Selecting the best measures to discover quantitative association rules

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

Keywords:
Quantitative association rules
Quality measures
Optimal fitness function
Evolutionary algorithms

A B S T R A C T

The majority of the existing techniques to mine association rules typically use the support and the confidence to evaluate the quality of the rules obtained. However, these two measures may not be sufficient to properly assess their quality due to some inherent drawbacks they present. A review of the literature reveals that there exist many measures to evaluate the quality of the rules, but that the simultaneous optimization of all measures is complex and might lead to poor results. In this work, a principal components analysis is applied to a set of measures that evaluate quantitative association rules' quality. From this analysis, a reduced subset of measures has been selected to be included in the fitness function in order to obtain better values for the whole set of quality measures, and not only for those included in the fitness function. This is a general-purpose methodology and can, therefore, be applied to the fitness function of any algorithm. To validate if better results are obtained when using the function fitness composed of the subset of measures proposed here, the existing QARGA algorithm has been applied to a wide variety of datasets. Finally, a comparative analysis of the results obtained by means of the application of QARGA with the original fitness function is provided, showing a remarkable improvement when the new one is used.

1. Introduction

Hybrid artificial intelligent systems are rapidly gaining relevance in the scientific community due to the ability shown to deal with real-life problems [1,13,14]. These systems combine the use of both extracted knowledge and raw data to solve problems.

High volume of data can be stored nowadays; therefore, the use of efficient computational techniques has become a task of the utmost importance. In this context, the discovery of association rules (AR) – and particularly of quantitative association rules (QAR) in this work – is a popular methodology that allows the discovery of significant and apparently hidden relations among variables that form databases [3,4,27,28].

The AR extraction process consists in using a non-supervised strategy to explore data properties. The main goal pursuit is, then, to find groups of attributes appearing frequently together in a dataset, so to provide comprehensive rules able to explain the existing relations among them.

A review of the literature reveals that there exist many algorithms to find AR. Most of them are based on the methods proposed by Agrawal et al. such as AIS [2], Apriori [3] or SETM [26].

Nonetheless, there is another big group of techniques to extract AR that are based on evolutionary algorithms (EA). EA are search algorithms that generate solutions for optimization problems using techniques inspired in natural evolution [18,23], in which a population of abstract representations (chromosomes) of candidate solutions (individuals) evolves toward better solutions. EA can be used to discover AR, since they offer several advantages for knowledge extraction and for rule induction processes [7].

The algorithms that discover AR are normally assessed by means of certain interestingness measures that are able to evaluate the quality of a rule. From all of them, support and confidence highlight although lift, gain, certainty factor or leverage are also indicators that provide useful information about the extracted rules.

A review on AR learning based on the use of EA applied to boolean, categorical, quantitative and fuzzy variables has been described in [16]. However, as this work is focused on quantitative variables only the works using this kind of data are reviewed in this section. Table 1 summarizes the measures used for both evaluation and optimization in several works recently published. From the observation of this table, one conclusion can be easily drawn: There is no uniformity on the selection of measures to assess the algorithms' performance.

For instance, an EA called EARMGA was used in [45] to obtain QAR. The confidence was the only objective to be optimized in the fitness function. To achieve this goal, the authors avoided the

Table 1
Quality measures used in the literature.

Algorithm	Quality measures considered					
	Support	Confidence	Re-covered	Comprehensibility or # Attributes	Amplitude	Interest
GENAR [33]	√	√	√			
GAR [34]	√	√	√	√	√	
Tong et al. [43]	√	√				
QuantMiner [40]	√	√				
Kaya and Alhajj [27]	√	√				√
Kaya and Alhajj [28]	√	√				√
Dehuri et al. [15]		√		√		√
Alatas and Akin [5]	√	√	√	√	√	
MODENAR [6]	√	√		√	√	
Orriols-Puig et al. [36]	√	√				
Ayubi et al. [9]	√	√				
EARMGA [45]		√				
Quodmanan et al. [39]	√	√				
NSGA-II-QAR [29]	√	√		√	√	√
GAR-plus [37]	√	√	√	√	√	

specification of the actual minimum support, which can be considered the main contribution of the work.

The combination of confidence and support as only quality measures can be found in several works. Hence, the work introduced in [43] proposes an approach to discover QAR by clustering items of a dataset and projecting the clusters into the domains of the quantitative attributes to form meaningful intervals. Also, the algorithm called QuantMiner [40] proposed a genetic algorithm to mine QAR and optimize support and confidence, by using a fitness function based on the gain measure proposed in [19].

The extraction of QAR has also been applied to the data streams field. A classifier, whose main novelty lied on its adaptability to online gathered data was presented in [36]. By contrast, a multi-objective approach was proposed in [39]. The algorithm did not consider the minimum support and confidence and applied the FP-tree algorithm [25]. The fitness function maximized both support and confidence of the rule. Finally, some works such as [9] have proposed the use of an extended set of operators to mine general association rules and have evaluated the proposal in terms of confidence and support.

Additionally to support and confidence, the authors of the work introduced in [33] used the number of recovered instances to evaluate their approach, called GENAR. GENAR is an EA-based approach capable of obtaining an undetermined number of quantitative attributes in the antecedent of the rule. The same quality measures plus the comprehensibility and the amplitude of the intervals forming the rule were used to evaluate the GAR algorithm [34] (and in its extension [37]). The comprehensibility measure [22] is defined as the logarithm of the number of attributes in the consequent divided by the logarithm of the number of attributes in the rule. The amplitude measure is defined as the addition of the amplitudes for each interval of the attributes which belong to the rule divided by the number of attributes. The authors proposed another EA but, this time, it was necessary to select which attributes formed the antecedent and which one the consequent. Recently, a comparative analysis of the effectiveness in QAR extraction has been presented [7], in which the algorithms GENAR [33], GAR [34] and EARMGA [45] were applied to two different datasets showing their efficiency in terms of coverage and confidence. These five features (support, confidence, recovered, comprehensibility and amplitude) were also evaluated on a multi-objective Pareto-based EA called MODENAR [6]. The same authors proposed an optimization metaheuristic based on rough particle swarm techniques to mine QAR [5]. The fitness function was composed of four different objectives in both works: Support, confidence, comprehensibility of the rule (to be maximized) and the amplitude of the intervals that forms the rule (to be minimized).

Alternatively, the support and confidence have been combined with the interest to form fitness functions in some works [27,28]. Their main particularity lies on the use of genetic algorithms to mine fuzzy association rules. The authors in [29] went one step further and used, in addition to the three measures aforementioned, the amplitude of the intervals as well as the comprehensibility of the rule to form the fitness function.

Finally, the authors in [15] proposed a fast and scalable multi-objective GA for mining AR from large datasets using parallel processing and a homogeneous dedicated network of workstations. The confidence, comprehensibility and interest were the measures maximized.

There is no unanimity in choosing the set of quality measures to be optimized, thus it becomes essential to propose a methodology to automatically select a subset of them whose optimization leads to the optimization of the entire set. Therefore, this work is focused on finding relations among different quality measures in order to determine which measures must be optimized in the fitness function. This way, it is expected that better rules can be extracted, regarding the whole set of measures and not only those included in the fitness function. To fulfill this task, this subset is generated according to a principal component analysis (PCA). The QARGA algorithm [31] has been used to check the new fitness function composed of the selected measures versus the original fitness function based on a weighting scheme that involved several evaluation measures such as support, confidence, number of attributes and amplitude of intervals of the attributes belonging to the rules. In particular, datasets from the public Bilkent University Function Approximation (BUFA) repository [24] have been used. Likewise, four different real-world datasets have been analyzed, specifically, datasets from biological, meteorological and seismological nature.

