
M_{m1}	68.2	M_{h1}	83.2	M_{v1}	94.5
M_{m2}	86.7	M_{h2}	89		

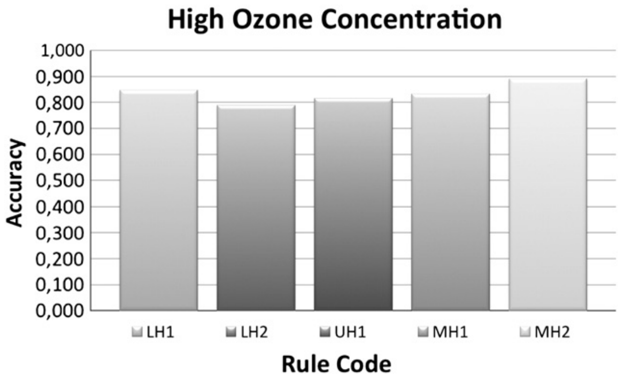


Fig. 10. Accuracy of association rules for high TOC concentration trained in Lisbon, Murcia and Madrid and tested in a dataset containing all location (Lisbon, Murcia. Madrid, Arenosillo and Montlouis).

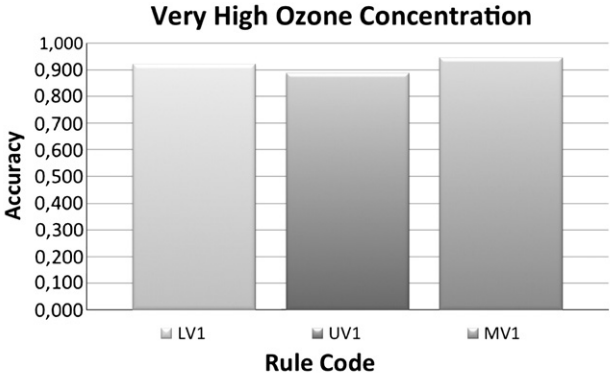


Fig. 11. Accuracy of association rules for very high TOC concentration trained in Lisbon, Murcia and Madrid and tested in a dataset containing all location (Lisbon, Murcia. Madrid, Arenosillo and Montlouis).

data with respect to training data. Therefore the accuracy obtained is stable, since there are no distinct differences among the datasets, which indicates that there is not over-fitting among the rules learned and datasets used as training data.

Accuracy values of QAR tested in a dataset containing the TOC concentration of Arenosillo and Montlouis have been displayed in Table 13. Rules for each location used as training data have been separated by level of TOC concentration and accuracy values are shown graphically in Figs. 9 to 11. Fig. 9 represents the rules discovered in Lisbon, Murcia and Madrid for medium TOC concentration and their accuracy values obtained in the test dataset. Similarly, Fig. 10 describes the rules discovered in Lisbon, Murcia and Madrid for high TOC concentration and their accuracy values obtained in the test dataset. Finally, Fig. 11 shows the rules discovered in Lisbon, Murcia and Madrid for very high TOC concentration and their accuracy values obtained in the test dataset. These tests prove that the rules obtained separately for a particular location are valid for locations analyzed together. The results show accuracy rates above 80% except in one case (M_{m1}), and over 90% in many cases.

After this validation study, we can conclude that there is no over-fitting among rules obtained and the dataset used as training data and we can confirm that the EQAR approach has been really good in terms of the quality of the QAR found because the accuracy values are very high (exceeding 80%) and are very similar in all datasets used as test data.

6. Conclusions

In this paper we have described the application of the EQAR algorithm (Evolutionary Quantitative Association Rules) to a problem Total Ozone Content (TOC) modeling in the Iberian Peninsula. Different improvements in the initial population generation and fitness function have been incorporated to EQAR in order to improve its performance in this problem of TOC modeling. Experimental results have been carried out in TOC data from the Total Ozone Mapping Spectrometer (TOMS) measuring at three locations (Lisbon, Madrid and Murcia) of the Iberian Peninsula. As prediction variables for the association rules we have considered several meteorological variables, such as Outgoing Long-wave Radiation (OLR), Temperature at 50 hPa level, Tropopause height, and wind vertical velocity component at 200 hPa. The results obtained with the EQAR approach have been really good in terms of the quality of the association rules found. Also, the analysis of these rules agrees with the results obtained in other works dealing with TOC modeling, so we can conclude that the use of association rules in TOC modeling could be an interesting analysis method for the future in this and similar problems.

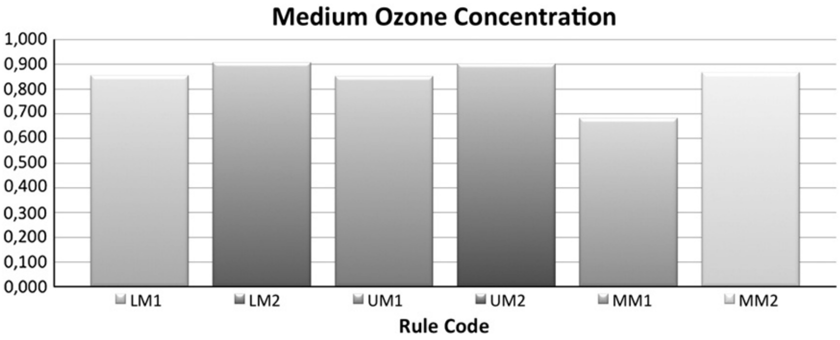


Fig. 9. Accuracy of association rules for medium TOC concentration trained in Lisbon, Murcia and Madrid and tested in a dataset containing all location (Lisbon, Murcia. Madrid, Arenosillo and Montlouis).

References

[1] S. Bronnimann, J. Luterbacher, C. Schmutz, H. Wanner, J. Staehelin, Variability of total ozone at Arosa, Switzerland, since 1931 related to atmospheric circulation indices, Geophysical Research Letters 27 (15) (2000) 2213–2216.

[2] B. Massart, O.M. Kvalheim, L. Stige, R. Aasheim, Ozone forecasting from meteorological variables: part I. Predictive models by moving window and partial least squares regression, Chemometrics and Intelligent Laboratory Systems 42 (1–2) (1998) 179–190.

[3] B. Massart, O.M. Kvalheim, L. Stige, R. Aasheim, Ozone forecasting from meteorological variables: part II. Daily maximum ground-level ozone concentration from local weather forecasts, Chemometrics and Intelligent Laboratory Systems 42 (1–2) (1998) 191–197.

[4] X. Jin, J. Li, C.C. Schmidt, T.J. Schmit, J. Li, Retrieval of total column ozone from imagers onboard geostationary satellites, IEEE Transactions on Geoscience and Remote Sensing 46 (2) (2008) 479–488.

[5] M. Palacios, F. Kirchner, A. Martilli, A. Clappier, F. Martín, M.E. Rodríguez, Summer ozone episodes in the Greater Madrid area. Analyzing the ozone response to abatement strategies by modelling, Atmospheric Environment 36 (2002) 5323–5333.

[6] J.W. Krzyscin, J.L. Borkowski, Variability of the total ozone trend over Europe for the period 1950–2004 derived from reconstructed data, Atmospheric Chemistry and Physics 8 (2008) 2847–2857.

[7] W. Steinbrecht, U. Köhler, K.P. Hoinka, Correlation between tropopause height and total ozone: implication for long-term trends, Journal of Geophysical Research 103 (1998) 19 183–19 192.

[8] W. Steinbrecht, B. Hassler, H. Claude, P. Winkler, R.S. Stolarski, Global distribution of total ozone and lower stratospheric temperature variations, Atmospheric Chemistry and Physics 3 (2003) 1421–1438.

[9] M.A. Barrero, J.O. Grimalt, L. Cantón, Prediction of daily ozone concentration maxima in the urban atmosphere, Chemometrics and Intelligent Laboratory Systems 80 (1) (2006) 67–76.

[10] G. Christakos, A. Kolovos, M.L. Serre, F. Vukovich, Total ozone mapping by integrating databases from remote sensing instruments and empirical models, IEEE Transactions on Geoscience and Remote Sensing 42 (5) (2004) 991–1008.

[11] M. Felipe-Sotelo, L. Gustems, I. Hernández, M. Terrado, R. Tauler, Investigation of geographical and temporal distribution of tropospheric ozone in Catalonia (North-East Spain) during the period 2000–2004 using multivariate data analysis methods, Atmospheric Environment 40 (2004) 7421–7436.

[12] P.J. Crutzen, The influence of nitrogen oxide on the atmospheric ozone content, Quarterly Journal of the Royal Meteorological Society 96 (1970) 320–327.

[13] P.J. Crutzen, Ozone production rates in an oxygen–hydrogen nitrogen oxide atmosphere, Journal of Geophysical Research 76 (1971) 7311–7327.

[14] R.S. Stolarski, R.J. Cicerone, Stratospheric chlorine: a possible sink for ozone, Canadian Journal of Chemistry 52 (1974) 1610–1615.

[15] M.J. Molina, F.S. Rowland, Stratospheric sink for chlorofluoromethanes: chlorine atom catalyzed destruction of ozone, Nature 249 (1974) 810–812.

[16] J.C. Farman, B.G. Gardiner, J.D. Shanklin, Large losses of total ozone in Antarctica reveal seasonal ClOx/NOx interaction, Nature 315 (1985) 207–210.

[17] S. Solomon, Stratospheric ozone depletion: a review of concepts and history, Reviews of Geophysics 37 (3) (1999) 275–316.

[18] United Nations Environment Programme, Environmental Effects Assessment Panel, Environmental effects of ozone depletion and its interactions with climate change: progress report 2005, Photochemical & Photobiological Sciences 5 (13) (2006).

[19] K. Bramstedt, J. Gleason, D. Loyola, W. Thomas, A. Bracher, M. Weber, J.P. Burrows, Comparison of total ozone from the satellite instruments GOME and TOMS with measurements from the Dobson network 1996–2000, Atmospheric Chemistry and Physics 3 (2003) 1409–1419.

[20] A.A. Silva, A quarter century of TOMS total column ozone measurements over Brazil, Journal of Atmospheric and Solar-Terrestrial Physics 69 (12) (2007) 1447–1458.

[21] V. Savastiouk, C.T. McElroy, Brewer spectrophotometer total ozone measurements made during the 1998 middle atmosphere nitrogen trend assessment (MANTRA) Campaign, Atmosphere-Ocean 43 (4) (2005) 315–324.

[22] K.M. Latha, K.V. Badarinath, Impact of aerosols on total columnar ozone measurements a case study using satellite and ground-based instruments, Atmospheric Research 66 (4) (2003) 307–313.

[23] Z. Xiangdong, Z. Xiuji, T. Jie, Q. Yu, C. Chuenyu, A meteorological analysis on a low tropospheric ozone event over Xining, North Western China on 26–27 July 1996, Atmospheric Environment 38 (2) (2004) 261–271.

[24] A. Pérez, I. Aguirre de Cárcer, F. Jaque, Low ozone event at Madrid in November 1996, Journal of Atmospheric and Solar-Terrestrial Physics 64 (3) (2002) 283–289.

[25] C. Appenzeller, A.K. Weiss, J. Staehelin, North Atlantic oscillation modulates total ozone winter trends, Geophysical Research Letters 27 (8) (2000) 1131–1134.

[26] T.G. Shepherd, A.I. Jonsson, On the attribution of stratospheric ozone and temperature changes to changes in ozone-depleting substances and well-mixed greenhouse gases, Atmospheric Chemistry and Physics 8 (2008) 1435–1444.

[27] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006.

[28] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 1993, pp. 207–216.

[29] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, Proceedings of the International Conference on Very Large Databases, 1994, pp. 478–499.

[30] M. Houtsma, A. Swami, Set-Oriented Mining for Association Rules in Relational Databases, IEEE Computer Society, 1995, pp. 25–33.

[31] A.E. Eiben, J.E. Smith, Introduction to Evolutionary Computing, Springer-Verlag, 2003.

[32] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, J.C. Riquelme, Mining quantitative association rules based on evoluationary computation and its application to atmospheric pollution, Integrated Computer-Aided Engineering 17 (3) (2010) 227–242.

[33] R.D. McPeters, P.K. Bhartia, A.J. Krueger, J.R. Herman, B.M. Schlesinger, C.G. Wellemeyer, C.J. Seftor, G. Jaross, S.L. Taylor, T. Swissler, O. Torres, G. Labow, W. Byerly, R.P. Cebula, Nimbus-7 Total Ozone Mapping Spectrometer (TOMS) Data Products User's Guide, NASA Reference Publication, 1996.

[34] M. Anton, J.M. Vilaplana, M. Kroon, A. Serrano, M. Parias, M.L. Cancillo, B.A. de la Morena, The empirically corrected EP-TOMS total ozone data against Brewer measurements at El Arenosillo (Southwestern Spain), IEEE Transactions on Geoscience and Remote Sensing 48 (7) (2010) 3039–3045.

[35] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K.C. Mo, C. Ropelewski, J. Wang, R. Jenne, D. Joseph, The NCEP/NCAR reanalysis 40-year project, Bulletin of the American Meteorological Society 77 (1996) 437–471.

[36] B. Liebmann, C.A. Smith, Description of a complete (Interpolated) outgoing longwave radiation dataset, Bulletin of the American Meteorological Society 77 (1996) 1275–1277.

[37] C. Varotsos, C. Cartalis, A. Vlamakis, C. Tzanis, I. Keramitsoglou, The long-term coupling between column ozone and tropopause properties, Journal of Climate 17 (2004) 3843–3854.

[38] K. Hoinka, H. Claude, U. Kohler, On the correlation between tropopause pressure and ozone above central Europe, Geophysical Research Letters 23 (1996) 1753–1756.

[39] World Meteorological Organization, Meteorology — a three dimensional science, World Meteorological Organization Bulletin 6 (1957) 134–138.

[40] G. Zangl, K. Hoinka, The tropopause in the polar regions, Journal of Climate 14 (14) (2001) 3117–3139.

[41] A. Rabbe, S.H. Larse, Ozone variations in the northern-hemisphere due to dynamic processes in the atmosphere, Journal of Atmospheric and Solar-Terrestrial Physics 54 (9) (1992) 1107–1112.

[42] A. Rabbe, S.H. Larse, 'On the low ozone values over Scandianvia during the winter of 1991–1992', Journal of Atmospheric and Solar-Terrestrial Physics 57 (4) (1995) 367–373.

[43] K. Henriksen, V. Roldugin, Total ozone variations and meteorological processes, Atmospheric Ozone Dynamics, in: Costas Varotsos (Ed.), Series I: Global Environmental Change, vol. 53, Springer Verlag, 1997.

[44] P.W. Menzel, Applications with meteorological satellites, World Meteorological Organization (WMO), Technical Document No. 1078, 2001.

[45] V. Williams, R. Toumi, The correlation between tropical total ozone and outgoing long-wave radiation, Quarterly Journal of the Royal Meteorological Society 127 (2001) 989–1003.

[46] B. Alatas, E. Akin, An efficient genetic algorithm for automated mining of both positive and negative quantitative association rules, Soft Computing 10 (3) (2006) 230–237.

[47] O. Berzal, I. Blanco, D. Sánchez, M. Vila, Ios press measuring the accuracy and interest of association rules: a new framework, , 2001.

[48] L. Geng, H. Hamilton, Interestingness measures for data mining: a survey, ACM Computing Surveys 38 (3) (2006) 9.

[49] S. Brin, R. Motwani, C. Silverstein, Beyond market baskets: generalizing association rules to correlations, Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, vol. 26, no. 2, 1997, pp. 265–276.

[50] G. Piatetsky-Shapiro, Discovery, analysis and presentation of strong rules, Knowledge Discovery in Databases, 1991, pp. 229–248.

[51] J. Mata, J.L. Álvarez, J.C. Riquelme, Discovering numeric association rules via evolutionary algorithm, Lecture Notes in Artificial Intelligence 2336 (2002) 40–51.

[52] X. Yan, C. Zhang, S. Zhang, Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support, Expert Systems with Applications: An International Journal 36 (2) (2009) 3066–3076.

[53] B. Alatas, E. Akin, A. Karci, MODENAR: multi-objective differential evolution algorithm for mining numeric association rules, Applied Soft Computing 8 (1) (2008) 646–656.

[54] J. Alcala-Fdez, N. Flugy-Pape, A. Bonarini, F. Herrera, Analysis of the effectiveness of the genetic algorithms based on extraction of association rules, Fundamenta Informaticae 98 (1) (2010) 1001–1014.

