Presentación: Minería de Datos

Roberto Ruiz[(1)], Karina Gilbert[(2)], José C. Riquelme[(3)]

[(1)] Área de Lenguajes y Sistemas Informáticos
Universidad Pablo de Olavide, Sevilla
rruisan1@upo.es
[(2)] Departamento de Estadística e Investigación Operativa
Universidad Politécnica de Cataluña, Barcelona
karina.gibert@upc.edu
[(3)] Departamento de Lenguajes y Sistemas Informáticos
Universidad de Sevilla
riquelme@lsi.us.es

Resumen

La Minería de Datos ha experimentado en los últimos años una notable explosión de interés tanto en ámbitos académicos como industriales. Se trata de un área interdisciplinar fuertemente relacionada con el aprendizaje automático y la estadística, así como con la gestión de bases de datos y la visualización de datos o la optimización. En esta introducción se presenta el orígen de este número y su estructura, así como una breve descripción de los artículos que lo constituyen y que ofrecen la panorámica más actual de lo que en estos momentos es la investigación en Minería de Datos en nuestro país.

Palabras clave: Minería de Datos.

1. Introduction

La minería de datos ha experimentado en los últimos años una notable explosión de interés tanto en ámbitos académicos como industriales. Se trata de un área interdisciplinar fuertemente relacionada con el aprendizaje automático y la estadística, así como con la gestión de bases de datos y la visualización de datos o la optimización. Tanto en aprendizaje automático como en estadística, problemas del tipo de aprender cómo es la estructura subyacente de un conjunto de datos, han sido objeto de estudio intensivo durante décadas.

Es más, las primeras formulaciones en clasificación de datos [1] en clustering (o análisis de conglomerados) [2] o aprendizaje inductivo [3] se remontan a casi 40 años atrás.

Partiendo de esos primeros trabajos, el reciente incremento por el interés en la minería de datos ha hecho inmersión en áreas como las bases de datos o la visualización y el enfoque principalmente aplicado de la minería de datos lo ha puesto en el punto de mira de una extensa gama de dominios, como la economía, el márqueting, la medicina, la biología, las ciencias medioambientales o la astronomía, para mencionar solo algunas de ellas.

La intersección entre el aprendizaje automático, la estadística y la minería de datos es muy amplia y ello trata de ilustrarse en los artículos de este número especial.

Los artículos de este número son una selección de los mejores trabajos presentados en el III Taller de Minería de Datos y Aprendizaje (TAMIDA), celebrado en Granada el 13 de Setiembre de 2005, en el marco del I Congreso Español de Informática

(CEDI). TAMIDA 2006 ha supuesto el tercer encuentro de la Red Española de Minería de Datos y Aprendizaje (MIDAS) (www.lsi.us.es/redmidas), financiada por el gobierno español. En el workshop, se seleccionaron 42 artículos para ser incluidos en las actas del congreso. TAMIDA 2006 tuvo más de 70 participantes. Todos los autores fueron invitados a someter versiones de sus trabajos a este número especial. Finalmente, se sometieron 25 trabajos procedentes de toda la geografía española, y tras un riguroso proceso de revisión en doble ciego, se seleccionaron los 10 mejores trabajos para su publicación.

Este número pretende ofrecer una perspectiva de la investigación que alrededor de la Minería de Datos tiene lugar en España en la actualidad, mostrando una variedad de temas relacionados con Minería de Datos y sus aplicaciones en distintos dominios.

Los editores quieren agradecer a todos los autores que sometieron sus artículos para ser considerados para este número especial. Asimismo a los revisores por sus comentarios constructivos y sugerencias, que ayudaron a mejorar la calidad de los artículos que finalmente constituyen este número. Los trabajos varían desde el caràcter eminentemente metodológico al esencialmente práctico; y de ámbitos tan tecnológicos como el control de procesos batch en un sistema informático a otros de interés localizado en nuestro país, como la producción de la oveja manchega.

1.1. Los trabajos

El primer artículo escrito por los editores de este número trata de introducir los principales conceptos usados en Minería de Datos y mostrar diferentes campos de aplicación. Finalmente se presentan los retos actuales de la disciplina.

En el trabajo "Elección de Operadores Lógicos para la Inducción de Conocimiento Comprensible" de Pedro G. Espejo y otros se desarrolla un sistema de minería de datos orientado a la tarea de clasificación, utilizando reglas como formalismo de representación. El objetivo principal es analizar el balance entre precisión y comprensibilidad, centrándose en un aspecto de la comprensibilidad: el que viene determinado por la elección de los operadores lógicos que pueden aparecer en el antecedente de las reglas. El sistema desarrollado se basa en la programación genética gramatical, ya que otro objetivo del trabajo es estudiar la utilidad de esta técnica evolutiva para llevar a cabo tareas de minería.

En el trabajo "A meta-learning framework for pattern classification by means of data complexity measures" de J. M. Sotoca y otros estudian el problema de cómo influyen las características de los datos en el comportamiento de los clasificadores. En este trabajo se presenta un entorno de meta-aprendizaje definiendo una medida de la complejidad de los datos y su aplicación a diversos problemas de análisis de patrones.

En el siguiente trabajo "Aprendizaje discriminativo de clasificadores Bayesianos" de Guzmán Santafé y otros, se estudian los modelos bayesianos desde un enfoque discriminativo en el que se busca maximizar la verosimilitud condicional, frente al enfoque generativo tradicional donde se maximiza la verosimilitud conjunta de los datos.