The remainder of the paper is as follows. Section 2 introduces the foundations underlying QAR. It also explores the most used measures found in the literature as well as some of their inherent

drawbacks. Section 3 provides the statistical analysis conducted to select the target measures and a brief description of the QARGA's main features. The methodology introduced in previous Section is applied to a wide variety of datasets in order to determine the fitness function in Section 4. The results obtained by QARGA using both original and new fitness functions along with statistical tests can also be found in this section. Finally, Section 5 summarizes the achievements reached in this work and the conclusions drawn.

2. Quantitative association rules

This section provides a brief description on QAR including some definitions. In addition, some measures of interest proposed in the literature and some of their flaws are presented.

2.1. Definitions

Formally, AR were first defined by Agrawal et al. in [2] 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 $X, Y \subseteq I$ 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 AR 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 $A \cap C = \varnothing$. 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.

2.2. Quality measures for QAR

This section details the most popular quality measures used to evaluate QAR. Note that it is very important to measure the quality of the rules to evaluate the results obtained by any methodology. In Table 2 the mathematical definition, the description of the evaluated properties and the range of variability of these measures are shown. Note that $n(X)$ is the number of occurrences of the itemset X in the dataset and N is the total number of instances in the dataset.

2.3. Drawbacks of the quality measures

The support and the confidence are the most used measures for QAR optimization. However, if only these measures were optimized, some inconveniences might occur. The optimization of the support may not be enough since very general QAR could be obtained, and the amplitude of the intervals could be enlarged until reaching the whole domain of each attribute. Moreover, to optimize only the confidence also presents some problems because this measure does not consider the support of the consequent of the rule, therefore it is not able to detect negative dependence among items.

To overcome these issues, the measures lift and leverage can be optimized as they consider the antecedent and the consequent of the rule. Furthermore, the lift evaluates the interest of the rule and the leverage measures the degree of dependence between the antecedent and the consequent of the rule. However, the lift is sensitive to noise in small databases since rare set of items with low probability to occur together can produce huge lift values. On the other hand, the measures gain, certainty factor and conviction also consider the support of the antecedent and the consequent of the rule and therefore, these measures also consider the rule's direction of the implication. Despite the accuracy does not consider the support of the rule nor the direction of the implication of the rule, this measure is very useful and powerful to validate QAR.

Table 3
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

Table 2
Quality measures for quantitative association rules.

Measures	Equation	Description	Range
$Sup(X)$	$n(X)/N$	Coverage of X	[0, 1]
$Sup(X \Longrightarrow Y)$	$n(X \cap Y)/N$	Generality of the rule	[0, 1]
$Conf(X \Longrightarrow Y)$	$sup(X \Longrightarrow Y)/sup(X)$	Reliability of the rule	[0, 1]
$Lift(X \Longrightarrow Y)$ [10]	$sup(X \Longrightarrow Y)/(sup(X) \cdot sup(Y))$	Interest of the rule • Value <1: X and Y negatively dependent • Value $=1$: X and Y independent • Value >1: X and Y positively dependent	$[0, +\infty)$
$Conviction(X \Longrightarrow Y)$ [11]	$(1-sup(Y))/(1-conf(X \Longrightarrow Y))$	Implication of the rule • Value <1: X and Y negatively dependent • Value $=1$: X and Y independent • Value >1: X and Y positively dependent	$(0, +\infty)$
$Gain(X \Longrightarrow Y)$ [21]	$conf(X \Longrightarrow Y)-sup(Y)$	Added value or change of support	$[-0.5, 1]$
$Certainty\ Factor(X \Longrightarrow Y)$ [41]	• If $conf(X \Longrightarrow Y) > sup(Y)$: $(conf(X \Longrightarrow Y)-sup(Y))/(1-sup(Y))$ • If $conf(X \Longrightarrow Y) <= sup(Y)$: $(conf(X \Longrightarrow Y)-sup(Y))/sup(Y)$	Gain normalized, strength of the rule • Value <0: X and Y negatively dependent • Value $=0$: X and Y independent • Value >0: X and Y positively dependent	$[-1, 1]$
$Leverage(X \Longrightarrow Y)$ [38]	$sup(X \Longrightarrow Y)-sup(X)sup(Y)$	Strength of the rule • Value <0: X and Y negatively dependent • Value $=0$: X and Y independent • Value >0: X and Y positively dependent	$[-0.25, 0.25]$
$Accuracy(X \Longrightarrow Y)$ [21]	$sup(X \Longrightarrow Y) + sup(\neg X \Longrightarrow \neg Y)$	Veracity of the rule	[0, 1]

Table 4
Quality measures for the example rules 1 and 2.

Measure	Rule 1	Rule 2	Best
Antecedent support	0.38	0.38	Tie
Consequent support	0.38	0.88	Rule 2
Rule support	0.25	0.25	Tie
Confidence	0.66	0.66	Tie
Lift	1.76	0.75	Rule 1
Leverage	0.23	−0.57	Rule 1
Accuracy	0.75	0.25	Rule 1
Gain	0.29	−0.21	Rule 1
Certainty factor	0.46	−0.34	Rule 1
Conviction	1.87	0.37	Rule 1

For a better understanding of the quality measures meaning an illustrative example is given in Table 3, by using a dataset that comprises eight instances and three features. Consider then two example rules, henceforth called Rules and 2, respectively:

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

Table 4 describes the values of the quality measures related to Rules 1 and 2. As can be observed, Rules 1 and 2 have the same support of the antecedent and the support of the rule. Therefore, the confidence for both rules is also the same although the support of the consequent is higher for Rule 2.

Note that Rule 1 is a refinement of Rule 2 because the amplitude of the consequent is lower and the percentage of covered records is the same in both rules. It can be clearly seen, that the first rule has higher quality although both rules have the same support and confidence values.

The *lift* of Rules 1 and 2 is 1.76 and 0.75, respectively. Here, *lift* of Rule 1 is larger than the *lift* of Rule 2, confirming the intuition that the first rule is more interesting than the second one. Regarding the values of *accuracy* and *leverage*, they are also higher in Rule 1. Therefore, it can be concluded that the first rule has better quality, accuracy, interest and strong degree of dependence between the antecedent and consequent than the second one even if they have the same confidence. In the case of the two example rules, *gain*, *certainty factor* and *conviction* also show that Rule 1 is better since the values of the measures are higher and positive, contrary to Rule 2 that presents negative values in *gain* and *certainty factor* measures.

To base the QAR optimization on only one measure is not sufficient in most cases, and an adequate combination of some of them is required. A study to select the target measures that summarize the whole set is conducted in Section 3.

3. Methodology

This section presents the methodology based on PCA in order to determine a fitness function, which simultaneously optimizes the maximum number of quality measures possible. Also, a brief description of QARGA and its fitness function is provided.

3.1. Principal component analysis

A statistical analysis of the relationships among the measures of interest described in Section 2 has been carried out to select a set of measures to be included in the fitness function of any algorithm to obtain association rules. In other words, the goal of the statistical analysis is to find the set of measures that characterize the full set of measures and which will form the fitness function.