[55] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, J. Riquelme, Quantitative association rules applied to climatological time series forecasting, Intelligent Data Engineering and Automated Learning — IDEAL 2009, ser. Lecture Notes in Computer Science, vol. 5788, 2009, pp. 284–291.

[56] L. Eshelman, The CHC Adaptative Search Algorithm: How to Have Safe Search when Engaging in Nontraditional Genetic Recombination, Morgan Kaufmann, 1991.

[57] G. Venturini, SIA: a supervised inductive algorithm with genetic search for learning attribute based concepts, Proceedings of the European Conference on Machine Learning, 1993, pp. 280–296.

[58] W.J. Randel, J.B. Cobb, Coherent variation of monthly mean total ozone and lower stratospheric temperature, Journal of Geophysical Research 99 (1994) 5433–5447.

Capítulo 5

Otras publicaciones relevantes

Quantitative Association Rules Applied to Climatological Time Series Forecasting

M. Martínez-Ballesteros[1], F. Martínez-Álvarez[2], A. Troncoso[2], and J.C. Riquelme[1]

[1] Department of Computer Science, University of Seville, Spain
{mariamartinez,riquelme}@us.es

[2] Area of Computer Science, Pablo de Olavide University of Seville, Spain
{fmaralv,ali}@upo.es

Abstract. This work presents the discovering of association rules based on evolutionary techniques in order to obtain relationships among correlated time series. For this purpose, a genetic algorithm has been proposed to determine the intervals that form the rules without discretizing the attributes and allowing the overlapping of the regions covered by the rules. In addition, the algorithm has been tested on real-world climatological time series such as temperature, wind and ozone and results are reported and compared to that of the well-known Apriori algorithm.

Keywords: Time series, forecasting, quantitative association rules.

1 Introduction

The prediction of the temporal evolution of variables –time series forecasting– is typically carried out by means of statistical methods. Despite the good performance and inherent simplicity presented by these methods in synthetic data, when applying to real-world time series the results are not as satisfactory as expected due to the non-linear features that such data exhibit.

The existence of other time series correlated with the one under study is an usual phenomenon. In the field of climatological times series, for instance, it is necessary to evaluate time series such as temperature, humidity or atmospheric pressure in order to forecast if it will rain or not. Thus, the problem faced in this work consists in forecasting the behavior of a time series by obtaining association rules among all the existing correlated time series. Concretely, the time series aimed to be forecasted is the tropospheric ozone, which is an atmospheric constituent classed as pollutant when it exceeds a certain threshold. The variation of the concentration of this agent in the air is under continuous analysis, since it is well known the noxious effects that may cause in both human beings and nature [5].

The goal of the association rules extraction process consists, basically, in discovering the presence of pair conjunctions (attribute (A) – value (v)) that appear in a dataset with a certain frequency in order to formulate the rules that display the existing relationship among the attributes. Formally, an association rule is

E. Corchado and H. Yin (Eds.): IDEAL 2009, LNCS 5788, pp. 284–291, 2009.

a relationship between attributes in a database in the way $C_1 \Rightarrow C_2$, where C_1 and C_2 are pair conjunctions such as $A = v$ if $A \in \mathbb{Z}$ or $A \in [v_1, v_2]$ if $A \in \mathbb{R}$. Generally, the antecedent C_1 is formed by a the conjunction of multiple pairs and the consequent C_2 is usually a single pair.

There exist many efficient algorithms that find these rules. However, many researchers are focused on databases with discrete attributes while most real-world databases comprise essentially continuous attributes, as it happens in time series analysis. Moreover, the majority of the tools said to work in the continuous domain just discretize the attributes using a specific strategy and, then, treat these attributes as if they were discrete [6]. The main motivation of this research is to develop a genetic algorithm (GA) able to find association rules over databases with continuous attributes avoiding the discretization as a previous step of the process.

A revision of the recently published literature reveals that the amount of works that provide metaheuristics and search algorithms related to association rules with continuous attributes is scant. Thus, a classifier was presented in [4] with the aim of extracting quantitative association rules over unlabeled data streams. The main novelty of this approach lied on its adaptability to on-line gathered data. An optimization metaheuristic based on rough particle swarm techniques was presented in [1]. In this case, the singularity was the obtention of the values that determine the intervals for the association rules. They also evaluated and tested several new operators in synthetic data. A multi-objective pareto-based GA was presented in [2]. The fitness function was formed by four different objectives: support, confidence, comprehensibility of the rule (aimed to be maximized) and the amplitude of the intervals that forms the rule (intended to be minimized). The work published in [9] presented a new approach based on three novel algorithms: value-interval clustering, interval-interval clustering and matrix-interval clustering. The application of them was found specially useful when mining complex information. Finally, another GA was used in [8] in order to obtain numeric association rules. However, the unique objective to be optimized in the fitness function was the confidence. To fulfill this goal, the authors avoided the specification of the actual minimum support, which is the main contribution of this work.

The rest of the paper is divided as follows. Section 2 provides the methodology used in this work. The results of the approach are discussed in Section 3. Finally, Section 4 describes the achieved conclusions.

2 Description of the Search of Rules

In a continuous domain, it is necessary to group certain sets of values that share same features and, as a consequence, it is required to be able to express the membership of the values to each group. No fixed ranges but intervals of confidence have been chosen to represent the membership of such values in this work. The search of the most appropriate intervals is carried out by means of a GA. Thus, the intervals are adjusted to find the association rules with high values for both support and confidence, together with other measures used in order to quantify the quality of the rule.

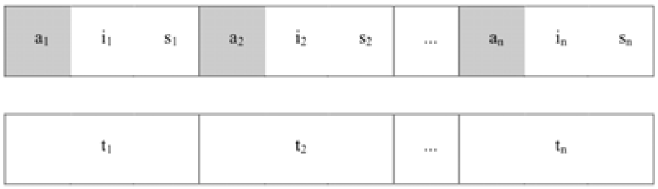


Fig. 1. Representation of an individual of the population

In the population, each individual constitutes a rule. These rules are then subjected to an evolutionary process in which both mutation and crossover operators are applied and, at the end of the process, the individual that presents the best fitness is designated as the best rule. Moreover, the fitness function has been provided with a set of parameters in order to the user can drive the process of search depending on the desired rules. The punishment of the covered instances allows the subsequent rules found with the GA to try to cover those instances that were still uncovered, by means of an Iterative Rule Learning (IRL) [7].

The following subsections detail the general scheme of the algorithm as well as the fitness function, the representation of the individuals and the genetic operators.

2.1 Codification of the Individuals

Each gene of an individual represents the upper and lower limit of the intervals of each attribute. The individuals are represented by a real codification since the values of the attributes are continuous. Each individual is formed by a variable number of attributes, which has to be lower than n, where n is the number of attributes belonging to the database.

Two structures are available for the representation of an individual, as it is shown in Fig. 1. Note that all the attributes included in the database are depicted in the upper structure. The limits of the intervals of each attribute are stored in this structure, where i_i is the inferior limit of the interval and s_i the superior one.

Nevertheless, not all the attributes will be present in the rules that describe an individual. A second structure indicating the type of each attribute, shown in the lower part of the Fig. 1, has been developed with the aim of improving the efficiency. Note that t_i can have three different values: 0 when the attribute does not belong to any individual, 1 when the attribute belongs to the antecedent and 2 when it belongs to the consequent. Therefore, if an attribute is wanted to be retrieved for a specific rule, it can be done by modifying the value equal to 0 of the type by a value equal to 1 or 2.

2.2 Generation of the Initial Population

The number of attributes is randomly generated for each individual taking into consideration the desired structure of the rules, the maximum and minimum number of allowed antecedents and consequents and the maximum and minimum number of attributes forming an individual.

It is important to remark that the generation of the limits of the intervals is not arbitrary. On the contrary, it is performed so that at least one sample of the dataset is covered and that the size of the intervals is less than a given maximum amplitude.

2.3 Genetic Operators

The genetic operators used in the proposed algorithm are: selection, crossover and mutation

1. *Selection.* An elitist strategy is used replicating thus the individual with the best fitness and a roulette selection-based method for the remaining individuals rewarding the best individuals according to their fitness.
2. *Crossover.* Two parent individuals, chosen by means of the roulette selection, are combined to generate a new individual. First, all the attributes associated to each parent are analyzed in order to discover their type. Then, if the same attribute in both parents belonged to the same type of attribute, this type of attribute would be assigned to the descendent and the interval is obtained generating two random numbers among the limits of the intervals of both parents. Thus, the lower interval is generated by means of a random number that belongs to the interval formed by both lower intervals of the parents; the upper interval is analogously calculated. Otherwise, one of the two types would be randomly chosen between both parents, without modifying the intervals of such attribute.
3. *Mutation.* It consists in varying one gene of the individuals. The mutation can be focused on the type of the attribute (antecedent to consequent, consequent to antecedent or antecedent or consequent to null) or on the intervals, in which three different cases are possible: equiprobable mutation of the upper limit, of the lower limit or of both limits of the interval. For this aim, a random value between 0 and the maximum amplitude is generated and it will be added or subtracted to the limit of the interval which is randomly selected.

2.4 The Fitness Function

The fitness of each individual allows to decide which are the best candidates to remain in subsequent generations. In order to make this decision, it is desirable that the support is high since this fact implies that more samples from the database are covered. Nevertheless, to take into consideration only the support is not enough to calculate the fitness because the algorithm would try to enlarge the amplitude of the intervals until the whole domain of each attribute would be completed. For this reason, it is necessary to include a measure to limit the growth of the intervals during the evolutionary process. The chosen fitness function to be maximized is:

$$f(i) = w_s \cdot sup + w_c \cdot conf - w_r \cdot recov + w_n \cdot nAttrib - w_a \cdot ampl \qquad (1)$$

where *sup* is the support, *conf* is the confidence, *recov* is the number of recovered instances, *nAttrib* is the number of attributes appearing in the rule, *ampl* is the average size of intervals of the attributes that compose the rule and w_s, w_c, w_r, w_n and w_a are weights in order to drive the search depending on the required rules.

The support rewards the rules with a high value of support, that is, rules fulfilled by many instances and the weight w_s can increase or decrease its effect.

The confidence together with the support are the most widely measures used in order to evaluate the quality of the association rules. The confidence is the grade of reliability of the rule. High values of w_c may be used when rules without error are desired, and viceversa.

The number of recovered instances is used to indicate that a sample has already been covered by a previous rule. Thus, rules covering different regions of the search space are preferred. The process of punishing the covered instances is now described. Every time the evolutive process ends and the best individual is chosen as the best rule, the database is processed in order to find those instances already covered by the rule. Hence, each instance has a counter that increases its value by one every time a rule covers it.

The rules with a high number of attributes provide more information but also, in many cases, it is difficult to find rules in which a high number of attributes appears. The number of attributes of a rule can be adjusted by means of the weight w_n.

Finally, the amplitude controls the size of the intervals of the attributes that compose the rules and those individuals with large intervals are penalized by means of the factor w_a, which allows the rules be more or less permissive regarding the amplitude of the intervals.

3 Results

The proposed algorithm has been applied to discover association rules between temperature, wind and ozone time series from June 2003 to September 2003. Note that for the prediction task, the temperature and wind are forced to be in the antecedent and the ozone in the consequent, obtaining thus an approximate prediction on the basis of these rules.

Several experiments have been carried out in order to validate the behavior of the proposed operators. The parameters of the algorithm are initially set with default values although a more exhaustive analysis should be performed to establish the optimum set of values. The main parameters of the GA are as follows: 100 for the size of the population, 100 for the number of generations; 20 for the number of rules to be obtained and 0.8 for the mutation probability. The weights of the fitness function are: 2 for w_s, 0.5 for w_c, 1 for w_r, 0.2 for w_n and 1.2 for w_a.

The reason for assigning a high value to the weight w_s is to cover the maximum number of examples by the obtained rules. However, the weight associated to the confidence is lower since it is impossible to obtain rules with a great confidence

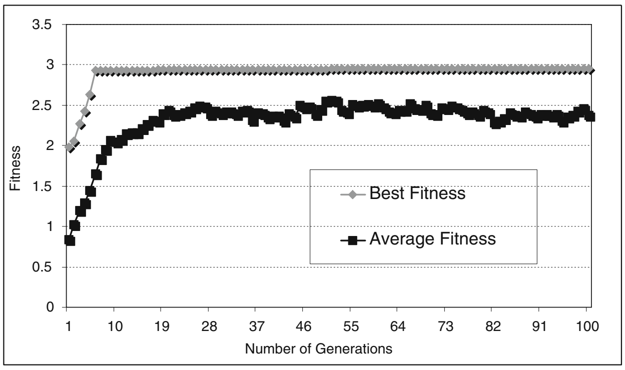


Fig. 2. Evolution of the best rule and the average population

Table 1. Description of the rules found by the proposed GA

Rules	Description
R1	temperature ∈ [28.5,32.2] $\Longrightarrow$ ozone ∈ [112.7,139.3]
R2	temperature ∈ [31.1,34.8] $\Longrightarrow$ ozone ∈ [119.0,145.8]
R3	temperature ∈ [25.3,29.0] $\Longrightarrow$ ozone ∈ [97.7,124.0]
R4	temperature ∈ [22.6,26.3] $\Longrightarrow$ ozone ∈ [103.0,128.7]
R5	temperature ∈ [20.4,23.0] and wind ∈ [13.0,15.5] $\Longrightarrow$ ozone ∈ [91.5,115.5]

due to the existence of a lot of noise in the dataset. The weight associated to the instances covered by other rules as well as the amplitude of the intervals are moderately high in order to penalize the rules whose intervals are too large and are also covering samples already covered by other rules (remind that the goal is to cover all the dataset). The weight associated to the number of attributes has been set with a small value in order to allow the rule to comprise as many attributes as necessary.

Figure 2 shows the evolution of the fittest individual rule and the average of the population throughout the evolutionary process for 10 runs. It can be noticed that the initial set of rules improves its quality all over the generations.

Table 1 shows the five rules selected among the twenty rules found by the GA. It can be noticed that four of them have just two attributes, which are the temperature (in the antecedent) and the ozone (in the consequent). This fact reveals that the temperature provides more information about the ozone than the wind. The fifth selected rule has two attributes in the antecedent –the temperature and the wind– and the ozone in the consequent. Equally remarkable is the possibility of finding rules that have overlapping but covering the whole domain of the consequent, to which the majority of instances belong to. On the other hand, the amplitude of the intervals is similar for all the discovered rules, showing the stability of the proposed algorithm.

Table 2 presents three measures for each rule shown in Table 1. The *Confidence* column indicates the percentage of samples covered by the rule among those samples that only cover the antecedent. The second column, *Covered*, shows the number of samples covered by each rule which is directly related to the support. The *Amplitude* column presents the average amplitude of the intervals for each rule. As it can be observed, the confidence in most cases, despite the small associated weight, is greater than 50% (and even greater than 70% in some cases),

Table 2. Measures for the rules obtained using the GA

Rules	Confidence (%)	Covered	Amplitude
R1	50.8	159	15.1
R2	47.8	143	15.2
R3	54.8	135	15.0
R4	56.0	84	14.7
R5	72.7	8	9.7

Table 3. Description of the rules found by the Apriori algorithm

Rules	Description
R1	temperature $\in$ [24.49,27.42] $\Longrightarrow$ ozone $\in$ [100.76,121.54]
R2	temperature $\in$ [30.35,33.28] $\Longrightarrow$ ozone $\in$ [121.54,142.32]
R3	temperature $\in$ [27.42,30.35] $\Longrightarrow$ ozone $\in$ [100.76,121.54]
R4	wind $\in$ [11.36,14.2] $\Longrightarrow$ ozone $\in$ [121.54,142.32]
R5	wind $\in$ [11.36,14.2] $\Longrightarrow$ ozone $\in$ [100.76,121.54]

Table 4. Measures for the rules obtained using the Apriori algorithm

Rules	Confidence (%)	Covered	Amplitude
R1	42	71	11.8
R2	41	93	11.8
R3	39	88	11.8
R4	29	59	11.8
R5	27	55	11.8

which means that the reached error by the rules can be considered satisfactory. The number of covered samples is much greater with two-attributes rules (more than 100 samples in most cases) than with those with three attributes. Moreover, the average amplitude of the intervals is approximately 14, which is a good result when predicting ozone.