En el trabajo se presenta un nuevo método para el aprendizaje discriminativo, tanto de la estructura como de los parámetros, de clasificadores Bayesianos. Esta aproximación está basada en la adaptación del algoritmo TM a modelos de clasificación Bayesianos. Asimismo se presenta una evaluación experimental del método propuesto aplicado a diferentes bases de datos estándares para clasificación supervisada.

Gonzalo Ramos-Jiménez y otros en su trabajo "Sistemas multiclasificadores y de aprendizaje por capas basados en CIDIM" describen dos sistemas multiclasificadores basados en un enfoque de aprendizaje por capas particularizado con el algoritmo CIDIM como base. CIDIM (Control de Inducción por DIvisión Muestral) es un algoritmo que ha sido desarrollado para inducir árboles de decisión precisos y pequeños y para conseguirlo, intenta reducir el sobreajuste usando un control de inducción local. Los sistemas multiclasificadores presentados aprovechan ciertas características de CIDIM, pero los enfoques desarrollados se pueden extender a otros algoritmos que compartan esas mismas características.

El trabajo "Similarity Functions for Structured Data. An Application to Decision Trees" de Vicente Estruch-Gregori y otros, trata de para datos estructurados, comenzando con una descripción de funciones de similaridad definidas sobre datos estructurados (listas, conjuntos, …). Posteriormente se presenta un método propio para construir un clasificador basado en árboles de decisión adaptado a la utilización de una función de similaridad y el

estudio experimental correspondiente con datos proposicionales y complejos.

Ricardo Blanco-Vega y otros en "La Técnica Mimética en Ausencia de Datos Originales: Aprendizaje y Revisión de Modelos" estudian la técnica mimética consistente en usar un modelo, generalmente preciso pero incomprensible, como paso previo para generar un conjunto de datos aleatorios que luego se utiliza junto al conjunto de datos inicial para entrenar a un segundo modelo comprensible. Esta técnica se ha empleado para dotar de comprensibilidad a modelos de caja negra sin sacrificar considerablemente su precisión. En este trabajo se estudia la aplicación del mimetismo en un escenario en el que los datos originales de entrenamiento no están disponibles. Finalmente se aplica la técnica mimética a la revisión de modelos mostrando que, en determinadas situaciones de cambio, el modelo mimético puede usarse como modelo de transición entre el modelo original y el nuevo.

En el trabajo "Selección genética para la mejora de la raza ovina manchega mediante técnicas de Minería de Datos" de M. Julia Flores y otros se presenta una aplicación de técnicas de clasificación a valores genéticos con el objetivo de mejorar las cifras de producción de la oveja manchega. Se aplican técnicas de aprendizaje de redes Bayesianas que capten el modelo y las relaciones entre los distintos factores influyentes en la determinación del mérito genérico (modelo descriptivo). El objetivo no es el de sustituir el uso de los métodos tradicionales, sino al contrario, aprender tanto con técnicas supervisadas como no supervisadas de los resultados obtenidos mediante éstas. Además, a partir de los modelos aprendidos se extrae información preliminar sobre el valor genético de un animal, que puede resultar de gran utilidad, empleándose técnicas de clasificación con el fin de identificar buenos subconjuntos de predictores.

Otro trabajo de aplicación es "Minería y visualizacion de datos del mercado electrico español" de F. E. Sánchez-Ubeda y otros. En concreto se estudian los datos disponibles para las empresas que participan en el mercado eléctrico español, con un volumen muy elevado y con una información potencialmente rica. La extracción de conocimiento sobre el comportamiento estratégico de la competencia en el mercado, condensado en las curvas de oferta, supone una ventaja competitiva. En este artículo se propone una metodología de análisis de las curvas de oferta basada en el empleo de técnicas de minería de datos, presentándose numerosos ejemplos del tipo de conocimiento que se puede obtener.

El área de *web mining* está representado en este número por el trabajo "Un modelo de minería de consultas para el diseño del contenido y la estructura de un sitio Web" de Ricardo Baeza y otros. En este trabajo se presenta un modelo para hacer minería de consultas en sitios Web, mediante la relación entre la información aportada por las consultas de un sitio, con los datos de uso, contenido y estructura. El principal objetivo del modelo es descubrir en forma simple, información valiosa acerca de cómo mejorar la estrutura y contenido del sitio, permitiendo que éste sea más intuitivo y adecuado a las necesidades de sus usuarios. Se propone el análisis de los diferentes tipos de consultas registradas en los logs de uso de un sitio Web, tales como las consultas formuladas por los usuarios en el motor de búsqueda interno del sitio y las consultas realizadas en buscadores externos, que condujeron hacia documentos en el sitio. Este modelo además propone una validación visual de la organización jerárquica del sitio, dada por los enlaces entre sus documentos y sus contenidos, además de su relación con las consultas.

Finalmente, Magda Ruiz y otros nos presentan su trabajo "Combination of statistical process control (SPC) methods and classification strategies for situation assessment of batch process". En él se estudia el desarrollo de una estrategia de clasificación para identificar situaciones críticas en un control de procesos *batch*. Se hace una reducción de los datos mediante un análisis de componentes principales. El objetivo es categorizar mediante un algoritmo de clasificación las causas de un mal comportamiento del sistema.