As a preliminary step, the QARGA algorithm [31] has been applied to several datasets of different nature to avoid the dependence between the results and the datasets. In particular, the real-world and public datasets described in Section 4.2 have been used to obtain five hundred QAR for each dataset (5 executions of 100 rules per dataset). In the case of the BUFA repository composed of thirty-five datasets, the average values of the measures for the five hundred rules have been obtained for each dataset. Thus, a matrix composed of 2035 rows (5 executions × 100 QAR × 4 datasets + 35 average values of BUFA repository) and 10 columns (one column for each measure) has been generated as PCA's input.

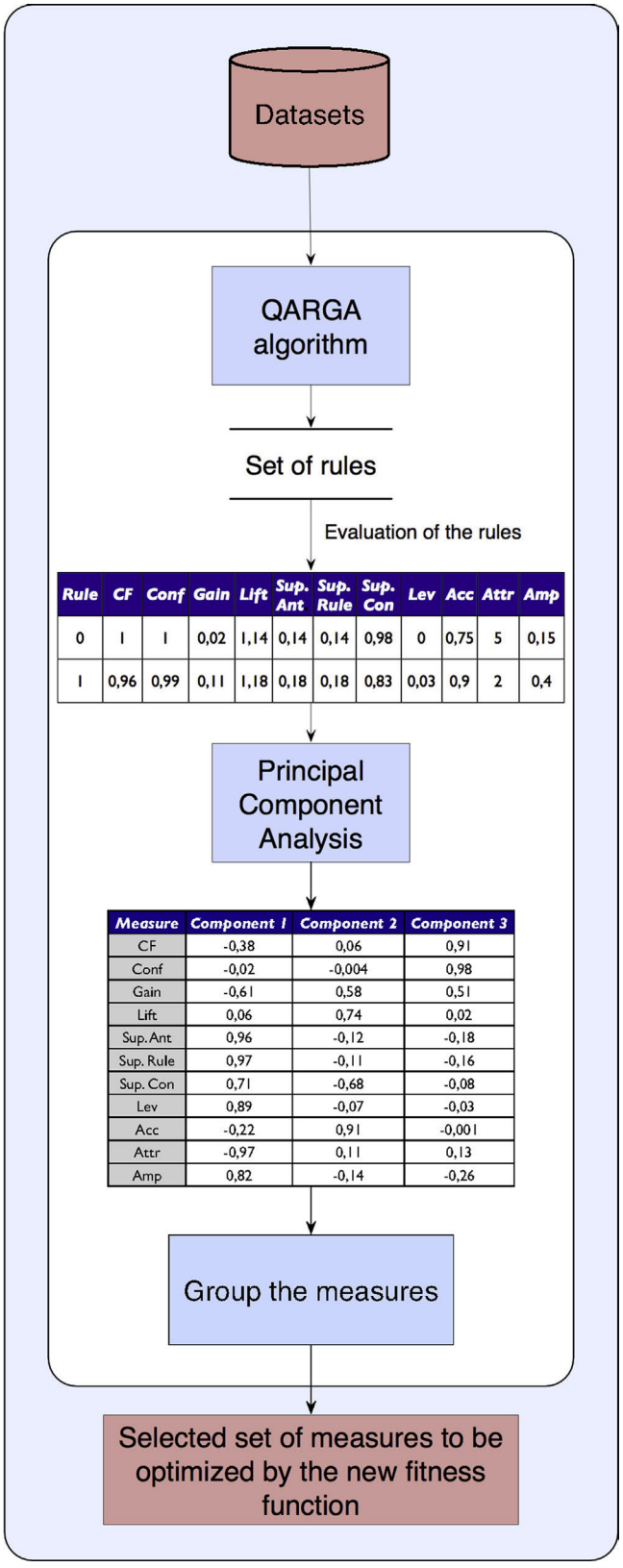
For this purpose, the relations and dependencies existing among the measures have been obtained by the application of PCA. This method identifies the components that can synthesize the maximum information possible contained in the original set of variables. In particular, PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables called principal components. In this work, the well-known rotation method Varimax with Kaiser Normalization has been used. Once PCA has been applied to the measures computed for the QAR obtained for each dataset, one measure representing each component is selected. Thus, all selected measures will be the objectives to optimize by the fitness function of any approach. The complete process of the application of PCA can be observed in Fig. 1.

3.2. Description of QARGA

This section describes the main features of QARGA algorithm, which is used to validate the suitability of using the selected subset of measures as fitness function. QARGA is a real-coded genetic algorithm designed to discover existing relationships, specifically QAR, among several variables. This algorithm uses adaptive intervals instead of fixed ranges to represent the membership of the values of the attributes and a particular codification of the individuals that does not perform a previous attribute discretization. Moreover, it is not necessary to set which variables belong to the antecedent or consequent.

The search of the most appropriate intervals is carried out by means of an evolutionary process and the intervals are adjusted to find high-quality QAR. 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 in order to indicate whether it belongs to the antecedent or to the consequent or if does not belong to the rule. Thus, the representation of an individual consists of two data structures where the first structure includes all the attributes of the database and the second structure indicates the membership of an attribute to the rule represented by an individual.

The individuals of the population are subjected to an evolutionary process in which the crossover and mutation operators are applied. At the end of this process, the individual with the highest fitness is designated as the best rule. Moreover, the fitness function detailed in Section 3.3 was provided with a set of weights (w_s, w_c, w_n, w_a and w_r) so that the user can direct the search process depending on the desired rules. Thus, high values of w_s



Rule	CF	Conf	Gain	Lift	Sup. Ant	Sup. Rule	Sup. Con	Lev	Acc	Attr	Amp
0	1	1	0,02	1,14	0,14	0,14	0,98	0	0,75	5	0,15
1	0,96	0,99	0,11	1,18	0,18	0,18	0,83	0,03	0,9	2	0,4

Measure	Component 1	Component 2	Component 3
CF	-0,38	0,06	0,91
Conf	-0,02	-0,004	0,98
Gain	-0,61	0,58	0,51
Lift	0,06	0,74	0,02
Sup. Ant	0,96	-0,12	-0,18
Sup. Rule	0,97	-0,11	-0,16
Sup. Con	0,71	-0,68	-0,08
Lev	0,89	-0,07	-0,03
Acc	-0,22	0,91	-0,001
Attr	-0,97	0,11	0,13
Amp	0,82	-0,14	-0,26

Fig. 1. General scheme of the application of PCA.

might be used when rules fulfilled by many instances are preferred. Similarly, high values of w_c may be used when rules with low error are desired.

To obtain a set of rules a scheme based on the Iterative Rule Learning (IRL) [44] was proposed. In QARGA, this scheme is implemented by penalizing those instances matching a rule. The fitness function includes a weight of penalization so the instances already covered are less probable to appear in subsequent rules.

3.3. The fitness function of QARGA

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. The fitness function has been designed to maximize a combination of different measures of association rules.

The fitness function proposed in [31] to be maximized by QARGA was

$$f(rule) = w_s \cdot sup + w_c \cdot conf + w_n \cdot nAttrib - w_a \cdot ampl - w_r \cdot recov \quad (1)$$

where *sup* is the support of the rule, *conf* is the confidence of the rule, *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 belonging to the rule and w_s, w_c, w_r, w_n and w_a are weights to drive the process of search of rules, and will vary depending on the required rules. The *recov* term indicates the ratio of instances which had already been covered, as defined in [31].

Note that this function takes into consideration the support and the confidence of the rule, and therefore, QAR with high support and confidence are obtained. High values of the weight w_s imply more covered samples from the database, and high values of the weight w_c imply rules with higher reliability. Moreover, this fitness function includes a measure to bound the growth of the intervals during the evolutionary process. This is to avoid that the algorithm enlarges the amplitude of the intervals until the whole domain of each attribute is completed to obtain a great support. 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 penalize an instance that has already been covered by a previous rule.