The Apriori algorithm [3] implemented in WEKA has been applied in order to obtain association rules with the purpose of establishing a comparison between the results of the proposed algorithm and that of the Apriori algorithm. Before applying the Apriori algorithm, the temperature, wind and ozone datasets have been discretized because this algorithm only can handle categorical attributes. The rules obtained by this algorithm are shown in Table 3. Note that all the generated rules comprise only two attributes. It can be observed that there are different rules with the same prediction interval for the ozone, e. g., $R1$, R_3 and R_5, and R_2 and R_4. Finally, it is worth noting that these rules do not cover the interval from 90 to 100 in which the dataset has many instances, while the proposed algorithm does.

Table 4 is equivalent to Table 2 but when applying the Apriori algorithm. With regard to the confidence, no rules have values greater than 50% which implies that the rules provide a prediction error greater than that of the proposed algorithm in most cases. The number of instances covered by the rules provided by the proposed approach is greater than that of the Apriori algorithm, obtaining rules with better support. With reference to the average amplitude of the intervals, both algorithms have a similar behavior. Finally, no rules with three attributes have been found using the Apriori algorithm.

4 Conclusions

A new GA has been proposed in this work in order to discover association rules among correlated real-world time series. This algorithm has determined the intervals that form the rules without discretizing the attributes and allowing the overlapping of the regions covered by the rules. When predicting the ozone time series with the new approach, the obtained error is lower than the one provided by the well-known Apriori algorithm, since the confidence of the rules generated by the GA is greater than that of the Apriori algorithm.

Acknowledgments

The financial support from the Spanish Ministry of Science and Technology, project TIN2007-68084-C-00, and from the Junta de Andalucía, project P07-TIC-02611, is acknowledged.

References

1. Alatas, B., Akin, E.: Rough particle swarm optimization and its applications in data mining. Soft Computing 12(12), 1205–1218 (2008)
2. Alatas, B., Akin, E., Karci, A.: MODENAR: Multi-objective differential evolution algorithm for mining numeric association rules. Applied Soft Computing 8(1), 646–656 (2008)
3. Kotsiantis, S., Kanellopoulos, D.: Association rules mining: A recent overview. GESTS International Transactions on Computer Science and Engineering 32(1), 71–82 (2006)
4. Orriols-Puig, A., Casillas, J., Bernadó-Mansilla, E.: First approach toward on-line evolution of association rules with learning classifier systems. In: Proceedings of the 2008 GECCO Genetic and Evolutionary Computation Conference, pp. 2031–2038 (2008)
5. Sahua, S.K., Yipc, S., Hollandb, D.M.: Improved space-time forecasting of next day ozone concentrations in the eastern US. Atmospheric Environment 43(3), 494–501 (2009)
6. Vannucci, M., Colla, V.: Meaningful discretization of continuous features for association rules mining by means of a som. In: Proceedings of the European Symposium on Artificial Neural Networks, pp. 489–494 (2004)
7. Venturini, G.: SIA: a Supervised Inductive Algorithm with genetic search for learning attribute based concepts. In: Brazdil, P.B. (ed.) ECML 1993. LNCS, vol. 667, pp. 280–296. Springer, Heidelberg (1993)
8. Yan, X., Zhang, C., Zhang, S.: Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. Expert Systems with Applications: An International Journal 36(2), 3066–3076 (2009)
9. Yin, Y., Zhong, Z., Wang, Y.: Mining quantitative association rules by interval clustering. Journal of Computational Information Systems 4(2), 609–616 (2008)

MINING QUANTITATIVE ASSOCIATION RULES IN MICROARRAY DATA USING EVOLUTIVE ALGORITHMS

M. Martínez-Ballesteros, C. Rubio-Escudero, J. C. Riquelme
Department of Computer Science, University of Seville, Seville, Spain

F. Martínez-Álvarez
Department of Computer Science, Pablo de Olavide University of Seville, Seville, Spain

Keywords: Data mining, Evolutionary algorithms, Quantitative association rules, MicroArray.

Abstract: The microarray technique is able to monitor the change in concentration of RNA in thousands of genes simultaneously. The interest in this technique has grown exponentially in recent years and the difficulties in analyzing data from such experiments, which are characterized by the high number of genes to be analyzed in relation to the low number of experiments or samples available. In this paper we show the result of applying a data mining method based on quantitative association rules for microarray data. These rules work with intervals on the attributes, without discretizing the data before. The rules are generated by an evolutionary algorithm.

1 INTRODUCTION

The use of massive processing techniques has revolutionized the biotechnology research and it has highly increased the amount of data obtained(Durbin et al., 1998). In particular, microarray technology has revolutionized the biological research due to its ability to monitor changes in RNA concentration in thousands of genes simultaneously (Durbin et al., 1998). Research in molecular biology has traditionally focused on the study gene to gene, but nowadays we are in the genomic era and genes are studied in thousands or even whole genomes. Besides the genes, it is necessary to know the relationships between them.

In this context we present the result of applying a data mining technique, specifically, association rules, to gene expression data from experiments using microarray technology. The aim of this process of mining association rules is to discover the presence of pairs (attribute - value), which appear in a dataset with a certain frequency. This technique is applied to discover associations between genes from microarray datasets, in which gene expression is linked to another gene expression, $Gen1 \Rightarrow Gen2$.

There are many efficient algorithms to find these rules, most focused on discrete data. However in the real world, particularly in the problem to deal in this paper, datasets consists of continuous attributes. In addition, the tools that work in continuous domains just discretize the attributes using a specific strategy and treat these attributes as if they were discrete (Vannucci and Colla, 2004). In this paper, the result of applying a genetic algorithm (GA) is presented. The algorithm can find association rules in databases with continuous attributes from microarray data, avoiding the discretization as a step in the process. The results will show that the rules obtained have been able to successfully characterize the data underlying and also to group relevant genes for the problem studied.

The rest of the paper is divided as follows. Section 2 provides the methodology used in this work. The results obtained by the algorithm developed are discussed in Section 3. Finally, Section 4 describes the achieved conclusions.

2 METHODOLOGY

2.1 Search of Rules

This work is focused on a continuous domain. It is necessary to group the sets of values in intervals to be able to express the membership of the values to each

Martinez Ballesteros M., Rubio Escudero C., C. Riquelme J. and Martíınez-Álvarez F. (2011).
MINING QUANTITATIVE ASSOCIATION RULES IN MICROARRAY DATA USING EVOLUTIVE ALGORITHMS.
In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence*, pages 574-577
DOI: 10.5220/0003152705740577

group. Ranges have not been fixed for intervals. The Genetic Algorithm finds and adjust the most appropriate intervals to find quantitative association rules. Each individual in the population is a rule. The set of rules comprising the population undergo an evolutionary process in which mutation and crossover operators will be applied. The individual with the best fitness at the end of the process represents the best rule. The user can drive the search process because the fitness function has been provided with a set of parameters. Our proposal performs an IRL process (Iterative Rule Learning) (Venturini, 1993) to penalize instances already covered by rules in order to emphasize the covering of instances still not covered.

In the following sections we provide details of the general scheme of the algorithm, the fitness function, representation of individuals and genetic operators.

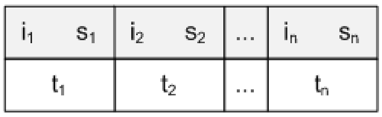
2.2 Scheme of the Algorithm

First, the rules population is initialized and evaluated. All rules are evaluated according to equation 1. Thus, in each iteration the selection operator is applied to select the best rules on the basis of the fitness function. Then, the crossover operator is applied to the selected rules while the population size is not completed. Individuals are randomly selected according to p_mut in order to apply the mutation operator. Finally, the new population is again evaluated by the fitness function and the evolutive process restarts. Note that the process will be repeated as many times as the maximum number of preset generations indicates.

2.3 Individuals Codification

The lower and upper limits of the intervals of each attribute will be represented by the different genes of an individual. Because the attributes are continuous, individuals are represented by an real coding. An individual consists of a not fixed number of attributes less than n, which represents the number of attribute in the database.

The representation of an individual consists in two data structures as shown in Figure 1. The upper structure includes all the attributes of the database, where i_j is the lower limit of the range and s_j is the upper limit. The bottom structure indicates the membership of an attribute to the rule represented by a individual. The type of each attribute t_j, can have three values: 0 when the attribute does not belong to the rule, 1 if it belongs to the antecedent of the rule and 2 when it belongs to the consequent part. If an attribute is wanted to be retrieved for a specific rule, it can be done by modifying the value equal to 0 of the type by a value equal to 1 o or 2 depending on the antecedent or consequent.

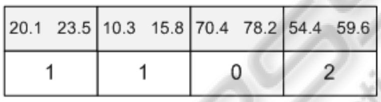


i_1 s_1	i_2 s_2	…	i_n s_n
t_1	t_2	…	t_n

Figure 1: Representation of an individual of the population.

An example of one individual of the population is shown in Figure 2.

$A_1 \in [20.1, 23.5]$ and $A_2 \in [10.3, 15.8] \Longrightarrow A_4 \in [54.4, 59.6]$.



20.1 23.5	10.3 15.8	70.4 78.2	54.4 59.6
1	1	0	2

Figure 2: Example of an individual of the population.

2.4 Initial Population

The number of attributes for each individual is randomly chosen to generate the initial population taking into account the desired format for the rules. In addition, the minimum and maximum numbers in the antecedents and consequents, the minimum and maximum number of attributes that belong to rule represented by an individual are controlled.

2.5 Genetic Operators

The genetic operators implemented in the propose genetic algorithm are: Selection, Crossover and Mutation.

- **Selection.** An elitist strategy replicating the individual with best fitness and a roulette selection-based method for the remaining individuals according to their fitness are used .
- **Crossover.** Two parents are chosen by the roulette selection-based method and they are combined to generate a new individual. The type of all the relevant attributes in both parents are analyzed.

  If both parents have an equal type for the same attribute, it will assigned to offspring. The interval is obtained as a random value between the limits of the intervals of both parents.

  Nevertheless, if both parents have a different type for the same attribute, one of the two parents is randomly chosen and offspring have the intervals and type attribute of the selected parent.
- **Mutation.** Individuals of the population are randomly selected in order to apply the mutation de-

pending on a mutation probability p_{Mut}. The mutation process consists in modifying individuals genes, according to a probability p_{MutGen} in the individuals selected. The mutation can be focused on the attribute type or on the intervals, in which are possible three separate cases: mutation of the upper limit, lower limit or both limits of the interval.

For this aim, a random value between 0 and 10% of the total domain in the attribute is generated and it is added or subtracted to the limit of the interval randomly selected.

2.6 Fitness Function

The fitness function calculation involves several measures that give us information about the rules. In particular, the most representative are the support and confidence that will positively affect the rule. However, it is necessary to take into account a number of factors with negative affect in the quality of the rule. In the amplitude of the intervals, the algorithm may try to extend the intervals to complete the domain of each attribute. For this aim, it is necessary to include a measure limiting growth of the intervals during the evolutive process.

The evaluation function f should be maximized in the evolutionary process is given by the equation . It consists in several parameters which values are calculated from the individual multiplied by a weight to calibrate the effect of each parameter in the overall evaluation.

$$f(i) = w_s \cdot sup + w_c \cdot conf - w_r \cdot recov \\ +w_n \cdot nAttrib - w_a \cdot ampl \tag{1}$$

where sup is the support, $conf$ is the confidence, *recov* is the number of recovered instances, *nAttrib* is the number of attributes in the rule, *ampl* is the average size of intervals of the attributes belong to the rule and w_s, w_c, w_r, w_n and w_a are weights in order to drive the process of search of rules.

The meaning of each parameters in the equation is:

- **Support (sop).** Percentage of records in the dataset covered by the rule.
- **Confidence (conf).** Conditional probability of consequent given the antecedent. Confidence is calculated dividing the support of the rule and the support of the antecedent.
- **Number of Recovered Instances (recub).** It is used to indicate a sample has already been covered by a previous rule. Rules covering different regions of search of space are preferred.
- **Number of Attributes (natrib).** Number of attributes (genes) belong to the rule (individual).
- **Amplitude (ampl).** Average of intervals size of the attributes belong to the rule.

3 RESULTS

The results of applying the algorithm proposed in Section 2 to a dataset acquired from a microarray experiment related to inflammation and immune response are presented. Inflammation is a critical process because the human body uses to protect itself from infections and lesions. In this experiment, conducted at the University of St. Louis, Missouri(Calvano et al., 2005), the blood of eight volunteers is analyzed, four treated with a toxin produces an inflammatory process and 4 with placebo. Samples has been taken at 6 time points over 24 hours, obtaining a total of 48 microarrays.

The algorithm was tested with the following parameters of AG: 100 for the size of the population, 100 for the number of generations, 20 for the number of rules to obtain, 0.8 for the mutation probability p_{Mut} of the individuals and 0.2 for the mutation probability p_{MutGen} of each gene in the individual. The fitness function weights are: 1 for w_s, 0.5 for w_c, 0.3 for w_r, 0.1 for w_n, and 0.1 for w_a. The reason to assign a high value to the weight w_s is to cover the maximum number of examples obtained by the rules. The weight associated to the instances covered by other rules and the size of the intervals are set to penalize rules whose intervals are too large and covering examples already covered by other rules.

The algorithm has been executed 10 times, and only those rules that cover a minimum of 6 samples out of 48 (support 12.5 %) have been taken into account, obtaining a total of 76 rules of the 200 possible rules (10 executions x 20 rules in each run). The limit of the support has been set at that value because 6 samples shows data from a complete volunteer, and such low limit for the support has been chosen because in this type of experiments we are interested both in frequent relations, but also in the not so frequent ones (McIntosh and Chawla, 2007).

The average support obtained for the 151 rules has been 47.17% with a confidence close to 100% for most of them. The average amplitude of intervals in the rules was 24.7%, which justifies the use of quantitative rules in place of the classical rules in which the whole domain of the attribute is taken into account.

The rules obtained have accurately characterized dataset treated, having two types of rules: those with a support value between 75 % and 100 %, and those

Table 1: Analyzed Rules.

Id	**Rule**	**Sup. (%)**	**Conf. (%)**	**Ampl. (%)**
1	$215091\_s\_at \in [98.35 , 376.99]$ and $215760\_s\_at \in [527.04 , 1168.82]$ $\Longrightarrow 203944\_x\_at \in [890.80 , 5308.61]$	52	100	17
2	$205119\_s\_at \in [783.83 , 1527.60]$ and $215597\_x\_at \in [8301.78 , 9819.85]$ $\Longrightarrow 212967\_x\_at \in [2076.59 , 2592.60]$	20	100	16
3	$222099\_s\_at \in [859.491 , 1425.210]$ $\Longrightarrow 49327\_at \in [1517.45 , 2239.45]$	55	100	17

with support values less than 50 % where in almost cases cover records or endotoxin-treated group or placebo group.

The number of rules covering the placebo group is significantly higher, which makes sense because this group has gene expression values more stable and frequent than the group treated with endotoxin(Rubio-Escudero, 2007). To examine the relevance of the rules obtained in the studied problem, we have used the Onto-CC software (Romero-Zliz et al., 2008), which retrieves information regarding the functionality of a set of genes that is passed as a query, and a PI value (the probability of intersection) associated with the relevance of these genes appear together in one rule. PI is a value to minimize between 0 and 1 and considered relevant those obtained under 0.05.

The results of only 3 rules are listed in Table 1 for readability. When Onto-CC is applied, the PI values obtained for every rule are quite low, indicating the relevance of grouping these three genes with respect to these terms, immune response and related terms are explicitly included.

4 CONCLUSIONS

In this paper we present the result of applying an evolutive technique for extracting association rules from microarray data. We have seen the rules obtained are able to successfully characterize the dataset applied, either covering almost all samples, or covering samples only one of the two groups in the data: treated with endotoxin or treated with placebo. In addition, the mean amplitude of the intervals was 24.7%, which justifies the use of quantitative rules in place of the classical rules.

We have shown the relevance of the rules obtained for the problem studied using the Onto-CC program. The PI values obtained show significance in the group of genes found in the rules, and secondly the terms obtained querying Gene Ontology are closely related to the problem of inflammation.

Thus, we conclude that the use of quantitative association rules, in particular those obtained by the algorithm proposed, is a valid method for analyzing microarray data, and we consider it a starting point for future work, applying this technique to other microarray data, comparing with other analytical techniques and seeing the importance of the influence antecedent-consequent obtained by the rules with regard to genetic networks.

ACKNOWLEDGEMENTS

The financial support from the Spanish Ministry of Science and Technology, project TIN 2007-68084-C-02, and from the Junta de Andaluca, project P07-TIC-02611, is acknowledged.

REFERENCES

Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., Chen, R. O., Brownstein, B. H., Cobb, J. P., Tschoeke, S. K., Miller-Graziano, C., Moldawer, L. L., Mindrinos, M. N., Davis, R. W., Tompkins, R. G., Lowry, S. F., and Large Scale Collab Res Program, I. A. (2005). A network-based analysis of systemic inflammation in humans. *Nature*, 437.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

McIntosh, T. and Chawla, S. (2007). High-confidence rule mining for microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(4):611–623.

Romero-Zliz, R., del Val, C., Cobb, J., and Zwir, I. (2008). Onto-cc: a web server for identifying gene ontology conceptual clusters. *Nucleic Acids Res*, 36(4):W352–W357.

Rubio-Escudero, C. (2007). *Fusion of Knowledge towards Identification of Genetic Profiles in the Systemic Inflammation Problem. Ph.D Thesis*. Unversidad de Granada.

Vannucci, M. and Colla, V. (2004). Meaningful discretization of continuous features for association rules mining by means of a som. In *Proceedings of the European Symposium on Artificial Neural Networks*, pages 489–494.

Venturini, G. (1993). SIA: a Supervised Inductive Algorithm with genetic search for learning attribute based concepts. In *Proceedings of the European Conference on Machine Learning*, pages 280–296.

Analysis of Measures of Quantitative Association Rules

M. Martínez-Ballesteros and J.C. Riquelme

Department of Computer Science, University of Seville, Spain
{mariamartinez,riquelme}@us.es

Abstract. This paper presents the analysis of relationships among different interestingness measures of quality of association rules as first step to select the best objectives in order to develop a multi-objective algorithm. For this purpose, the discovering of association rules is based on evolutionary techniques. Specifically, a genetic algorithm has been used in order to mine quantitative association rules and determine the intervals on the attributes without discretizing the data before. The algorithm has been applied in real-word climatological datasets based on Ozone and Earthquake data.

Keywords: Data mining, evolutionary algorithms, quantitative association rules.

1 Introduction

The use of massive processing techniques has revolutionized the scientific research and it has highly increased the amount of data obtained. Data mining is the most used instrumental tool in discovering knowledge from transactions.

In this context we present the result of applying a data mining technique, specifically, association rules (ARs), to data from several experiments. The aim of this process of mining ARs is discover the presence of pairs (attribute (A) - value (v)), which appear in a dataset with certain frequency in order to formulate the rules that display the existing relationship among the attributes.

A revision of the published literature reveals that there are many algorithms to find these rules. Most of the association rule (AR) algorithms are based on methods proposed by Agrawal et al. such as AIS [1] and Apriori [16], SETM[11], etc. Many tools that work in continuous domains just discretize the attributes using a specific strategy and treating these attributes as if they were discrete. Many others are based on evolutionary algorithms. Genetic Algorithms (GAs) [10] are used to solve AR problems because they offer a set of advantages for knowledge extraction and specifically for rule induction processes. Authors of [14] proposed a genetic algorithm (GA) to discover numeric ARs, dividing the process in two phases. Another GA was used in [17] in order to obtain quantitative ARs and confidence was optimized in the fitness function.

Some researches tried to visualize AR mining as a multi-objective problem rather than a single objective. Therefore, several measures can be considered as

E. Corchado, M. Kurzyński, M. Woźniak (Eds.): HAIS 2011, Part II, LNAI 6679, pp. 319–326, 2011.

an objective. In [3] a multi-objective pareto-based GA was presented where the fitness function was formed by four different objectives.

In preliminary work [12][13], authors of this paper developed several single-objective GA that use a weighting scheme for the fitness function which involved some evaluation measures. It is known that a scheme of this nature is not ideal compared to multi-objective schemes, so that could reduce the features used in the fitness function for applying a multi-objective technique. So we expected to extend these algorithms to multi-objective algorithms. However the problem arises when choosing the right objectives to optimize the condition being treated.

Thus, the main motivation of this paper is to analyze the relationship among different evaluation measures of the ARs, in order to classify them in positively correlated, negatively correlated or not correlated. The study is the first step to select the best objectives involved in the subsequent development of a multiobjective GA for extracting ARs. To carry out the study a GA [12] is used for mining quantitative ARs. The algorithm has been applied in two real-world datasets, concretely in ozone data and earthquake data.

The rest of the paper is organized as follows. Section 2 provides a brief preliminary on ARs and some interestingness measures proposed in the literature. Section 3 describes an introduction of multi-objective algorithms. The results obtained are discussed in Section 4. Finally, Section 5 provides the achieved conclusions.

2 Association Rules

In the field of data mining and machine learning, ARs are used to discover common events in a dataset. Several methods have been extensively researched for learning ARs that have been proven to be very interesting to discover relationships among variables in large datasets [16][2]. ARs are classified as unsupervised learning in machine learning.

The AR mining finds interesting associations and/or correlation relationships among elements of large datasets. A typical example is the market-basket analysis [1]. In addition they are widely used in other many fields. It is also useful in the healthcare environment to identify risk factors in the onset or complications of diseases. This form of knowledge extraction is based on statistical techniques such as correlation analysis and variance. One of the most widely used algorithms is the Apriori algorithm.

Formally, an AR is a relationship among attributes in a dataset in the way $A \Rightarrow B$, where A and B are pair conjunctions such as $A = v$ if $A \in \mathbb{Z}$ or $A \in [v_1, v_2]$ if $A \in \mathbb{R}$. Generally, the antecedent A is formed by the conjunction of multiple pairs and the consequent B is usually a single pair.

2.1 Interestingness Measures for Association Rules

The following paragraphs detail the popular measures used to characterize an AR. It is important evaluate the quality of the rule in order to select the best ones and evaluate the results obtained.

Support(A)[9]: The support of an itemset A is defined as the ratio of transactions in the dataset that contain A. Usually, the support of A is named as the probability of A.

$$sup(A) = P(A) = \frac{n(A)}{N}. \tag{1}$$

where $n(A)$ is the number of occurrences of antecedent A in the dataset, and N is the number of transactions forming such dataset.

Support($A \Longrightarrow B$)[9]: The support of the rule $A \Longrightarrow B$ is the percentage of transactions in the dataset that contain A and B simultaneously.

$$sup(A \Longrightarrow B) = P(A \cap B) = \frac{n(AB)}{N}. \tag{2}$$

where $n(AB)$ is the number of instances that satisfy the conditions for the antecedent A and consequent B simultaneously.

Confidence($A \Longrightarrow B$)[9]: The confidence is the probability that transactions containing A, also contain B. In other words, it is the support of the rule divided by the support of the antecedent.

$$conf(A \Longrightarrow B) = \frac{sup(A \Longrightarrow B)}{sup(A)} \tag{3}$$

Lift($A \Longrightarrow B$)[4]: Lift or interest is defined as how many times A and B are together in the dataset more often than expected, assuming that the presence of A and B in transactions are occurrences statically independent. Lift greater than one involves statistical dependence in simultaneous occurrence of A and B. In other words, the rule provides valuable information about A and B occurring together in the dataset.

$$lift(A \Longrightarrow B) = \frac{P(A \mid B)}{P(B)} = \frac{sup(A \Longrightarrow B)}{sup(A)sup(B)} = \frac{conf(A \Longrightarrow B)}{sup(B)} \tag{4}$$

Conviction($A \Longrightarrow B$)[4]: Conviction was introduced as an alternative to confidence for mining ARs in relational databases. Values in the range (0, 1) means negative dependence, higher than 1 means positive dependence and a value equals to 1 means independence. Conviction is directional and gets its maximum value (infinity) when the implication is perfect, that is, if whenever A occurs also happens B.

$$conv(A \Longrightarrow B) = \frac{P(A)P(\neg B)}{P(A \cap \neg B)} = \frac{sup(A)sup(\neg B)}{sup(A \Longrightarrow \neg B)} = \frac{1 - sup(B)}{1 - conf(A \Longrightarrow B)} \tag{5}$$

Gain($A \Longrightarrow B$)[9]: Gain is calculated from the difference between the confidence of the rule and consequent support. It is also known as added value or change of support.

$$Gain(A \Longrightarrow B) = P(A \mid B) - P(B) = conf(A \Longrightarrow B) - sup(B) \tag{6}$$

Certainty Factor($A \Longrightarrow B$)[8]: Certainty factor was introduced by Shortliffe and Buchanan to represent uncertainty in the MYCIN expert system. It is a measure of the variation of the probability that B is in a transaction when we consider only those transactions where A is. A similar interpretation can be done for negative CFs. The certainty factor takes values in [-1, 1] and achieves its maximum possible value, 1, if and only if the rule is totally accurate.
$Conf(A \Longrightarrow B) > Sup(B)$

$$CF(A \Longrightarrow B) = \frac{P(A \mid B) - P(B)}{1 - P(B)} = \frac{conf(A \Longrightarrow B) - sup(B)}{1 - sup(B)} \tag{7}$$

$Conf(A \Longrightarrow B) <= Sup(B)$

$$CF(A \Longrightarrow B) = \frac{P(A \mid B) - P(B)}{P(B)} = \frac{conf(A \Longrightarrow B) - sup(B)}{sup(B)} \tag{8}$$

Leverage($A \Longrightarrow B$) [15]: Leverage measures the proportion of additional cases covered by both A and B above those expected if A and B were independent of each other. Values above 0 are desirable. In addition, leverage is a lower bound for support, so optimizing only leverage guarantees a certain minimum support (contrary to optimizing only confidence or only lift).

$$lev(A \Longrightarrow B) = P(A \cap B) - P(A)P(B) = sup(A \Longrightarrow B) - sup(A)sup(B) \tag{9}$$

In most cases, it is sufficient to focus on a combination of support, confidence, and either lift or leverage to quantitatively measure the "quality" of the rule. However, the real value of a rule, in terms of usefulness and actionability is subjective and depends heavily of the particular domain and business objectives.

3 Multi-objective Optimization

GAs are search algorithms which generate solutions to optimization problems using techniques inspired by natural evolution [10]. They are implemented as a computer simulation in which a population of abstract representations (chromosomes) of candidate solutions (individuals) to an optimization problem evolves toward better solutions. In this context, a classical real-coded GA (RCGA) is used due to the domain of the ARs is continuous, thus, the algorithm deals with numeric data during the whole rule extraction process.

Evolutionary algorithms were originally designed for solving single objective optimization problems. However, many real world optimization problems have more than one objective in conflict with each other. Since multi-objective optimization searches for an optimal vector (rules in data mining) an not just a single value (one rule), one solution often cannot be said to be better than another and there exists not only a single optimal solution, but a set of optimal solutions, called the Pareto-optimal set [19]. The presence of multiple conflicting objectives and the need of using decision-making principles cause a number of different problem scenarios to emerge in practice.

In the last two decades an increasing interest has been developed in the use of GAs for multiobjective optimization. There are multiple proposals of multiobjective GAs [5] as the algorithms MOGA [7], NSGA II [6] or SPEA2 [18] for instance.

The mining process of ARs can be considered as a multi-objective problem rather than a single objective one, in which the measures used for evaluating a rule can be thought as different objectives. There are two goals in multi-objective optimization in the mining of ARs. First, discover rules as close to the Pareto-optimal as possible, and second, find rules as diverse as possible in the obtained non-dominated set. For this purpose, it is necessary to define the best objectives in order to get rules with high accuracy, comprehensible and interesting. In this proposal, several experiments have been carried out and the results are shown in Section 4. The aim of this study is to analyze the correlation and relationships among different evaluation measures of the ARs to define the objectives in order to design a multi-objective GA in this context.

4 Experimental Study

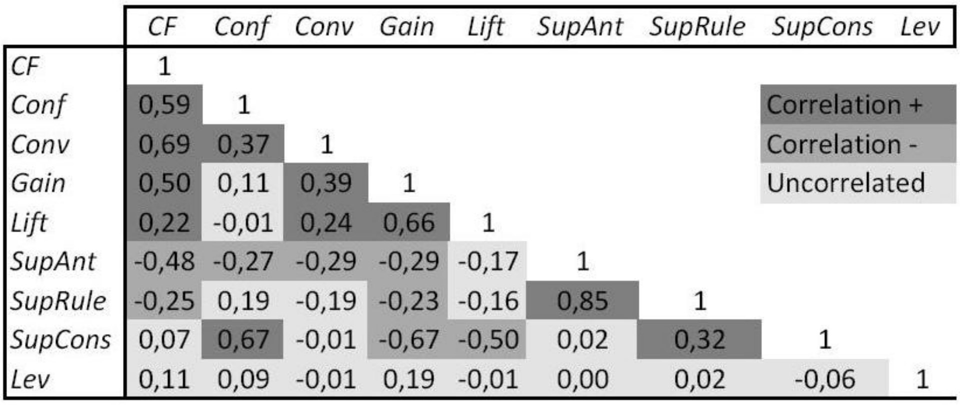
Several experiments have been carried out in this paper to evaluate the relationship among different interestingness measures of ARs. As a preliminary step, the proposed algorithm in [12] by the authors of this work was applied in order to achieve the AR mining task. Two kind of real-world datasets are considered in this work to prevent the resulting set of measures are not dependent on the datasets:

- Ozone concentration: Four datasets have been used containing a compact monthly average values including total ozone content (TOC), over different sites at Iberian Peninsula: Madrid, Arenosillo, Lisbon and Murcia. TOC series are based on ozone data from the Total Ozone Mapping Spectrometer (TOMS) on board the NASA Nimbus-7 satellite from 1st November 1978 to 6th May 1993. Each dataset consists of eight quantitative attributes and 172 samples.
- Earthquakes: The earthquake dataset was collected from the catalogue of Spanish's Geographical Institute (SGI). This dataset consists of four attributes related to location and magnitude of Spanish earthquakes from 1981 to 2008 and 873 samples.

Afterwards, the interestingness measures described in Subsection 2.1 were calculated for the quantitative ARs obtained by the algorithm for each dataset and included in a single database. A statistical study has been carried out to analyze the relationships and dependencies among measures. Specifically, correlation coefficient and principal component analysis (PCA) was applied among the measures.

Correlation coefficient is a measure of the correlation (linear dependence) between two variables X and Y, giving a value between +1 and -1 inclusive. Correlation is +1 in the case of a perfect positive (increasing) linear relationship (correlation), -1 in the case of a perfect decreasing (negative) linear relationship (anticorrelation) [5], and some value between -1 and 1 in all other cases, indicating the degree of linear dependence among the variables. The closer the coefficient is to either -1 or 1, the stronger the correlation among the variables.

PCA is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables called principal components.



	CF	*Conf*	*Conv*	*Gain*	*Lift*	*SupAnt*	*SupRule*	*SupCons*	*Lev*
CF	1								
Conf	0,59	1							
Conv	0,69	0,37	1						
Gain	0,50	0,11	0,39	1					
Lift	0,22	-0,01	0,24	0,66	1				
SupAnt	-0,48	-0,27	-0,29	-0,29	-0,17	1			
SupRule	-0,25	0,19	-0,19	-0,23	-0,16	0,85	1		
SupCons	0,07	0,67	-0,01	-0,67	-0,50	0,02	0,32	1	
Lev	0,11	0,09	-0,01	0,19	-0,01	0,00	0,02	-0,06	1

Correlation +

Correlation -

Uncorrelated

Fig. 1. Correlation coefficients

Table 1. Rotated Components

Measure	Component 1	Component 2	Component 3	Component 4
CF	**0,837**	0,228	-0,276	0,094
Conf	**0,887**	-0,283	0,081	0,101
Conv	**0,721**	0,308	-0,149	-0,120
Gain	0,308	**0,862**	-0,138	0,189
Lift	0,166	**0,808**	-0,019	-0,089
SupAnt	-0,322	-0,035	**0,913**	-0,008
SupRule	0,057	-0,164	**0,973**	0,024
SupCons	0,434	**-0,857**	0,164	-0,050
Lev	0,037	0,052	0,015	**0,985**

Figure 1 shows a table of correlation coefficient among measures: certainty factor (*CF*), confidence (*conf*), conviction (*conv*), gain, lift, support of antecedent (*supAnt*), support of rule (*supRule*), support of consequent (*supCons*) and leverage (*lev*). In the table three cases of correlation have been distinguished: Positive correlation (correlation +) when the coefficient is greater than 0.2, negative correlation (correlation -) when the coefficient is less than -0.2, and not correlation (uncorrelated) in other case. Some interesting conclusions can be extracted from these results.