However, the use of only support and confidence implies some drawbacks as previously discussed in Section 2.3. Thus, a new fitness function has been proposed based on the analysis described in Section 4.3.

4. Experimental results

The results obtained by the application of the QARGA algorithm with both original and new fitness functions to the datasets described in Section 4.2 are presented in this section. The goal of this experimentation is to show that better rules could be obtained when the measures selected with the help of PCA are considered in the fitness function.

First, a summary of the main parameters of configuration for QARGA can be found in Section 4.1. Section 4.2 provides a detailed description of all used datasets. The statistical analysis carried out to select the measures to be included in the function fitness is described in Section 4.3, according to the methodology described in the previous section. Section 4.4 includes the results obtained by QARGA when using the new and original fitness function and, finally, non-parametric statistical tests have been conducted to evaluate the significance of these results in Section 4.5.

4.1. Parameters configuration

It is noteworthy that all the weights appearing in both original and new fitness functions have been set to one to avoid the influence of the weights and thus to make possible the comparison of the results. QARGA has been executed five times for each dataset and the average results are shown for each fitness function. The main parameters of the QARGA algorithm are 100 for the number of rules, 100 for the size of the population, 100 for the number of generations, 0.0 for the minimum support and the minimum confidence and 0.8 for the probability of mutation.

4.2. Datasets description

This section presents the main features of the datasets used in the statistical study carried out by PCA. Several datasets have been tested from the public BUFA repository. Likewise, four different real-world datasets have been analyzed, specifically, datasets from biological, meteorological and seismological context.

- *Public datasets*
  Relevant information about thirty-five public datasets from BUFA repository [24] is summarized in Table 5. Note that Buying, Country, College, Education, Read and Usnews Colleges have been preprocessed using K-means Imputation method proposed in [17] (available in the KEEL tool [8]) in order to deal with missing values.
- *Biological dataset*
  The microarray dataset of Spellman [42] and Cho [12] for the budding yeast (Saccharomyces cerevisiae) cell-cycle has also been selected. This dataset considers a set of well-described genes, which encode important proteins for cell-cycle regulation. The budding yeast cell cycle microarray dataset consists of 20 attributes and 23 samples.
- *Tropospheric ozone*
  The tropospheric ozone of Seville (TOS) dataset is composed of climatological time series such as temperature, humidity, direction and speed of the wind, several variables such as the hour of the day and the day of the week and, finally, the tropospheric ozone. These variables have an influence on the ozone concentration in the atmosphere, the target agent. All variables have been retrieved from the meteorological station of the city of Seville in Spain for the months July to August during the years 2003 and 2004, generating a dataset with 7 quantitative attributes and 1488 instances. The reason for selecting such periods is due to the highest concentration of ozone there reported.
- *Total ozone content*
  Four datasets have been used containing a set of monthly average values including Total Ozone Content (TOC) [32] 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 8 quantitative attributes and 172 samples.
- *Earthquakes*
  The earthquake dataset [30,35] has been retrieved from the Spanish Geographical Institute (SGI). This dataset consists of four quantitative attributes and 873 instances related to the location and the magnitude of Spanish earthquakes from 1981 to 2008.

Table 5
Public datasets.

Dataset	Records	Attributes
Ailerons (AI)	7154	41
Baseball (BA)	337	17
Basketball (BK)	96	5
Bodyfat (FA)	252	18
Bolts (BL)	40	8
Buying (BU)	100	40
Computer activity (CA)	8192	22
Country (CN)	122	21
College (CO)	236	21
Education (ED)	1500	44
Elevators (EV)	16,599	19
Fried (FR)	40,768	11
House_16H (HH)	22,784	17
Kinematics (KI)	8192	9
Longley (LO)	16	7
Mortgage (MO)	1049	17
Normal body temperature (NT)	130	3
Plastic (PL)	1650	3
2Dplanes (PN)	40,768	11
Pw Linear (PW)	200	11
Pollution (PO)	60	16
Pole telecomm (PT)	9065	49
Pyramidines (PY)	74	28
Quake (QU)	2178	4
Read (RE)	681	26
School (SC)	62	20
Sleep (SL)	57	8
Stock price (SP)	950	10
Televisions (TV)	40	5
Treasury (TR)	1049	17
Triazines (TZ)	186	61
Usnews College (US)	1269	32
Vineyard (VY)	52	4
Weather Ankara (WA)	1609	11
Weather Izmir (WI)	1641	11

4.3. Selecting the measures to be included in the new fitness function

This section details the statistical analysis performed by the application of PCA. This analysis has been applied to the following quality measures: Support of the rule, support of the antecedent, support of the consequent, confidence, leverage, accuracy, lift, gain, certainty factor, amplitude and finally, number of attributes of the rule. Note that the conviction measure has been excluded because this measure reached the value infinity many times and, therefore, its use is not suitable for PCA. Only the analysis of principal components for the TOS dataset is now described, as the results obtained for the other datasets were similar. However, several relevant tables regarding the other datasets can be found in Appendix Appendix A.

Table 6 presents the value of the eigenvalues, the percentage of variance and the percentage of cumulative variance explained for each component obtained by PCA when it was applied to TOS dataset. In this case, three components were extracted and the total variance explained by the three components was 89.424%. In particular, the components one, two and three explain the 48.785%, 20.371% and 20.267% of the variance, respectively.

Table 7 describes the principal components extracted by PCA for the TOS dataset. It can be noticed that three principal components, each of them corresponding to an independent group of measures, have been obtained. Support of the antecedent, support of the rule, support of the consequent, leverage, amplitude and the number of

Table 6
Total explained variance according to three components for the TOS dataset.

Component	Eigenvalues	Variance (%)	Cumulative variance (%)
1	5.366	48.785	48.785
2	2.241	20.371	69.156
3	2.229	20.267	89.424

Table 7
Matrix of rotated components obtained by PCA for the TOS dataset.

Measure	Comp. 1	Comp. 2	Comp. 3
Accuracy	−0.224	**0.910**	−0.001
Certainty factor	−0.381	0.062	**0.914**
Confidence	−0.018	−0.004	**0.984**
Gain	**−0.606**	**0.579**	**0.515**
Lift	0.061	**0.738**	0.020
Antecedent support	**0.966**	−0.117	−0.187
Rule support	**0.972**	−0.117	−0.169
Consequent support	**0.702**	−0.681	−0.082
Leverage	**0.896**	−0.075	−0.029
Amplitude	**0.825**	−0.140	−0.267
Attributes	**−0.973**	0.113	0.128

Table 8
Summary of measures according to the principal components.

Group 1	Group 2	Group 3	Group 4
Anteced. support	Accuracy	Certainty factor	Gain
Rule support	Lift	Confidence	
Conseq. support			
Leverage			
Amplitude			
Attributes			

attributes of the rule belong to the first group because they are the most correlated measures in the first component. On the other hand, the second component is composed of the accuracy and lift measures. Finally, the certainty factor and the confidence belong to the third group due to the high correlation they present in it. Nevertheless, the measure gain has a similar correlation in the three components, and therefore, it could be considered the least dependent to the measures of the other groups.