It can be observed that *CF* is positively correlated to *conf*, *conv*, *gain* and *lift*, and negatively correlated to *supAnt* and *supRule*. *CF* and *conf* are the measures that best correlates positively with other measures. Also, *supRule* is strong correlated with *supAnt*. *supCons* is correlated with *gain*, *lift* and *conf*. However, *lev* is uncorrelated with other measures, thus, independent of other measures.

Table 1 presents the matrix of components with PCA as extraction method and Varimax with Kaiser Normalization as rotation method. The aim of this table is to find groups among the measures, and select the best representative for each group. This study may be useful to choose the various objectives that could be optimized in a multi-objective algorithm for mining ARs. It can be noticed that there are four principal components corresponding each to an independent group of measures. *CF*, *conf* and *conv* belongs to the first group because they are the most correlated in the *Component* 1. *Gain*, *lift* and *supCons* belongs to the second group due to the highest correlation in the *Component* 2. The third group contains *supAnt* and *supCons* and finally, *lev* is only measure of the group 4. In order to select the best objectives, we can study the most correlated for each group. Therefore, *conf*, *gain*, *supRule* and *lev* could be good candidates to optimize the mining of ARs by a multi-objective algorithm.

5 Conclusions

A method of analysis of quality measures of ARs has been proposed in this work. The ARs mining process can be considered as a multi-objective problem rather than a single objective. However, the selection of the best objectives candidates is not arbitrary. Several experiments have been carried out in order to analyze the relationship among different evaluation measures as a previous step before implementing a multi-objective algorithm for association rules. The results have determined that correlation coefficient and principal component analysis can be useful to define dependencies and grouping the interestingness measures of ARs.

Acknowledgments. The financial support from the Spanish Ministry of Science and Technology, project TIN2007-68084-C-00, and from the Junta de Andalucía, project P07-TIC-02611, is acknowledged.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207–216 (1993)
2. Alatas, B., Akin, E.: An efficient genetic algorithm for automated mining of both positive and negative quantitative association rules. Soft Computing 10(3), 230–237 (2006)
3. Alatas, B., Akin, E., Karci, A.: MODENAR: Multi-objective differential evolution algorithm for mining numeric association rules. Applied Soft Computing 8(1), 646–656 (2008)
4. Brin, S., Motwani, R., Silverstein, C.: Beyond market baskets: generalizing association rules to correlations. In: Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, vol. 26, pp. 265–276 (1997)

5. Deb, K.: Multi-Objective Optimization Using Evolutionary Algorithms. John Wiley & Sons, Inc., Chichester (2001)
6. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: Nsga-ii. IEEE Transactions on Evolutionary Computation 6(2), 182–197 (2002)
7. Fonseca, C.M., Fleming, P.J.: Genetic Algorithms for Multiobjective Optimization: Formulation, Discussion and Generalization. In: Proceedings of the Fifth International Conference on Genetic Algorithms, pp. 416–423. Morgan Kaufmann, San Francisco (1993)
8. Fu, L.M., Shortliffe, E.H.: The application of certainty factors to neural computing for rule discovery. IEEE Transactions on Neural Networks 11(3), 647–657 (2000)
9. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Comput. Surv. 38(3), 9 (2006)
10. Goldberg, E.D.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Publishing Company, Reading (1989)
11. Houtsma, M., Swami, A.: Set-Oriented Mining for Association Rules in Relational Databases, pp. 25–33. IEEE Computer Society Press, Los Alamitos (1995)
12. Martínez-Ballesteros, M., Martínez-Álvarez, F., Troncoso, A., Riquelme, J.C.: Quantitative association rules applied to climatological time series forecasting. In: Corchado, E., Yin, H. (eds.) IDEAL 2009. LNCS, vol. 5788, pp. 284–291. Springer, Heidelberg (2009)
13. Martínez-Ballesteros, M., Martínez-Álvarez, F., Troncoso, A., Riquelme, J.C.: Mining quantitative association rules based on evoluationary computation and its application to atmospheric pollution. Integrated Computer-Aided Engineering 17(3), 227–242 (2010)
14. Mata, J., Alvarez, J.-L., Riquelme, J.-C.: Discovering numeric association rules via evolutionary algorithm. In: Chen, M.-S., Yu, P.S., Liu, B. (eds.) PAKDD 2002. LNCS (LNAI), vol. 2336, p. 40. Springer, Heidelberg (2002)
15. Piatetsky-Shapiro, G.: Discovery, analysis and presentation of strong rules. In: Knowledge Discovery in Databases, pp. 229–248 (1991)
16. Srikant, R., Agrawal, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the International Conference on Very Large Databases, pp. 478–499 (1994)
17. Yan, X., Zhang, C., Zhang, S.: Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. Expert Systems with Applications: An International Journal 36(2), 3066–3076 (2009)
18. Zitzler, E., Laumanns, M., Thiele, L.: Spea2: Improving the strength pareto evolutionary algorithm. In: EUROGEN, vol. 3242(103), pp. 95–100 (2001)
19. Zitzler, E., Thiele, L.: Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. IEEE Transactions on Evolutionary Computation 3(4), 257–271 (1999)

Inferring Gene-Gene Associations from Quantitative Association Rules

M. Martínez-Ballesteros, I. Nepomuceno-Chamorro, J.C. Riquelme
Department of Computer Science
University of Seville
Seville, Spain
Email: mariamartinez,inepomuceno,riquelme@us.es

***Abstract*—The microarray technique is able to monitor the change in concentration of RNA in thousands of genes simultaneously. The interest in this technique has grown exponentially in recent years and the difficulties in analyzing data from such experiments, which are characterized by the high number of genes to be analyzed in relation to the low number of experiments or samples available. Microarray experiments are generating datasets that can help in reconstructing gene networks. One of the most important problems in network reconstruction is finding, for each gene in the network, which genes can affect it and how. Association Rules are an approach of unsupervised learning to relate attributes to each other. In this work we use Quantitative Association Rules in order to define interrelations between genes. These rules work with intervals on the attributes, without discretizing the data before and they are generated by a multi-objective evolutionary algorithm. In most cases the extracted rules confirm the existing knowledge about cell-cycle gene expression, while hitherto unknown relationships can be treated as new hypotheses.**

***Keywords*-Data mining; evolutionary algorithms;quantitative association rules; gene networks**

I. Introduction

Microarray technology has revolutionized the biological research due to its ability to monitor changes in RNA concentration in thousands of genes simultaneously [1]. Research in molecular biology has traditionally focused on the study gene to gene, but nowadays we are in the genomic era and genes are studied in thousands or even whole genomes. Standard approaches to microarray analysis (biomarker discovery) are based on the identification of differentially expressed genes and the assumption that genes act independently. However, it is known that powerful prognostic biomarkers may be encoded by genes that are not highly differentially expressed across control and disease patients [2]. Therefore, a systems-level approach can provide insights into the interplay of genes and their association with clinical phenotypes.

In this context we present the result of applying a data mining technique, specifically, association rules, to gene expression data from experiments using microarray technology. The aim of mining association rules is to discover the sets of attributes which appear in a dataset with a certain frequency in order to obtain rules that show the existing relationships among the attributes, specifically, this technique is applied to discover associations between genes from microarray datasets, in which gene expression is linked to another gene expression.

A revision of the published literature reveals that exist many algorithms such as Apriori [3] to find ARs. However, many of these tools that work in continuous domains just discretize the attributes by using a specific strategy and deal with these attributes as if they were discrete [4]. Many algorithms are based on evolutionary algorithms (EAs) [5] which have been extensively used for the optimization and adjustment of models in data mining tasks. EAs are used to discover ARs due to they offer a set of advantages for knowledge extraction and specifically for rule induction processes [6]. In [7] the authors proposed an EA to obtain numeric ARs, dividing the process in two phases. Another EA was used in [8] to obtain ARs where the confidence was optimized in the fitness function.

The mining process of ARs can be considered as a multi-objective problem rather than a single objective one, in which the measures used for evaluating a rule can be thought as different objectives. In the last two decades an increasing interest has been developed in the use of EAs for multi-objective optimization [9]. There are multiple proposals such as the algorithms NSGA II [10] or SPEA2 [11] for instance. In [12] a multi-objective pareto-based EA was presented and another multi-objective GA to AR mining is proposed in [13].

In preliminary works such as the proposed algorithms in [14] and [15], henceforth called QARGA (Quantitative Association Rules by Genetic Algorithm), authors of this paper developed several single-objective EA that use a weighting scheme for the fitness function which involved some evaluation measures.

However, it is known that a scheme of this nature is not ideal compared to multi-objective schemes, so that could reduce the features used in the fitness function for applying a multi-objective technique.

Thus, the main motivation of this paper is to extend these algorithms to a multi-objective approach based on the NSGA-II algorithm. The non-dominated multi-objective evolutionary algorithm proposed in this work can find quantitative association rules in databases with continuous attributes from microarray data, avoiding the discretization as a step in the process. The results will show that the rules obtained have been able to successfully characterize the data underlying and also to group relevant genes for the problem studied.

The rest of the paper is organized as follows. Section II provides a brief preliminary on ARs. Section III describes the methodology used in this work. The results obtained by the developed algorithm are discussed in Section IV. Finally, Section V provides the achieved conclusions.

II. Association rules

Data mining is one of the most used instrumental tools for discovering knowledge from transactions. In the field of data mining, the learning of ARs is a popular and well-known research method for discovering interesting relations among variables in large databases [3]. The discovery of ARs is, unlike classification, a non-supervised learning tool as ARs are descriptive. Descriptive mining tasks identify patterns that explain or summarize the data, that is, they are used to explore the properties of the data, instead of predicting the class of new data [16].

This form of knowledge extraction is based on statistical techniques such as correlation analysis and variance. One of the most widely used algorithms is the Apriori algorithm.

Formally, AR were first defined by Agrawal et al. in [17] as follows. Let $I = \{i_1, i_2, ..., i_n\}$ be a set of n items and $D = \{t_1, t_2, ..., t_N\}$ a set of N transactions, where each t_j contains a subset of items. Thus, a rule can be defined as $X \Rightarrow Y$, where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. Finally, X and Y are called antecedent (or left side of the rule) and consequent (or right side of the rule), respectively.

When the domain is continuous, the ARs are known as Quantitative Association Rules (QAR). In this context, let $F = \{F_1, ..., F_n\}$ be a set of features, with values in $\mathbb{R}$. Let A and C be two disjoint subsets of F, that is, $A \subset F$, $C \subset F$, and $A \cap C = \emptyset$. A QAR is a rule $X \Rightarrow Y$, in which features in A belong to the antecedent X, and features in C belong to the consequent Y, such that X and Y are formed by a conjunction of multiple boolean expressions of the form $F_i \in [v_1, v_2]$. The consequent Y is usually a single expression. In this proposal, QAR are used because the domain is a continuous domain.

It is important measure the quality of the rule in order to select the best rules and evaluate the results obtained by the proposed algorithm. In the ARs mining process, probability-based measures that evaluate the generality and reliability of ARs have been selected [18][19]. In particular, support is used to represent the generality of the rule and confidence, lift and leverage are used to represent the reliability of the rule. Others popular measures are conviction, gain, certainty factor and accuracy.

In most cases, it is sufficient to focus on a combination of support, confidence, and either lift or leverage to quantitatively measure the ”quality” of the rule. However, the real value of a rule, in terms of usefulness and actionability is subjective and depends heavily of the particular domain and business objectives.

III. Methodology

In this section we describes the main features of the proposed algorithm in order to discover ARs from datasets whose attribute are real data.

A. Search of Rules

In a continuous domain, it is necessary to group certain sets of values that share same features and therefore it is required to express the membership of the values to each group. Adaptive intervals instead of fixed ranges have been chosen to represent the membership of such values in this work. The search for the most appropriate intervals has been carried out by means of the proposed algorithm. Thus, the intervals are adjusted to find QAR with high values for support and confidence, together with other measures used in order to quantify the quality of the rule.

Our proposal is based on the NSGA-II approach [10], and its main purpose is to evolve the population based on the non-dominated sort of the solutions in fronts of dominance. The first front is composed of the non-dominated solutions of the population (the Pareto front), the second is composed of the solutions dominated by one solution, the third of solutions dominated by two, and so on. The operating scheme of the algorithm proposed can be seen in Figure 1. The overall complexity of the algorithm NSGA-II is $O(MN^2)$, which is governed by the nondominated sorting part of the algorithm.

In the population, each individual constitutes a rule. These rules are then subjected to an evolutionary process, in which the mutation and crossover operators are applied and, at the end of the process the best individual the Pareto front is designated as the best rule. Our proposal performs an IRL process (Iterative Rule Learning) [21] to penalize instances already covered by rules found by the algorithm, in order to emphasize the covering of instances still not covered. The IRL affects the generation of initial population in each evolutionary process which is described in Subsection III-C.

In order to optimize the mining of AR by the proposed algorithm, thus, rules with high quality and precision, two interestingness measures are selected as objectives:

- *Confidence*($X \implies Y$)[18]: Confidence is defined as the probability that instances satisfying X, also satisfy Y. In other words, it is the support of the rule divided by the support of the antecedent.

$$Conf(X \Longrightarrow Y) = P(X \mid Y) = \frac{sup(X \Longrightarrow Y)}{sup(X)} \quad (1)$$

where $sup(X)$ is the support of the antecedent that is defined as the ratio of instances in the dataset that satisfy the antecedent X, and $sup(X \implies Y)$ is the

```
**Inputs**: Maximum number of rules ($MaxNumRules$),
Maximum number of generations ($MaxNumGen$)
**Output**: Population of the last generation

**Multi-objective Algorithm**(MaxNumRules, MaxNumGen)

Initialize the rule counter $r = 0$
**Repeat**
  1) Initialize the generation counter $t = 0$
  2) Initialize parent population $P_{t=0}$ based on instances covered by
     fewer rules.
  3) Evaluate the individuals of $P_{t=0}$ based on the measures selected
     as objectives.
  4) $P_{t=0}$ is ranked using the Fast non dominated Sort [10] that
     consists in sorting the individuals of a population in different
     Pareto fronts ($F$) according to their non dominance.
     **Repeat**
       a) an offspring population $Q_t$ of same size as $P_t$ is generated
          using crossover and mutation operators over the individuals
          of $P_t$ selected using binary Tournament selection-based
          method [20]
       b) The individuals of $P_t$ and $Q_t$ are merged into $R_t$ and the
          Fast Non dominated Sort is carried out.
       c) The next population $P_{t+1}$ consists of the $N$ best individuals
          of $R_t$.
          Initialize the front counter $i = 0$.
          **Repeat**
              If the current level of $R_t$ ($F_i$, $i-th$ Pareto front) has
              less than or equal to $N$ individuals, the individuals
              of $F_i$ are added to the population $P_{t+1}$.
              In other case,
                  if the current level of $R_t$ ($F_i$, $i-th$ Pareto
                  front) has more than $N$ individuals, the best
                  individuals are used to fill the population of
                  next generation ($P_{t+1}$), and for that purpose,
                  the Crowding distance assignment [10]is
                  used in order to sort the population of the
                  current level and select the best individuals
                  that represent the best rules.

              Increment the front counterr ($i = i + 1$)

          **While** the next population $P_{t+1}$ is not complete.
       d) Increment the generation counter ($t = t + 1$)
     **While** the maximum number of generations is not reached.
  5) **Return** best individual, thus, the rule in the first Pareto front
     ($F_1$) which reach a higher crowding distance value.
  6) Penalize the instances covered by the best rule found.
  7) Increment the rule counter ($r = r + 1$)
**While** the number of desired rules is not reached.
**Return** the best rules found.
```

Figure 1. General scheme of the algorithm.

support of the rule, thus, the percentage of instances in the dataset that satisfy X and Y simultaneously.

- *Leverage*($X \implies Y$)[19]: Leverage measures the proportion of additional cases covered by both X and Y above those expected if X and Y were independent of each other. Leverage takes values inside [-1, 1]. Values equal or under value 0, indicate a strong independence between antecedent and consequent. On the other hand values near 1 are expected for an important association rule. Values above 0 are desirable. In addition, leverage is a lower bound for support, and therefore, optimizing only the leverage guarantees a certain minimum support (contrary to optimizing only the confidence or only the lift).

$$Lev(X \implies Y) = sup(X \implies Y) - sup(X)sup(Y) \quad (2)$$

where $sup(Y)$ is the support of the consequent of the rule, that is, the ratio of instances in the dataset that satisfy the consequent Y.

The proposed algorithm doesn't use a threshold for minimum support and minimum confidence.

The different parts of the algorithm are defined in the following subsections.

B. Individuals Codification

The lower and upper limits of the intervals of each attribute will be represented by the different genes of an individual. Because the attributes are continuous, individuals are represented by a real coding. An individual consists of a not fixed number of attributes less than n, which represents the number of attribute in the database. The representation of an individual consists in two data structures as shown in Figure 2. The upper structure includes all the attributes of the database, where l_j is the lower limit of the range and u_j is the upper limit. The bottom structure indicates the membership of an attribute to the rule represented by an individual. The type of each attribute t_j, can have three values: 0 when the attribute does not belong to the rule, 1 if it belongs to the antecedent of the rule and 2 when it belongs to the consequent part. If an attribute is wanted to be retrieved for a specific rule, it can be done by modifying the value equal to 0 of the type by a value equal to 1 o or 2 depending on the antecedent or consequent.

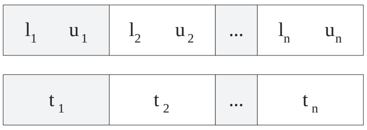


Figure 2. Representation of an individual of the population.

C. Initial Population

The generation of the initial population in the proposed algorithm was carried out at the beginning of each evolutionary process and is perform such at least one chosen sample or instance of the dataset was covered. The samples of the dataset are selected based on their level of hierarchy. The hierarchy is organized according to the number of rules which cover a sample. Thus, the records are sorted by the number of rules that are covered and the samples covered by a few rules have a higher priority.

A sample is selected according to the inverse of the number of rules which cover such sample. Intuitively, the process is similar to roulette selection method where the parents are selected depending on their fitness.

Thus, the samples covered by a few rules have a greater portion of roulette and, therefore, they will be more likely selected. In the first evolutionary process, all samples have the same probability to be selected. Constraints to generate individuals are given by the number of attributes that belong to rule represented by an individual, the number of attributes in the antecedents and consequents and the structure of the rule (attributes fixed or not fixed in consequent).

D. Genetic Operators

The genetic operators implemented in the genetic algorithm proposed are Crossover and Mutation described in [15]. In addition, a new Mutation operator has been added. Concretely, the Antecedent $\Longleftrightarrow$ Consequent Mutation that works as follow: If the type t_i of the selected attribute is antecedent (1), changed to consequent (2), else if the type t_i of the selected attribute is consequent (2), changed to antecedent (1).

IV. Results

We applied our methodology to the microarray datasets of Spellman and Cho for the budding yeast (Saccharomyces cerevisiae) cell-cycle [22] and [23]. These data were synchronized by three different methods: cdc15, cdc28, and alpha-factors. Therefore, these three gene expression data sets may be defined as statistically independent [24].

The same training experiments with cdc15 dataset used by Soinov et al. in [24] were analyzed to achieve a comparison between the two methods. We considered a set of well-described genes, which encode proteins important for cell-cycle regulation.

We selected these genes for the performance analysis of the proposed method in order to establish comparisons with the previous study [24].

A. Parameters configuration

As the proposed algorithm is non-deterministic, it has been executed five times for the dataset. The main parameters are as follows: 100 for the number of the rules to obtain, 50 for the size of the population, 50 for the number of generations, 0.1 for the mutation probability p_{Mut} of the individuals, 0.2 for the mutation probability p_{MutGen} of each gene in the individual.

B. Discussion of Results

In order to choose the best individual (rule) of each generation, the individual with the highest support value in the first Pareto front has been selected in order to cover the maximum number of examples by the obtained rules. We have extracted the relationships between attributes belonging to the antecedent and attributes belonging to the consequent for each AR found by the proposed algorithm in each run. For example, if we have the following rule:

$$A \in [0.2,\ 1.3] \Longrightarrow B \in [0.3,\ 1.2] \wedge C \in [0.5,\ 1.9]$$

the relationships or associations between the attributes of the antecedent and consequent of the rule are:

$$A \Longrightarrow B \text{ and } A \Longrightarrow C$$

Then, we have built a graph with associations derived from the rules, where each attribute that belongs to the rule is a graph node and each association obtained between attributes is an edge of the graph.

For the resulting graph, we performed the intersection between the graphs obtained in each of the five executions carried out by the algorithm in order to find the frequent interrelations between genes.

Table I shows some of the QAR obtained by the algorithm resulting after performing the intersection of the graphs constructed for each algorithm execution. The *Sup. Rule* column, shows the support of the rule that is the percentage of samples covered by the rule. The *Conf* column indicates the probability that instances satisfying the antecedent, also satisfy the consequent. The *Lev* column presents the leverage of the rule and measures the proportion of additional cases covered by both antecedent and consequent above those expected if they were independent of each other. The *Acc* column describes the accuracy of the rule and means the percentage success of the rule. The *CF* column presents the Certainty Factor of the rule. The interest of the rule is shown in column *Lift* and the *Amp* column presents the average amplitude of the intervals of the attributes belonging to each rule.

It is important that the values of all interestingness measures of the AR are as high as possible.

For better understanding, Table I shows rules containing 2 attributes, one attribute in the antecedent and one in the consequent. Rules formed by 3 attributes are shown only for the relationships of genes that are not obtained in any rule of 2 attributes. Because the format of the rules obtained by the algorithm is not fixed, that is, any attribute may belong to the antecedent or the consequent, rules have been obtained with the same attributes but the sense of the implication of the association is different. For example, rules 0 and 1, rules 3 and 4, which are represented as directed edges in the graph in Figure 3.

We can see that the support value of all rules, between 25 % and 50 %, is good enough for the problem at hand. Equally remarkable, the values of confidence, certainty factor and accuracy for most of the rules is equal to 1 or very close to 1, which means that these measures have their highest value and indicates that the rule is totally accurate and the implication of the rule is perfect. The lift and leverage values are quite high, and this means that the rules are interesting and provides valuable information about antecedent and consequent occurring together in the dataset. In addition, the proportion of instances covered by

Table I
QUANTITATIVE ASSOCIATION RULES AND GENE-GENE ASSOCIATIONS INFERRED BY THE PROPOSED ALGORITHM

ID	Rule	Sup. Rule	Conf	Lev	Acc	CF	Lift	Amp	Gene-Gene associations inferred by our method	Soinov
0	$CLN1 \in$ [0.23, 1.21] $\Longrightarrow CLN2 \in$ [0.84, 1.72]	0.292	1	0.207	1	1	3.429	0.26	CLN1 CLN2	
1	$CLN2 \in$ [0.61, 1.72] $\Longrightarrow CLN1 \in$ [0.2, 1.21]	0.333	1	0.222	1	1	3	0.296	CLN2 CLN1	$\surd$
2	$CDC20 \in$ [-0.23, 0.91] $\Longrightarrow CLN1 \in$ [-1.34, -0.28]	0.5	0.857	0.184	0.875	0.688	1.582	0.332	CDC20 CLN1	$\surd$
3	$CLB1 \in$ [-1.37, -0.23] $\Longrightarrow CLB2 \in$ [-1.74, -0.07]	0.5	1	0.25	1	1	2	0.496	CLB1 CLB2	$\surd$
4	$CLB2 \in$ [-1.74, -0.15] $\Longrightarrow CLB1 \in$ [-1.37, -0.23]	0.458	1	0.229	0.958	1	2	0.483	CLB2 CLB1	$\surd$
5	$CLB6 \in$ [-0.92, 0.09] $\Longrightarrow CLB5 \in$ [-0.58, 0.02]	0.375	1	0.219	0.958	1	2.4	0.285	CLB6 CLB5	$\surd$
6	$CLB5 \in$ [-0.58, -0.11] $\Longrightarrow CLB6 \in$ [-0.92, 0.09]	0.333	1	0.208	0.958	1	2.667	0.254	CLB5 CLB6	$\surd$
7	$CLN2 \in$ [0.61, 1.72] $\Longrightarrow CLB5 \in$ [0.25, 1.08]	0.333	1	0.222	1	1	3	0.352	CLN2 CLB5	
8	$CLB2 \in$ [0.42, 1.24] $\Longrightarrow CLB5 \in$ [-1.02, 0.08]	0.458	1	0.172	0.833	1	1.6	0.399	CLB2 CLB5	
9	$CLB2 \in$ [-0.24, 0.93] $\Longrightarrow SW15 \in$ [-0.56, 0.75]	0.542	1	0.226	0.958	1	1.714	0.418	CLB2 SW15	$\surd$
10	$CDC34 \in$ [-1.17, 0.06] $\Longrightarrow MBP1 \in$ [0.28, 1.27]	0.458	1	0.248	1	1	2.182	0.45	CDC34 MBP1	$\surd$
11	$MBP1 \in$ [0.52, 1.27] $\Longrightarrow CDC34 \in$ [-1.17, -0.19]	0.417	1	0.243	1	1	2.4	0.352	MBP1 CDC34	$\surd$
12	$MBP1 \in$ [0.52, 1.13] $\Longrightarrow SKP1 \in$ [-1.47, -0.13]	0.375	1	0.203	0.917	1	2.182	0.358	MBP1 SKP1	$\surd$
13	$SKP1 \in$ [-0.83, -0.24] $\Longrightarrow MBP1 \in$ [0.52, 1.27]	0.33	1	0.194	0.917	1	2.4	0.241	SKP1 MBP1	
14	$SW15 \in$ [0.3, 0.77] $\Longrightarrow CLN2 \in$ [-1.88, -0.14]	0.375	1	0.18	0.875	1	2	0.321	SW15 CLN2	$\surd$
15	$CLB1 \in$ [0.07, 1.27] $\Longrightarrow CLN2 \in$ [-1.88, -0.14]	0.458	0.917	0.208	0.917	0.833	1.833	0.469	CLB1 CLN2	
16	$CLB1 \in$ [-1.37, -0.53] $\Longrightarrow SW15 \in$ [-1.46, -0.34]	0.333	1	0.194	0.917	1	2.4	0.349	CLB1 SW15	$\surd$
17	$CLB2 \in$ [-1.74, -0.15] $\Longrightarrow$ $CLB1 \in$ [-1.37, -0.23] $\wedge CLN2 \in$ [0, 1.72]	0.458	1	0.248	1	1	2.182	0.481	CLB2 CLN2	$\surd$
18	$MBP1 \in$ [-1.12, 0] $\wedge$ $CDC53 \in$ [-0.62, 0.09] $\Longrightarrow$ $SKP1 \in$ [-0.21, 0.74]	0.458	1	0.21	0.917	1	1.846	0.387	CDC53 SKP1	
19	$SW14 \in$ [-0.14, 0.3] $\wedge$ $CLB4 \in$ [0.09, 0.95] $\Longrightarrow$ $CDC34 \in$ [0.06, 0.68] $\wedge CLN1 \in$ [-0.59, 0.99]	0.417	1	0.243	1	1	2.4	0.387	SW14 CDC34	

both antecedent and consequent is greater than ones covered by antecedent and consequent separately. Leverage is a lower bound for support, so optimizing leverage guarantees a certain minimum support (contrary to optimizing only confidence or only lift).

C. Biological Relevance

The associations inferred by our approach are summarized in the tenth column of Table I. The eleventh column of Table I indicates gene-gene associations that were also inferred by the proposed methods by Soinov in [24] using the same dataset. The Gene Regulatory Network corresponding to the rules inferred by our approach and Soinov is shown in Figure 3 and 4, respectively. In summary, all rules inferred by the decision-tree-based method [24] (13 in total) were also inferred by our approach, with the addition of new seven rules inferred only by our proposal. The biological relevance of the rules inferred by our approach was verified by analyzing whether such rules reflect functional properties relating to the different cell-cycle phase. The rules which are supported by the literature are: 3, 4, 5, 6, 9, 10, 11, 12, 14, 16. The rules 1 and 2 are consistent with the prior knowledge and are detected by Soinov. The rules which are not supported by the literature, i.e. 0, y and 7 are new hypothesis to analyze in the laboratory.

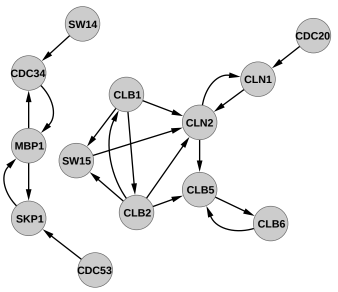


Figure 3. Directed graph obtained by the proposed algorithm.

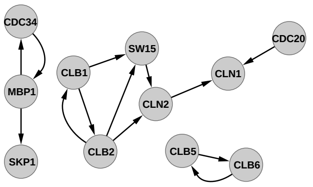


Figure 4. Directed graph obtained by Soinov.

V. CONCLUSION

A multi-objective evolutionary algorithm for mining quantitative association rules has been proposed in this work. The approach is based on the well-known NSGA-II and has determined the intervals that form the rules without discretizing the attributes as a first step of the process. In order to evaluate its performance, the approach has been applied in a dataset and compared to other published results. The results report the relevance and significance in the group of genes found in the rules obtained for the problem studied in terms of support, confidence, accuracy, interest and leverage.

As a conclusion, an advantage of network reconstruction using our approach is that the method is able to construct a network correctly, i.e. reproducing the logic of a network consistent with the data as [24]. The network reconstructed from cell cycle yeast dataset is consistent with the knowledge store in the literature. Furthermore, the method can be improve by adding prior knowledge and more gene expression profiles. Our method constitute an interactive expert system for gene association networks, where the expert decides when to stop adding new gene expression profiles and what biological meaning represent the network.

ACKNOWLEDGMENT

The financial support from the Spanish Ministry of Science and Technology, project TIN2007-68084-C-00, and from the Junta de Andaluca, project P07-TIC-02611, is acknowledged.

REFERENCES

[1] P. Brown and D. Botstein, "Exploring the new world of the genome with dna microarrays," *Nature Genet.*, vol. 21, no. Suppl., pp. 33–37, 1999.

[2] F. Azuaje and Y. D. adn DR Wagner, "Coordinated modular functionality and prognostic potential of a heart failure biomarker-driven interaction network," *BMC Syst. Biol.*, vol. 4, p. 60, 2010.

[3] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the International Conference on Very Large Databases*, 1994, pp. 478–499.

[4] M. Vannucci and V. Colla, "Meaningful discretization of continuous features for association rules mining by means of a som," in *Proceedings of the European Symposium on Artificial Neural Networks*, 2004, pp. 489–494.

[5] E. D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Publishing Company, 1989.

[6] J. Alcalá-Fdez, N. Flugy-Pape, A. Bonarini, and F. Herrera, "Analysis of the effectiveness of the genetic algorithms based on extraction of association rules," *Fundamenta Informaticae*, vol. 98, no. 1, pp. 1001–1014, 2010.

[7] J. Mata, J. L. Álvarez, and J. C. Riquelme, "Discovering numeric association rules via evolutionary algorithm," *Lecture Notes in Artificial Intelligence*, vol. 2336, pp. 40–51, 2002.

[8] X. Yan, C. Zhang, and S. Zhang, "Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support," *Expert Systems with Applications: An International Journal*, vol. 36, no. 2, pp. 3066–3076, 2009.

[9] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Inc., 2001.

[10] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *Evolutionary Computation, IEEE Transactions on*, vol. 6, no. 2, pp. 182 –197, 2002.

[11] E. Zitzler, M. Laumanns, and L. Thiele, "Spea2: Improving the strength pareto evolutionary algorithm," *EUROGEN*, vol. 3242, no. 103, pp. 95 – 100, 2001.

[12] B. Alatas, E. Akin, and A. Karci, "MODENAR: Multi-objective differential evolution algorithm for mining numeric association rules," *Applied Soft Computing*, vol. 8, no. 1, pp. 646–656, 2008.

[13] H. Qodmanan, M. Nasiri, and B. Minaei-Bidgoli, "Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence," *Expert Systems with Applications*, vol. 38, no. 1, pp. 288–298, 2011.

[14] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, and J. C. Riquelme, "Mining quantitative association rules based on evoluationary computation and its application to atmospheric pollution," *Integrated Computer-Aided Engineering*, vol. 17, pp. 227–242, 2010.