Table 8 presents the grouped measures according to the principal components obtained when PCA was applied to the TOS dataset. It can be observed that there are three remarkable independent groups. The first group, which corresponds to the Group 1 column, is formed by the support of the antecedent, support of the rule, support of the consequent, leverage, amplitude and the number of attributes measures. The accuracy and lift belong to the second group represented by Group 2, and finally, Group 3 is composed of certainty factor and confidence measures. Note that the gain measure does not clearly belong to any of the components obtained by PCA as it has a similar correlation in the three components. Therefore, the gain has been considered as a new independent group and denoted as Group 4.

From Tables 6 and 8, it can be concluded that the support of the antecedent, support of the rule, support of the consequent, leverage, amplitude and the number of attributes measures corresponding to the Group 1 explain the highest percentage of variance (48.785%) versus the lowest percentages of variance explained by the two remaining groups (20.371% for accuracy and lift measures and 20.267% for certainty factor and confidence measures).

The next step is to select a representative measure for each component. The support of the rule, the confidence and the accuracy measures are chosen because they are the measures with the highest correlation in each group (0.972, 0.984 and 0.910, respectively). Moreover, the gain must also be added because this measure is equally distributed among all components, and therefore, all the measures could be optimized by maximizing this measure too. Therefore, the support of the rule, the confidence, the gain and the accuracy are the measures that characterize the full set of measures and will eventually form the fitness function.

From all the analysis discussed before, it can be concluded that the new fitness function to be maximized is given by the following equation:

$$f(rule) = w_s \cdot sup + w_c \cdot conf + w_g \cdot gain + w_a \cdot acc - w_r \cdot recov \quad (2)$$

where *gain* is the gain measure of the rule, *acc* is the accuracy value of the rule and w_g and w_a are weights to drive the process of search according to required rules. High values of the weight w_g induce a higher gain of information on the rules regarding the consequent when the antecedent is also present and high values of the weight w_a imply a higher accuracy and precision in the rules to be obtained.

This function is composed of the measures selected by PCA, that is, the support of the rule, the confidence, the gain and the accuracy measures instead of the use of only the support and the confidence measures. Every measure belongs to an independent group, and therefore, this set summarizes all the measures and simultaneously optimizes different properties of the QAR. It can be observed that the amplitude of the intervals or the number of attributes are not included in this function according to the results obtained by PCA.

The amplitude of the intervals is inversely proportional to the gain and the accuracy measures. Thus, if these measures are maximized, it is ensured that the intervals of attributes do not extend to the whole domain. On the other hand, the number of attributes is not necessary because the optimal number of attributes could be reached by maximizing the selected measures. In particular, the number of attributes is minimum when the support of the rule is maximized since both measures belong to the same component but with negative correlation. Finally, note that this function also includes a measure to penalize the instances that have already been covered by a previous rule as in the former function fitness.

4.4. Analysis of results

In this section, the results obtained by QARGA when optimizing the original and the new fitness functions using the datasets described in Section 4.2 are discussed.

The average results obtained for the five executions are summarized in Tables 9–11. *QARGA* 1 denotes QARGA when using the original fitness function and *QARGA* 2 is used to refer the QARGA algorithm when the new fitness function is maximized. It can be noted that the first four rows represent the real-world datasets, specifically, TOS, TOC, microarray for the budding yeast cell-cycle and earthquake datasets and the thirty-five remaining datasets are those that belong to the BUFA repository.

Table 9 shows the percentage of covered records, the average support, the average amplitude and the average number of attributes for the rules obtained by the *QARGA* 1 and *QARGA* 2 algorithms for all datasets. It can be noted that the average percentage of the records covered by the rules obtained by *QARGA* 2 are greater than that of *QARGA* 1, reaching values close to 80% even though the weight to penalize the covered instances was the same for both cases. However, the average support obtained by *QARGA* 2 is less than that obtained by *QARGA* 1. Equally remarkable is that the greatest difference between the average support of *QARGA* 1 and *QARGA* 2 occurs in those datasets with a larger number of attributes and viceversa.

Then, it can be concluded that *QARGA* 2 mined narrower rules with respect to the amplitude and with less number of attributes than those of *QARGA* 1. In terms of average values, these results lead to conclude that *QARGA* 2 discovers more specific rules than *QARGA* 1 by obtaining a lower number of attributes which helps to improve the comprehensiveness of the rules.

Table 9
Percentage of covered records, support of the rule, amplitudes and number of attributes of the rules.

Dataset	Records (%)		Rule support (%)		Rule amplitude (%)		Rule size	
	QARGA 1	QARGA 2	QARGA 1	QARGA 2	QARGA 1	QARGA 2	QARGA 1	QARGA 2
Tropospheric ozone	6.85	57.00	0.08	0.63	8.10	6.39	7.00	5.72
Total ozone content	38.82	71.88	0.59	1.09	8.25	5.83	8.99	5.63
Budding yeast cell-cycle	95.83	99.17	4.17	4.20	8.00	5.01	17.10	3.98
Earthquakes	79.40	100.00	1.10	2.12	8.36	6.82	4.00	3.72
Ailerons (AI)	99.86	88.82	4.46	0.92	7.76	4.10	23.06	14.22
Baseball (BA)	69.38	8.13	1.00	0.30	8.10	5.02	14.84	7.12
Basketball (BK)	64.58	70.00	1.09	1.51	8.04	5.53	5.00	4.22
Bolts (BL)	96.50	89.50	2.82	3.15	8.05	5.12	7.60	4.64
Buying (BU)	84.00	57.60	1.71	1.01	8.12	5.01	17.09	9.56
Computer activity (CA)	87.40	99.10	0.91	1.01	8.48	5.42	9.09	8.93
Country (CN)	87.38	67.21	1.56	0.91	8.14	5.13	15.92	6.82
College (CO)	88.64	41.86	1.29	0.55	8.12	5.09	10.38	8.24
Education (ED)	95.08	64.69	1.02	0.68	8.09	5.10	14.85	8.64
Elevators (EV)	99.91	98.80	6.50	1.56	7.75	4.04	8.86	7.22
Bodyfat (FA)	83.10	66.59	1.34	0.94	8.46	5.32	6.98	3.45
Fried (FR)	0.23	77.21	0.00	0.77	8.03	6.16	14.14	7.10
House_16H (HH)	95.29	99.61	0.99	1.00	9.07	5.63	3.00	2.86
Kinematics (KI)	1.07	40.47	0.01	0.41	8.03	5.67	3.00	2.96
Longley (LO)	100.00	100.00	6.50	6.51	8.06	5.13	9.18	6.19
Mortgage (MO)	79.52	90.26	1.82	1.56	8.51	6.06	14.47	5.24
Normal body temp. (NT)	85.08	100.00	1.58	2.19	8.22	6.45	4.55	15.21
Plastic (PL)	68.58	99.96	1.27	1.89	10.41	7.94	9.06	6.61
2Dplanes (PN)	10.34	90.65	0.11	1.84	8.26	6.78	19.07	8.27
Pollution (PO)	99.00	87.00	2.55	1.88	8.09	5.06	3.99	3.76
Pole telecomm (PT)	100.00	100.00	66.50	1.06	0.89	0.00	16.72	5.69
Pw Linear (PW)	41.40	67.20	0.52	1.05	8.00	5.31	7.79	4.56
Pyramidines (PY)	99.46	96.76	5.32	2.38	7.23	3.87	9.98	6.30
Quake (QU)	64.28	97.65	0.74	1.16	8.26	6.75	14.84	7.36
Read (RE)	81.62	16.65	0.95	0.23	8.20	5.09	5.00	3.56
School (SC)	90.82	90.49	2.11	1.68	8.01	5.04	30.83	16.18
Sleep (SL)	100.00	85.10	3.09	2.51	8.01	5.12	4.00	3.63
Stock price (SP)	39.54	59.71	0.65	0.79	8.60	6.39	8.93	6.29
Treasury (TR)	74.64	93.46	1.77	1.59	8.41	6.11	9.93	6.50
Televisions (TV)	97.50	99.50	4.62	3.46	8.17	5.18	26.62	7.85
Triazines (TZ)	100.00	57.96	7.17	1.62	7.11	2.80	17.09	7.07
Usnews College (US)	86.44	45.50	0.96	0.48	8.26	5.13	18.53	8.31
Vineyard (VY)	97.31	81.15	2.51	2.40	8.37	5.31	25.79	13.54
Weather Ankara (WA)	93.76	99.75	1.06	1.08	8.47	6.05	19.93	9.52
Weather Izmir (WI)	89.61	99.04	1.06	1.08	8.63	5.96	21.42	11.11
Average	**76.21** (± 29.56)	**78.34** (± 24.10)	**3.68** (± 10.49)	**1.57** (± 1.19)	**8.03** (± 1.27)	**5.33** (± 1.27)	**12.53** (± 7.12)	**7.12** (± 3.31)

Table 10 presents the average of the percentage of the *confidence*, *certainty factor* and *leverage* measures for the rules obtained by *QARGA* 1 and *QARGA* 2.

In contrast to the previous table, the average confidence obtained by *QARGA* 1 is greater than that of *QARGA* 2. However, both certainty factor and leverage measures for the rules discovered by *QARGA* 2 are better than that of *QARGA* 1. Therefore, it can be concluded that the rules obtained by *QARGA* 2 presents a strong dependency between the antecedent and consequent, although *QARGA* 1 presents better results in confidence. As discussed previously, the confidence has some drawbacks because it is not able to find negative dependencies between the antecedent and the consequent, hence it should not be considered particularly relevant if the confidence is slightly higher in *QARGA* 1.

It can be appreciated that the initial goal has been achieved since leverage values of the rules obtained by *QARGA* 2 are better than those of *QARGA* 1. Although leverage is not explicitly optimized in the new fitness function described in Section 3.3, this measure is improved in *QARGA* 2 due to its correlation with the support of the rule, which is optimized by the new fitness function. As stated in Section 1, the optimization of the selected measures by the methodology proposed using PCA did involve the optimization of other measures too.

Finally, Table 11 summarizes the lift, accuracy and gain measures for the QAR obtained by both *QARGA* 1 and *QARGA* 2 algorithms. Regarding the first four datasets, the lift, accuracy and gain measures of the rules discovered by *QARGA* 2 are higher, and therefore better in these datasets. With respect to the datasets from BUFA repository, *QARGA* 2 obtained better results in all datasets in terms of lift and accuracy, even getting better values for the gain measure in almost 95% of datasets. From its observation, it can be concluded that *QARGA* 2 discovers more accurate and interesting rules, and reaches higher information gain on the rules regarding the consequent when the antecedent is also present.

Note that lift values are better in *QARGA* 2 since similar conclusions regarding the leverage and support measures can be drawn. That is, lift and accuracy measures belong to the same group due to the correlation existing between them. Thus, lift is also optimized even if it was not explicitly included in the new fitness function.

Some other interesting conclusions can be drawn from these results. *QARGA* 2 presents less number of attributes, in other words, the new fitness function obtains more comprehensible rules helping the user to easily understand them. As for the rest of quality measures, although the confidence is slightly lower, *QARGA* 2 obtains better results producing interesting and precise rules with a high degree of dependence between the antecedent

Table 10
Confidence, certainty factor and leverage measures for rules obtained by QARGA with both original and new fitness function.

Dataset	Confidence (%)		Certainty factor		Leverage	
	QARGA 1	QARGA 2	QARGA 1	QARGA 2	QARGA 1	QARGA 2
Tropospheric ozone	99.12	96.71	0.99	0.96	0.0007	0.0021
Total ozone content	99.53	97.44	1.00	0.97	0.0056	0.0090
Budding yeast cell-cycle	100.00	98.10	1.00	0.98	0.0397	0.0401
Earthquakes	90.25	94.71	0.81	0.94	0.0016	0.0012
Ailerons (AI)	95.12	96.13	0.93	0.96	0.0007	0.0016
Baseball (BA)	99.91	99.70	1.00	1.00	0.0033	0.0030
Basketball (BK)	99.70	98.34	1.00	0.98	0.0098	0.0109
Bolts (BL)	99.80	86.37	1.00	0.86	0.0236	0.0293
Buying (BU)	99.95	99.90	1.00	1.00	0.0100	0.0100
Computer activity (CA)	96.94	94.70	0.95	0.95	0.0003	0.0002
Country (CN)	99.96	99.47	1.00	0.99	0.0089	0.0090
College (CO)	99.99	99.25	1.00	0.99	0.0043	0.0053
Education (ED)	99.97	99.80	1.00	1.00	0.0006	0.0030
Elevators (EV)	86.23	96.76	0.81	0.97	0.0029	0.0055
Bodyfat (FA)	99.96	99.09	1.00	0.99	0.0050	0.0052
Fried (FR)	100.00	99.49	1.00	0.99	0.0000	0.0000
House_16H (HH)	97.00	95.24	0.95	0.95	0.0002	0.0001
Kinematics (KI)	99.90	98.18	1.00	0.98	0.0001	0.0001
Longley (LO)	100.00	97.90	1.00	0.98	0.0597	0.0602
Mortgage (MO)	97.14	92.02	0.97	0.92	0.0166	0.0145
Normal body temperature (NT)	91.57	81.79	0.85	0.80	0.0075	0.0101
Plastic (PL)	99.14	99.86	0.99	1.00	0.0093	0.0077
2Dplanes (PN)	99.26	87.39	0.99	0.85	0.0003	0.0079
Pollution (PO)	99.98	98.67	1.00	0.99	0.0162	0.0167
Pole telecomm (PT)	84.52	81.04	0.16	0.80	0.0001	0.0003
Pw Linear (PW)	99.90	97.28	1.00	0.97	0.0043	0.0080
Pyramidines (PY)	95.56	99.67	0.94	1.00	0.0112	0.0187
Quake (QU)	94.24	90.67	0.86	0.90	0.0005	0.0016
Read (RE)	99.92	100.00	1.00	1.00	0.0023	0.0022
School (SC)	99.99	99.03	1.00	0.99	0.0174	0.0165
Sleep (SL)	99.43	97.45	0.99	0.97	0.0148	0.0223
Stock price (SP)	96.26	92.84	0.96	0.93	0.0060	0.0071
Treasury (TR)	97.16	92.51	0.97	0.92	0.0161	0.0148
Televisions (TV)	96.56	96.78	0.95	0.97	0.0281	0.0320
Triazines (TZ)	95.89	100.00	0.91	1.00	0.0041	0.0120
Usnews College (US)	99.96	99.60	1.00	1.00	0.0011	0.0032
Vineyard (VY)	93.85	99.80	0.93	1.00	0.0204	0.0230
Weather Ankara (WA)	97.98	93.68	0.96	0.94	0.0007	0.0008
Weather Izmir (WI)	98.31	92.88	0.97	0.93	0.0009	0.0009
Average	**97.44** ($\pm$3.77)	**95.90** ($\pm$4.85)	**0.94** ($\pm$0.14)	**0.96** ($\pm$0.05)	**0.0091** ($\pm$0.0123)	**0.0107** ($\pm$0.0126)

and the consequent. Therefore, to optimize the measures selected by the methodology proposed in this work leads to obtain better results instead of optimizing only the support and the confidence as most algorithms of the literature propose.