[15] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, and J. Riquelme, "An evolutionary algorithm to discover quantitative association rules in multidimensional time series," *Soft Computing*, vol. 15, no. 10, pp. 2065–2084, 2011.

[16] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2006.

[17] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 1993, pp. 207–216.

[18] L. Geng and H. Hamilton, "Interestingness measures for data mining: A survey," *ACM Comput. Surv.*, vol. 38, no. 3, p. 9, 2006.

[19] G. Piatetsky-Shapiro, "Discovery, analysis and presentation of strong rules," in *Knowledge Discovery in Databases*, 1991, pp. 229–248.

[20] B. Miller and D. Goldberg, "Genetic algorithms, tournament selection, and the effects of noise," *Complex Systems*, vol. 9, pp. 193–212, 1995.

[21] G. Venturini, "SIA: A Supervised Inductive Algorithm with genetic search for learning attribute based concepts," in *Proceedings of the European Conference on Machine Learning*, 1993, pp. 280–296.

[22] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization," *Mol Biol Cell 1998,*, vol. 9, pp. 3273–3297, 1998.

[23] R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfsberg, A. Gabrielian, D. Landsman, D. Lockhart, and R. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Mol Cell*, vol. 2, pp. 65–73, 1998.

[24] L. Soinov, M. Krestyaninova, and A. Brazma, "Towards reconstruction of gene networks from expression data by supervised learning," *Genome Biology*, vol. 4, p. R6, 2003.

Parte III

Conclusiones y Trabajos futuros

Capítulo 6

Conclusiones

La principal motivación para realizar el trabajo desarrollado en esta tesis doctoral ha sido estudiar y desarrollar algoritmos capaces de encontrar QAR en bases de datos con atributos continuos. En concreto se han propuesto cuatro EA llamados QARGA, QARGA-CHC, EQAR y MOQAR para descubrir QAR capaces de encontrar los intervalos más adecuados sin realizar un proceso de discretización como paso previo a diferencia de muchos otros enfoques que sí llevan a cabo discretización para extraer reglas.

Cabe destacar que las distintas propuestas desarrolladas se pueden integrar en la herramienta KEEL [Alcalá-Fdez et al., 2009] que se trata de una herramienta software basada en Java cuyo objetivo es evaluar EA para problemas de DM tales como regresión, clasificación, aprendizaje no supervisado, etc. Los ficheros de configuración de los algoritmos propuestos en esta tesis doctoral, así como las bases de datos utilizadas han seguido el mismo esquema y formato de los ficheros en KEEL con el objetivo de poder compartir y reutilizar fuentes de datos, facilitar la inserción de los algoritmos desarrollados así como el uso de las funcionalidades disponibles dentro de esta herramienta. Del mismo modo, los ficheros de salida se han configurado de tal manera que su lectura y comprensión resulte cómoda para el usuario.

La parte inicial de la memoria se ha dedicado a la descripción del problema de las QAR, el contexto de investigación y los objetivos a perseguir en la tesis doctoral así como un resumen de la selección de trabajos y discusión conjunta de los resultados de los mismos. La segunda parte que ocupa la parte central y más relevante de la memoria ha recapitulado la selección de publicaciones vinculadas con la temática de la tesis doctoral que han motivado el desarrollo de la misma. En esta parte se han detallado en profundidad las distintas propuestas Qarga, Qarga-Chc, Eqar y Moqar así como los resultados más relevantes para cada una de ellas. Los resultados experimentales han mostrado un notable funcionamiento para Qarga puesto que ha obtenido mejores resultados cuando se ha comparado con otras propuestas evolutivas como Earmga y Genar en bases de datos públicas procedentes del repositorio BUFA. También ha sido capaz de extraer eficientemente reglas a partir de series temporales multidimensionales generadas sintéticamente. En cuanto a las series temporales multidimensionales del mundo real, Qarga ha descubierto QAR significativas y coherentes para el problema del ozono, ya que se han obtenido dependencias relevantes entre la concentración de ozono y otras series temporales climatológicas mejorando los resultados obtenidos por Apriori. Qarga también ha demostrado ser un método válido para analizar datos procedentes de microarrays por la relevancia de las reglas obtenidas que ha sido demostrada usando el programa Onto-CC.

La segunda propuesta Qarga-Chc se ha basado en el conocido algoritmo Chc y los resultados han demostrado su buen funcionamiento ya que ha obtenido mejores resultados que el algoritmo Modenar en bases de datos sintéticas bajo diferentes niveles de ruido. Ha resultado ser efectivo para propósitos predictivos en series temporales de agentes contaminantes (O_3, NO y SO_2) ya que las reglas extraídas han sido acordes con los procesos químicos asociados con estos agentes.

La tercera propuesta llamada Eqar ha sido una extensión mejorada de Qarga-Chc añadiendo nuevas características para mejorar su funcionamiento. Se ha aplicado al problema de modelado de TOC en la Península Ibérica (Lisboa, Madrid, Murcia, Arenosillo y Montlouis) y los resultados han sido bastante buenos en cuanto a la calidad de las QAR encontradas además de estar en consonancia con los resultados obtenidos en otros trabajos relacionados con el mismo problema. El descubrimiento de QAR en modelado TOC ha resultado ser un método de análisis interesante para el futuro en problemas similares.

La cuarta propuesta denominada como Moqar se ha basado en el conocido enfoque multiobjetivo Nsga-II y ha sido una mejora de las tres propuestas anteriores. Con el objetivo de seleccionar los objetivos más adecuados, se han analizado las relaciones entre las diferentes medidas de evaluación mediante métodos estadísticos. Las QAR obtenidas por este algoritmo han dado lugar a relevantes y significantes grupos de genes consistentes con el conocimiento existente en la literatura.

La calidad de los resultados obtenidos por las cuatro propuestas nos permite concluir que las propuestas desarrolladas han satisfecho notoriamente los objetivos a perseguir en esta tesis doctoral, además de ser herramientas válidas para problemas del mundo real por su capacidad de aplicación en distintos dominios como hemos podido comprobar.

Capítulo 7

Conclusions

The main motivation to carry out the work developed in this PhD dissertation has been to study and develop algorithms able to find QAR from datasets with continuous attributes. Specifically, four EA called QARGA, QARGA-CHC, EQAR and MOQAR have been proposed to discover QAR. These algorithms are able to find the most appropriate intervals without performing a previous discretization unlike many other approaches.

It can be noted that the different proposals developed can be integrated into the KEEL tool [Alcalá-Fdez et al., 2009] that is a Java-based software tool designed to evaluate problems such as regression, classification, unsupervised learning, and so on. The configuration files of the algorithms proposed in this PhD dissertation and the datasets used have followed the same scheme and format of the files on KEEL in order to share and reuse data sources, facilitating the integration of our algorithms and using the available features in this tool. Similarly, the output files have been configured making easier the reading and the comprehension for the user.

The initial part of this PhD dissertation has been devoted to describe the QAR problem, the research context and the objectives to be pursued. In addition, a summary of the selected research works and the discussion of the results obtained has been included. The second part is the central and most important part of the dissertation, has summed up the selection of publications related to the topics that have motivated the development of this PhD dissertation. The proposals QARGA, QARGA-CHC, EQAR, MOQAR and the results most relevant have been thoroughly detailed in this part. Experimental results have shown a remarkable performance for QARGA since it has obtained the best results when it is compared with other evolutionary proposals as EARMGA and GENAR in public datasets from BUFA repository. Furthermore, QARGA has been able to discover rules from multidimensional time series synthetically generated efficiently. In terms of real word multidimensional time series, QARGA also has discovered significant and consistent QAR for the ozone problem, since relevant dependences among the ozone concentration and other climatological time series have been discovered improving the results obtained by APRIORI. QARGA has proven to be a valid method to analyze microarray data because the relevance of the rules obtained has been demonstrated using the program Onto-CC.

The second proposal QARGA-CHC is based on the well known algorithm CHC and the results have shown a good performance since the results obtained have been better than that of the algorithm MODENAR in synthetic datasets under different noise levels. This algorithm has proved to be effective for predictive purposes in time series of pollutants (O_3, NO and SO_2) because the rules discovered have been in accordance with the chemical processes associated with these agents.

The third approach called EQAR has been an improved extension of QARGA-CHC adding new features to improve its performance. This proposal has been applied to the problem of TOC modeling in the Iberian Peninsula (Lisbon, Madrid, Murcia, Montlouis and Arenosillo) and the results have been good in terms of quality of the QAR and they are consistent with the results obtained in other works associated with the same problem. The discovery of QAR in TOC modeling has proven to be an interesting analysis method for the future in similar problems.

The fourth proposal referred as MOQAR was based on the well-known multi-objective approach NSGA-II and it has been an improvement of the three previous proposals. In order to select the appropriate objectives to optimize, the relationships among the different measures have been analyzed using statistical methods. The QAR obtained by this algorithm have provided important and significant groups of genes consistent with the existing knowledge in the literature.

From the quality of the results obtained by the four proposals, it can be concluded that the developed proposals have notoriously satisfied the aim pursued in this PhD dissertation, besides develop valid tools for real-world problems by its applicability in different domains.

Capítulo 8

Trabajos futuros

Una vez comentadas las conclusiones generales obtenidas tras el desarrollo de esta tesis doctoral, se detallan algunas propuestas y futuras tareas como mejoras de los actuales algoritmos y posibles líneas de investigación.

8.1. Mejora de la escalabilidad de algoritmos evolutivos para descubrir reglas de asociación cuantitativas

Durante el desarrollo de esta tesis doctoral se ha realizado una estancia de investigación desde Junio hasta Septiembre del 2011 en el grupo ASAP de la School of Computer Science en la Universidad de Nottingham para colaborar con el Dr. Jaume Bacardit con el objetivo de mejorar los algoritmos desarrollados para la extracción de QAR.

En este periodo de tiempo el principal trabajo ha consistido en estudiar y mejorar la escalabilidad de las propuestas presentadas en esta tesis doctoral, y por otro lado, la aplicación de las mismas a conjuntos de datos de bioinformática para extracción de relaciones entre aminoácidos. La motivación que ha conducido a mejorar la escalabilidad de los algoritmos QARGA, QARGA-CHC, EQAR y MOQAR es el gran coste computacional que supone la evaluación de los individuos en los EA cuando se aplican a dominios con dimensionalidad muy alta, es decir, bases de datos con un gran número de registros y bases de datos con un gran número de atributos como es el caso de las bases de datos procedentes de microarrays, que se caracterizan por cientos, miles o incluso millones de datos.

Por tanto el principal objetivo que se ha perseguido durante la estancia ha sido modificar nuestras propuestas evolutivas para mejorar su eficiencia sin utilizar un hardware especial o paralelo combinando una nueva representación para atributos continuos y un mecanismo de mejora de eficiencia basados en las técnicas propuestas en [Bacardit and Garrell, 2003], [Bacardit et al., 2004] y [Bacardit et al., 2009].

Actualmente estamos finalizando la fase experimental utilizando datos procedentes del repositorio Infobiotics PSP benchmarks [Bacardit and Krasnogor, 2008] con el fin de evaluar las técnicas aplicadas a las distintas propuestas para la mejora de la escalabilidad de las mismas en bases de datos de alta dimensionalidad. Del mismo modo, se plantea como trabajo futuro analizar a nivel biológico las dependencias extraídas entre aminoácidos a partir de las QAR obtenidas.

8.2. Medidas de calidad para un conjunto de reglas de asociación cuantitativas

Actualmente en la literatura existe una amplia gama de medidas de calidad para evaluar AR, sin embargo, todas ellas están dedicadas a la evaluación de cada regla a nivel individual. Podría suceder que las reglas obtenidas por cualquier individuo fueran de alta calidad pero no constituyeran un conjunto significativo de reglas debido a que presenten algunos problemas como es el caso de que todas estuvieran cubriendo el mismo espacio de búsqueda y por tanto, el porcentaje total de registros cubiertos de la base de datos fuera muy bajo.

Por ello, se plantea como trabajo futuro la redefinición de las medidas de interés tales como *soporte*, *confianza*, *leverage*, *gain*, etc, para evaluar la calidad de las AR a nivel de conjunto.

8.3. Potenciar la diversidad en MOQAR

La propuesta QARGA-CHC se caracteriza por estar basada en el esquema del algoritmo CHC cuya finalidad era establecer un equilibro entre la diversidad y la convergencia que influyen negativamente sobre la efectividad de los GA. Por otro lado, la propuesta MOQAR se basa en el esquema multiobjetivo del algoritmo NSGA-II debido a los inconvenientes que supone una función de agregación de objetivos como se ha comentado anteriormente. Podemos ver que cada problema que solventa cada una de las propuestas es distinto, por lo que sería ideal el desarrollo de una propuesta que hibride las características de ambos enfoques para el descubrimiento de QAR. Por tanto se propone como trabajo futuro introducir mecanismos que potencien la diversidad en MOQAR tales como la inclusión de un operador de re-inicialización cuando la diversidad en la población sea pobre, así como la aplicación de la prevención de incesto en el operador de cruce.

8.4. Otras perspectivas futuras

También se considera como tarea futura aumentar el ámbito de aplicación de las distintas propuestas desarrolladas en esta tesis doctoral a otros problemas del mundo real. En concreto, podría resultar interesante el descubrimiento de QAR en datos procedentes de terremotos, así como en series temporales de oferta y demanda dentro del mercado eléctrico. Del mismo modo, consideramos ampliar las actuales propuestas añadiendo nuevas funcionalidades, utilizando nuevos enfoques, desarrollando nuevos operadores genéticos y teniendo en cuenta más medidas de interés con el fin de mejorar la tarea de descubrimiento de QAR.

Bibliografía

[Agrawal et al., 1993] Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216.

[Agrawal and Srikant, 1994] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the International Conference on Very Large Databases*, pages 478–499.

[Alatas and Akin, 2008] Alatas, B. and Akin, E. (2008). Rough particle swarm optimization and its applications in data mining. *Soft Computing*, 12(12):1205–1218.

[Alatas et al., 2008] Alatas, B., Akin, E., and Karci, A. (2008). MODENAR: Multi-objective differential evolution algorithm for mining numeric association rules. *Applied Soft Computing*, 8(1):646–656.

[Alcalá-Fdez et al., 2009] Alcalá-Fdez, J., Alcalá, R., Gacto, M. J., and Herrera, F. (2009). Learning the membership function contexts forming fuzzy association rules by using genetic algorithms. *Fuzzy Sets and Systems*, 160(7):905–921.

[Alcalá-Fdez et al., 2010] Alcalá-Fdez, J., Flugy-Pape, N., Bonarini, A., and Herrera, F. (2010). Analysis of the effectiveness of the genetic algorithms based on extraction of association rules. *Fundamenta Informaticae*, 98(1):1001–1014.

[Alcalá-Fdez et al., 2009] Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M., Ventura, S., Garrell, J., Otero, J., Romero, C., Bacardit, J., Rivas, V., Fernández, J., and Herrera, F. (2009). Keel: A software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, 13(3):307–318.

[Aumann and Lindell, 2003] Aumann, Y. and Lindell, Y. (2003). A statistical theory for quantitative association rules. *Journal of Intelligent Information Systems*, 20(3):255–283.

[Ayubi et al., 2009] Ayubi, S., Muyeba, M. K., Baraani, A., and Keane, J. (2009). An algorithm to mine general association rules from tabular data. *Information Sciences*, 179:3520–3539.

[Bacardit et al., 2009] Bacardit, J., Burke, E. K., and Krasnogor, N. (2009). Improving the scalability of rule-based evolutionary learning. *Memetic Computing*, 1(1):55 – 67.

[Bacardit and Garrell, 2003] Bacardit, J. and Garrell, J. M. (2003). Incremental learning for pittsburgh approach classifier systems. In *Proceedings of the Segundo Congreso Español de Metaheurísticas, Algoritmos Evolutivos y Bioinspirados*, pages 303 – 311.

[Bacardit et al., 2004] Bacardit, J., Goldberg, D., Butz, M., Llorà, X., and Garrell, J. M. (2004). Speeding-up pittsburgh learning classifier systems: Modeling time and accuracy. In *Parallel Problem Solving from Nature - PPSN 2004*, pages 1021–1031. Springer-Verlag, LNCS 3242.