4.5. Statistical tests

Finally, a non-parametric statistical analysis [20] has been conducted to show if better results are really obtained when the set of selected measures are optimized. For this purpose, the support of the rule, confidence, leverage, lift, gain, accuracy and certainty factor measures obtained from the application of *QARGA* 1 and *QARGA* 2 to the real-world datasets and thirty-five datasets from BUFA repository have been calculated.

Specifically, the Wilcoxon test has been applied to detect significant differences in the measures of the rules obtained by *QARGA* 1 and *QARGA* 2. Let R^+ be the sum of ranks for the datasets in which the new fitness function (*QARGA* 2) outperformed the original one (*QARGA* 1), and R^- the sum of the opposite ranks. The results obtained by the Wilcoxon test for the level of significance $\alpha = 0.05$ are summarized in Table 12. The winner fitness function is stressed in bold in each row when the *p-value* associated is less than 0.05.

In the case of the confidence measure, the original fitness functions has presented better average results, the R^- value is greater than the R^+ and the *p-value* obtained is lower than the level of significance considered. Therefore, the test rejects the hypothesis concluding that the original fitness function is better than the new fitness function in terms of confidence.

Regarding the support of the rule and certainty factor, the R^- values are also greater than the R^+ values. However, the $p-$values obtained are greater than the level of significance considered, hence, the test accepts the hypothesis indicating that in terms of support and certainty factor the original fitness function and the new one do not present significant differences.

The new fitness function really outperforms the original fitness function in the rest of the measures considered in the Wilcoxon test, specifically, the measures accuracy, leverage, lift and gain. In all cases, the R^+ value is greater than the R^- and also, the $p-$ values obtained are less than the level of significance 0.05. Thus, the test rejects the hypothesis and it can be stated that there exist significant differences between the results obtained by both fitness functions.

5. Conclusions

A study based on the PCA method has been proposed to obtain the set of measures to be included in a fitness function to discover QAR. In particular, the support of the rule, confidence, gain and

Table 11
Lift, accuracy and gain measures for rules obtained by QARGA with both original and new fitness function.

Dataset	Lift		Accuracy (%)		Gain	
	QARGA 1	QARGA 2	QARGA 1	QARGA 2	QARGA 1	QARGA 2
Tropospheric ozone	42.02	489.20	89.56	99.27	0.89	0.95
Total ozone content	50.04	138.16	96.04	99.76	0.95	0.96
Budding yeast cell-cycle	22.83	23.40	99.48	99.83	0.95	0.94
Earthquakes	8.89	235.74	51.64	98.20	0.41	0.91
Ailerons (AI)	104.84	1739.72	72.49	99.36	0.66	0.95
Baseball (BA)	98.04	321.30	93.22	99.97	0.92	0.99
Basketball (BK)	30.63	86.94	91.62	99.82	0.90	0.97
Bolts (BL)	11.88	25.30	88.06	97.81	0.85	0.82
Buying (BU)	62.57	99.10	97.76	99.99	0.96	0.99
Computer activity (CA)	26.60	2110.06	58.25	99.65	0.54	0.93
Country (CN)	46.09	114.65	93.73	99.97	0.92	0.99
College (CO)	41.08	212.96	88.03	99.93	0.87	0.99
Education (ED)	45.01	1032.03	83.50	99.83	0.82	0.99
Elevators (EV)	107.84	1470.62	71.50	98.94	0.59	0.94
Bodyfat (FA)	27.24	211.03	92.37	99.82	0.91	0.98
Fried (FR)	198.03	4030.80	92.81	99.55	0.93	0.98
House_16H (HH)	7.64	3780.50	59.69	99.72	0.56	0.94
Kinematics (KI)	204.94	2757.58	93.64	99.72	0.94	0.97
Longley (LO)	14.20	15.39	98.63	99.68	0.92	0.91
Mortgage (MO)	23.83	116.48	94.52	99.47	0.90	0.90
Normal body temperature (NT)	2.97	16.36	57.17	92.97	0.47	0.73
Plastic (PL)	4.89	17.93	77.31	88.88	0.75	0.87
2Dplanes (PN)	3.26	9.58	60.04	85.49	0.59	0.71
Pollution (PO)	37.63	58.27	95.97	99.90	0.93	0.97
Pole telecomm (PT)	1.20	1457.33	70.72	97.04	0.03	0.77
Pw linear (PW)	13.89	66.72	83.25	97.89	0.83	0.94
Pyramidines (PY)	10.55	63.95	73.38	99.79	0.67	0.97
Quake (QU)	2.14	69.26	42.35	96.10	0.36	0.86
Read (RE)	66.26	571.19	94.34	99.90	0.93	1.00
School (SC)	44.02	59.14	97.61	99.93	0.95	0.97
Sleep (SL)	10.83	43.84	69.89	99.44	0.66	0.94
Stock price (SP)	31.06	222.12	93.85	99.57	0.89	0.92
Treasury (TR)	23.89	119.33	94.55	99.48	0.90	0.90
Televisions (TV)	13.03	31.83	73.54	99.18	0.66	0.93
Triazines (TZ)	5.29	70.50	51.61	96.64	0.44	0.95
Usnews College (US)	47.05	976.14	84.73	99.91	0.84	0.99
Vineyard (VY)	17.87	42.19	91.84	99.48	0.84	0.97
Weather Ankara (WA)	10.35	598.21	57.73	99.46	0.55	0.92
Weather Izmir (WI)	7.71	536.14	55.85	99.35	0.53	0.91
Average	**39.18** (± 47.07)	**616.44** (± 1009.51)	**80.31** (± 16.55)	**98.48** (± 3.02)	**0.75** (± 0.22)	**0.93** (± 0.07)

Table 12
Wilcoxon test to compare QARGA with the original and new fitness function.

Measure	QARGA 2 R^+	QARGA 1 R^-	*p-Value*
Confidence	189	**591**	4.2900E − 03
Accuracy	**780**	0	3.6380E − 12
Gain	**763**	17	7.5300E − 10
Leverage	**643**	137	2.3400E − 04
Lift	**780**	0	3.6380E − 12
Rule support	288	**492**	1.5821E − 01
Certainty factor	300	**480**	0.205031

accuracy are the measures that best summarize all the considered measures. Real-world climatological datasets, biological datasets and public datasets retrieved from the BUFA repository have been used to test the quality of the rules discovered by QARGA using a new fitness function that includes the set of selected measures. All the results show a remarkable performance of the new fitness function outperforming in many cases that of the original fitness function. The analysis of the quality measures has been very helpful to choose the most suitable objective function to be optimized by any algorithm, contrary to optimize only the support and confidence as most of the algorithms to discover QAR in the literature do. As future work, the authors want to analyze correlations between performance of different measures and attributes, as well as analyzing how some properties in datasets may have an influence on the quality measures performance.

Acknowledgments

The financial support from the Spanish Ministry of Science and Technology under project TIN2011-28956-C02 is acknowledged.

Appendix A. Statistical analysis performed by PCA in several datasets

This appendix extends Section 4.3 including the statistical analysis based on the PCA method, which has been undertaken on a wide range of data sets in addition to the TOS dataset.