[Bacardit and Krasnogor, 2008] Bacardit, J. and Krasnogor, N. (2008). The infobiotics psp benchmarks repository. (http://www.infobiotic.net/PSPbenchmarks).

[Back, 1996] Back, T. (1996). *Evolutionary Algorithms in Theory and Practice*, volume 2. Oxford University Press.

[Back et al., 1997] Back, T., Fogel, D., and Michalewicz, Z. (1997). *Handbook of Evolutionary Computation.* Oxford University Press.

[Bellazzi et al., 2005] Bellazzi, R., Larizza, C., Magni, P., and Bellazzi, R. (2005). Temporal data mining for the quality assessment of hemodialysis services. *Artificial Intelligence in Medicine*, 34:25–39.

[Berzal et al., 2001] Berzal, O., Blanco, I., Snchez, D., and Vila, M. (2001). Ios press measuring the accuracy and interest of association rules: A new framework.

[Brin et al., 1997a] Brin, S., Motwani, R., and Silverstein, C. (1997a). Beyond market baskets: generalizing association rules to correlations. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, volume 26, pages 265–276.

[Brin et al., 1997b] Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. (1997b). Dynamic itemset counting and implication rules for market basket data. In *Proc. of the ACM SIGMOD 1997*, pages 265–276.

[Calvano et al., 2005] Calvano, S. E., Xiao, W., Richards, D., Felciano, R., Baker, H., Cho, R., Chen, R., Brownstein, B., Cobb, J., Tschoeke, S., Miller-Graziano, C., Moldawer, L., Mindrinos, M., Davis, R., Tompkins, R., Lowry, S., and Program, I. L. S. C. R. (2005). A network-based analysis of systemic inflammation in humans. *Nature*, 437.

[Chen et al., 2009] Chen, C. H., Hong, T. P., and Tseng, V. (2009). Speeding up genetic-fuzzy mining by fuzzy clustering. In *Proceedings of the IEEE International Conference on Fuzzy Systems*, pages 1695–1699.

[Chen et al., 2010] Chen, C. H., Hong, T. P., and Tseng, V. (2010). Genetic-fuzzy mining with multiple minimum supports based on fuzzy clustering. *Soft Computing, in press.*

[Cho et al., 1998] Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D., and Davis, R. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell*, 2:65–73.

[Davis, 1991] Davis, L. (1991). *Handbook of Genetic Algorithms*. Van Nostrand Reinhold.

[Deb, 2001] Deb, K. (2001). *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Inc.

[Deb et al., 2002] Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: Nsga-ii. *Evolutionary Computation, IEEE Transactions on*, 6(2):182 –197.

[Dehuri et al., 2006] Dehuri, S., Jagadev, A., Ghosh, A., and Mall, R. (2006). Multi-objective genetic algorithm for association rule mining using a homogeneous dedicated cluster of workstations. *American Journal of Applied Science*, 3:2086 – 2095.

[del Jesús et al., 2007] del Jesús, M. J., González, P., Herrera, F., and Mesonero, M. (2007). Evolutionary fuzzy rule induction process for subgroup discovery: A case study in marketing. *IEEE Transactions on Fuzzy Systems*, 15(4):578–592.

[del Jesus et al., 2011] del Jesus, M., Gámez, J., González, P., and Puerta, J. (2011). On the discovery of association rules by means of evolutionary algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(5):397–415.

[Dunham, 2003] Dunham, M. H. (2003). *Data Mining: Introductory and Advanced Topics.* Prentice Hall.

[Eiben and Smith, 2003] Eiben, A. and Smith, J. (2003). *Introduction to Evolutionary Computing.* Natural Computing Series. Springer.

[Eiben and Smith, 2008] Eiben, A. E. and Smith, J. (2008). *Introduction to Evolutionary Computing (Natural Computing Series).* Springer.

[Eshelman, 1991] Eshelman, L. (1991). *The CHC Adaptative search algorithm: How to have safe search when engaging in nontraditional genetic recombination.* Morgan Kaufmann.

[Fogel, 1995] Fogel, D. (1995). *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence.* IEEE Press Series on Computational Intelligence. IEEE Press.

[Fogel, 1998] Fogel, D. (1998). *Evolutionary Computation: The fossil record.* Wiley-IEEE Press.

[Fonseca and Fleming, 1993] Fonseca, C. M. and Fleming, P. (1993). Genetic algorithms for multiobjective optimization: Formulation, discussion and generalization. In *Genetic Algorithms: Proceedings of the Fifth International Conference*, pages 416–423. Morgan Kaufmann.

[García et al., 2009] García, S., Fernández, A., Luengo, J., and Herrera, F. (2009). A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability. *Soft Computing*, 13(10):959–977.

[Geng and Hamilton, 2006] Geng, L. and Hamilton, H. (2006). Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38(3):9.

[Georgii et al., 2005] Georgii, E., Richter, L., Rckert, U., and Kramer, S. (2005). Analyzing microarray data using quantitative association rules. *BMC Bioinformatics*, 21(2):123–129.

[Ghosh and Nath, 2004] Ghosh, A. and Nath, B. (2004). Multi-objective rule mining using genetic algorithms. *Information Science*, 163:123 – 133.

[Giordana and Neri, 1995] Giordana, A. and Neri, F. (1995). Search-intensive concept induction. *Evolutionary Computation*, 3(4):375–416.

[Goldberg, 1989] Goldberg, E. D. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley Publishing Company.

[Greene and Smith, 1993] Greene, D. and Smith, S. (1993). Competition-based induction of decision models from examples. *Machine Learning*, 13(2):229–257.

[Guha et al., 1998] Guha, S., Rastogiand, R., and Shim, K. (1998). CURE: An efficient clustering algorithm for large databases. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 73–84.

[Gupta et al., 2006] Gupta, N., Mangal, N., Tiwari, K., and Mitra, P. (2006). Mining quantitative association rules in protein sequences. *Lecture Notes in Artificial Intelligence*, 3755:273–281.

[Guvenir and Uysal, 2000] Guvenir, H. A. and Uysal, I. (2000). Bilkent university function approximation repository. *http://funapp.cs.bilkent.edu.tr.*

[Hajela and Lin, 1992] Hajela, P. and Lin, C. (1992). Genetic search strategies in multicriterion optimal design. *Structural Optimization*, 4(2):99 – 107.

[Han and Kamber, 2006] Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques.* Morgan Kaufmann.

[Han et al., 2004] Han, J., Pei, J., Yin, Y., and Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach.

[Hand, 2007] Hand, D. J. (2007). Principles of data mining. *Drug safety an international journal of medical toxicology and drug experience*, 30(7):621–622.

[Horn and Nafpliotis, 1993] Horn, J. and Nafpliotis, N. (1993). Multiobjective optimization using the niched pareto genetic algorithm. *World*, 1(IlliGAl Report 93005):61801–2996.

[Houtsma and Swami, 1995] Houtsma, M. and Swami, A. (1995). Set-oriented mining for association rules. In *Proceedings of IEEE Data Engineering Conference*.

[Huang et al., 2008] Huang, Y. P., Kao, L. J., and Sandnes, F. E. (2008). Efficient mining of salinity and temperature association rules from ARGO data. *Expert Systems with Applications*, 35:59–68.

[Ishibuchi and Murata, 1998] Ishibuchi, H. and Murata, T. (1998). A multi-objective genetic local search algorithm and its application to flowshop scheduling. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 28(3):392 –403.

[Janikow, 1993] Janikow, C. Z. (1993). A knowledge-intensive genetic algorithm for supervised learning. *Machine Learning*, 13(2-3):189–228.

[Jong, 1975] Jong, K. D. (1975). *An Analysis of the Behavior of a Class of Genetic Adaptive Systems*. PhD thesis.

[Jong et al., 1993] Jong, K. D., Spears, W., and Gordon, D. (1993). Using genetic algorithms for concept learning. *Machine Learning*, 13(2-3):161–188.

[Kaya and Alhajj, 2005] Kaya, M. and Alhajj, R. (2005). Genetic algorithm based framework for mining fuzzy association rules. *Fuzzy Sets and Systems*, 152(3):587–601.

[Kaya and Alhajj, 2006] Kaya, M. and Alhajj, R. (2006). Utilizing genetic algorithms to optimize membership functions for fuzzy weighted association rules mining. *Applied Intelligence*, 24(1):7–152.

[Khan et al., 2010] Khan, M. S., Coenen, F., Reid, D., Patel, R., and Archer, L. (2010). A sliding windows based dual support framework for discovering emerging trends from temporal data. *Research and Development in Intelligent Systems*, Part 2:35–48.

[Lin and Lee, 2002] Lin, M. and Lee, S. (2002). Fast discovery of sequential patterns by memory indexing. In *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery*, pages 150–160.

[Martínez-Ballesteros et al., 2009] Martínez-Ballesteros, M., Martínez-Álvarez, F., Troncoso, A., and Riquelme, J. (2009). Quantitative association rules applied to climatological time series forecasting. In *Proceedings of the 10th international conference on Intelligent data engineering and automated learning (IDEAL'09)*, volume 5788 of *Lecture Notes in Computer Science*, pages 284–291.

[Martínez-Ballesteros et al., 2011a] Martínez-Ballesteros, M., Martínez-Álvarez, F., Troncoso, A., and Riquelme, J. (2011a). An evolutionary algorithm to discover quantitative association rules in multidimensional time series. *Soft Computing*, 15(10):2065–2084.

[Martínez-Ballesteros et al., 2010] Martínez-Ballesteros, M., Martínez-Álvarez, F., Troncoso, A., and Riquelme, J. C. (2010). Mining quantitative association rules based on evoluationary computation and its application to atmospheric pollution. *Integrated Computer-Aided Engineering*, 17:227–242.

[Martínez-Ballesteros and Riquelme, 2011] Martínez-Ballesteros, M. and Riquelme, J. (2011). Analisys of measures of quantitative association rules. In *International Conference on Hybrid Artificial Intelligent Systems (HAIS'11)*, volume 6679 of *Lecture Notes in Computer Science*, pages 319–326.

[Martínez-Ballesteros et al., 2011b] Martínez-Ballesteros, M., Rubio-Escudero, C., Riquelme, J., and Martínez-Álvarez, F. (2011b). Mining quantitative association rules in microarray data using evolutive algorithms. In *ICAART 2011 - Proceedings of the 3rd International Conference on Agents and Artificial Intelligence*, volume 1, pages 574–577.

[Martínez-Ballesteros et al., 2011] Martínez-Ballesteros, M., Salcedo-Sanz, S., Riquelme, J., Casanova-Mateo, C., and Camacho, J. (2011). Evolutionary association rules for total ozone content modeling from satellite observations. *Chemometrics and Intelligent Laboratory Systems*, 109(2):217 – 227.

[Mata et al., 2001] Mata, J., Álvarez, J., and Riquelme, J. C. (2001). Mining numeric association rules with genetic algorithms. In *Proceedings of the International Conference on Adaptive and Natural Computing Algorithms*, pages 264–267.

[Mata et al., 2002] Mata, J., Álvarez, J. L., and Riquelme, J. C. (2002). Discovering numeric association rules via evolutionary algorithm. *Lecture Notes in Artificial Intelligence*, 2336:40–51.

[Michalewicz, 1992] Michalewicz, Z. (1992). Genetic algorithms+data structures = evolution programs. *New York Springer Verlag.*

[Nam et al., 2009] Nam, H., Lee, K., and Lee, D. (2009). Identification of temporal association rules from time-series microarray data sets. *BMC Bioinformatics*, 10(3):1–9.

[Orallo et al., 2004] Orallo, J. H., Quintana, M. R., and Ramírez, C. F. (2004). *Introducción a la Minería de Datos.* Prentice Hall.

[Orriols-Puig et al., 2008] Orriols-Puig, A., Casillas, J., and Bernadó-Mansilla, E. (2008). First approach toward on-line evolution of association rules with learning classifier systems. In *Proceedings of the 2008 GECCO Genetic and Evolutionary Computation Conference*, pages 2031–2038.

[Pei et al., 2001] Pei, J., Han, J. W., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., and Hsu, M. C. (2001). Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings of IEEE Conference on Data Engineering*, pages 215–224.

[Piatetsky-Shapiro, 1991] Piatetsky-Shapiro, G. (1991). Discovery, analysis and presentation of strong rules. In *Knowledge Discovery in Databases*, pages 229–248.

[Pyle, 1999] Pyle, D. (1999). Data preparation for data mining. *Order A Journal On The Theory Of Ordered Sets And Its Applications*, 17(5 & 6):375–381.

[Qodmanan et al., 2011] Qodmanan, H., Nasiri, M., and Minaei-Bidgoli, B. (2011). Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence. *Expert Systems with Applications*, 38(1):288–298.

[Romero-Zaliz et al., 2008] Romero-Zaliz, R., Val, C. D., Cobb, J., and Zwir, I. (2008). Onto-cc: a web server for identifying gene ontology conceptual clusters. *Nucleic Acids Research*, 36(Web Server issue):W352–W357.

[Ruckert et al., 2004] Ruckert, U., Richter, L., and Kramer, S. (2004). Quantitative association rules based on half-spaces: An optimization approach. In *Proceedings of the IEEE International Conference on Data Mining*, pages 507–510.

[Sarker et al., 2002] Sarker, R., Liang, K., and Newton, C. (2002). A new multiobjective evolutionary algorithm. *European Journal of Operational Research*, 140(1):12 – 23.

[Shortliffe and Buchanan, 1975] Shortliffe, E. and Buchanan, B. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23:351–379.

[Soinov et al., 2003] Soinov, L., Krestyaninova, M., and Brazma, A. (2003). Towards reconstruction of gene networks from expression data by supervised learning. *Genome Biology*, 4:R6.

[Spellman et al., 1998] Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell 1998,*, 9:3273–3297.

[Srikant and Agrawal, 1996] Srikant, R. and Agrawal, R. (1996). Mining quantitative association rules in large relational tables. *ACM SIGMOD Record*, 25(2):1–12.

[Srinivas and Deb, 1994] Srinivas, N. and Deb, K. (1994). Muiltiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary Computation*, 2(3):221–248.

[Tong et al., 2005] Tong, Q., Yan, B., and Zhou, Y. (2005). Mining quantitative association rules on overlapped intervals. *Lecture Notes in Artificial Intelligence*, 3584:43–50.

[Tung et al., 2003] Tung, A., Lu, H., Han, J., and Feng, L. (2003). Efficient mining of intertransaction association rules. *IEEE Transactions on Knowledge and Data Engineering*, 15(1):43–56.

[Vannucci and Colla, 2004] Vannucci, M. and Colla, V. (2004). Meaningful discretization of continuous features for association rules mining by means of a som. In *Proceedings of the European Symposium on Artificial Neural Networks*, pages 489–494.

[Venturini, 1993] Venturini, G. (1993). SIA: a Supervised Inductive Algorithm with genetic search for learning attribute based concepts. In *Proceedings of the European Conference on Machine Learning*, pages 280–296.

[Wakabi-Waiswa and Baryamureeba, 2008] Wakabi-Waiswa, P. and Baryamureeba, V. (2008). Extraction of interesting association rules using genetic algorithms. *International Journal of Computing and ICT Research*, 2(1):26 – 33.

[Wan et al., 2007] Wan, D., Zhang, Y., and Li, S. (2007). Discovery association rules in time series of hydrology. In *Proceedings of the IEEE International Conference on Integration Technology*, pages 653–657.

[Winarko and Roddick, 2007] Winarko, E. and Roddick, J. F. (2007). ARMADA – An algorithm for discovering richer relative temporal association rules from interval-based data. *Data and Knowledge Engineering*, 63:76–90.

[Yan et al., 2009] Yan, X., Zhang, C., and Zhang, S. (2009). Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. *Expert Systems with Applications: An International Journal*, 36(2):3066–3076.

[Yang et al., 2010] Yang, Y., Webb, G., and Wu, X. (2010). Discretization methods. In Maimon, O. and Rokach, L., editors, *Data Mining and Knowledge Discovery Handbook*, pages 101–116. Springer US.

[Zaki, 2000] Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12:372–390.

[Zitzler et al., 2001] Zitzler, E., Laumanns, M., and Thiele, L. (2001). Spea2: Improving the strength pareto evolutionary algorithm. *EUROGEN*, 3242(103):95 – 100.

[Zitzler and Thiele, 1999] Zitzler, E. and Thiele, L. (1999). Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *Evolutionary Computation, IEEE Transactions on*, 3(4):257 –271.