Tables A1–A3 present the value of eigenvalues, the percentage of cumulative variance explained for each component obtained by PCA when it was applied to several datasets detailed in Section 4.2. Specifically, thirty-five datasets from BUFA datasets, TOC dataset and Earthquakes dataset, respectively.

Table A1
Total explained variance according to the three components for BUFA datasets.

Component	Eigenvalues	Variance (%)	Cumulative variance (%)
1	3.058	33.973	33.973
2	2.432	27.02	60.993
3	1.635	18.171	79.165

Table A2
Total explained variance according to the three components for the Total Ozone Content dataset.

Component	Eigenvalues	Variance (%)	Cumulative variance (%)
1	6.321	57.46	57.46
2	2.106	19.143	76.604
3	2.066	18.778	95.382

Table A3
Total explained variance according to the three components for the earthquakes dataset.

Component	Eigenvalues	Variance (%)	Cumulative variance (%)
1	4.261	38.737	38.737
2	2.55	23.183	61.92
3	2.254	20.487	82.407

Table A4
Matrix of rotated components obtained by PCA for BUFA datasets.

Measure	Comp. 1	Comp. 2	Comp. 3
Accuracy	0.238	0.237	**0.906**
Certainty factor	−0.099	**0.931**	0.117
Confidence	0.097	**0.863**	0.097
Gain	**−0.544**	**0.553**	**0.6**
Lift	−0.427	−0.208	**0.616**
Antecedent support	**0.809**	−0.525	0.021
Rule support	**0.854**	−0.425	0.012
Consequent support	0.598	−0.427	**−0.622**
Leverage	**0.789**	0.318	0.086
Amplitude	**0.779**	0.16	−0.178
Attributes	**−0.69**	0.108	0.302

Table A5
Matrix of rotated components obtained by PCA for the Total Ozone Content dataset.

Measure	Comp. 1	Comp. 2	Comp. 3
Accuracy	0.011	−0.012	**0.942**
Certainty factor	−0.069	**0.997**	−0.009
Confidence	−0.018	**0.999**	−0.015
Gain	**−0.808**	**0.334**	**0.468**
Lift	−0.157	−0.024	**0.835**
Antecedent support	**0.996**	−0.03	−0.04
Rule support	**0.997**	−0.024	−0.04
Consequent support	**0.852**	−0.01	−0.502
Leverage	**0.993**	−0.015	−0.04
Amplitude	**0.987**	−0.019	−0.071
Attributes	**−0.982**	0.031	0.018

Tables A4–A6 depict the principal components extracted after the application of PCA in thirty-five datasets from BUFA datasets, TOC dataset and earthquake dataset, respectively.

As can be observed from these tables, three principal components have been obtained, each of them corresponding to an independent group of measures. All measures are clearly in a specific component at least in any dataset except the gain measure that is not clear to which component does it belong in any dataset as it was described in Section 4.3.

Table A6
Matrix of rotated components obtained by PCA for the earthquakes dataset.

Measure	Comp. 1	Comp. 2	Comp. 3
Accuracy	0.29	**0.93**	−0.154
Certainty factor	−0.164	0.125	**0.944**
Confidence	0.118	−0.1	**0.962**
Gain	−0.18	**0.774**	**0.585**
Lift	−7.40E−05	**0.388**	0.064
Antecedent support	**0.97**	−0.07	−0.101
Rule support	**0.971**	−0.07	−0.095
Consequent support	0.293	**−0.94**	0.097
Leverage	**0.769**	0.103	0.163
Amplitude	**0.878**	−0.049	−0.014
Attributes	**−0.879**	0.055	0.11

References

[1] A. Abraham, E. Corchado, J.M. Corchado, Hybrid learning machines, Neurocomputing 72 (2009) 2729–2730.
[2] R. Agrawal, T. Imielinski, 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.
[3] R. Agrawal, 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] 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.
[5] B. Alatas, E. Akin, Rough particle swarm optimization and its applications in data mining, Soft Computing 12 (12) (2008) 1205–1218.
[6] 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.
[7] J. Alcalá-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.
[8] J. Alcalá-Fdez, L. Sánchez, S. García, M.J. del Jesus, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J.C. Fernández, F. Herrera, Keel a software tool to assess evolutionary algorithms for data mining problems, Soft Computing 13 (3) (2009) 307–318.
[9] S. Ayubi, M.K. Muyeba, A. Baraani, J. Keane, An algorithm to mine general association rules from tabular data, Information Sciences 179 (2009) 3520–3539.
[10] S. Brin, R. Motwani, C. Silverstein, Beyond market baskets: generalizing association rules to correlations, in: Proceedings of the ACM SIGMOD, vol. 26, 1997, pp. 265–276.
[11] S. Brin, R. Motwani, J.D. Ullman, S. Tsur, Dynamic itemset counting and implication rules for market basket data, in: Proceedings of the ACM SIGMOD, 1997, pp. 265–276.
[12] R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, R.W. Davis, A genome-wide transcriptional analysis of the mitotic cell cycle, Molecular Cell 2 (1998) 65–73.
[13] E. Corchado, A. Abraham, A. Carvalho, Hybrid intelligent algorithms and applications, Information Sciences 75 (14) (2010) 2633–2634.
[14] E. Corchado, M. Graña, M. Wozniak, New trends and applications on hybrid artificial intelligence systems, Neurocomputing 75 (2012) 61–63.
[15] S. Dehuri, A.K. Jagadev, A. Ghosh, R. Mall, Multi-objective genetic algorithm for association rule mining using a homogeneous dedicated cluster of workstations, American Journal of Applied Science 3 (2006) 2086–2095.
[16] M.J. del Jesús, J.A. Gámez, P. González, J.M. Puerta, On the discovery of association rules by means of evolutionary algorithms, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1 (5) (2011) 397–415.
[17] J. Deogun, W. Spaulding, B. Shuart, D. Li, Towards missing data imputation: a study of fuzzy k-means clustering method, in: 4th International Conference of Rough Sets and Current Trends in Computing(RSCTC'04), Lecture Notes on Computer Science, vol. 3066, 2004, pp. 573–579.
[18] A.E. Eiben, J.E. Smith, Introduction to Evolutionary Computing. Natural Computing Series, Springer, 2003.
[19] T. Fukuda, Y. Morimoto, S. Morishita, T. Tokuyama, Mining optimized association rules for numeric attributes, in: ACM Symposium on Principles of Database Systems, 1996, pp. 182–191.
[20] S. García, A. Fernández, J. Luengo, F. Herrera, A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability, Soft Computing 13 (10) (2009) 959–977.

[21] L. Geng, H.J. Hamilton, Interestingness measures for data mining: a survey, ACM Computing Surveys 38 (3) (2006) 1–42.
[22] A. Ghosh, B. Nath, Multi-objective rule mining using genetic algorithms, Information Science 163 (2004) 123–133.
[23] E.D. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley Publishing Company, 1989.
[24] H.A. Guvenir, I. Uysal, Bilkent university function approximation repository, ⟨http://funapp.cs.bilkent.edu.tr⟩, 2000.
[25] J. Han, J. Pei, Y. Yin, R. Mao, Mining frequent patterns without candidate generation: a frequent-pattern tree approach, 2004.
[26] M. Houtsma, A. Swami, Set-oriented mining for association rules, in: Proceedings of IEEE Data Engineering Conference, 1995.
[27] M. Kaya, R. Alhajj, Genetic algorithm based framework for mining fuzzy association rules, Fuzzy Sets and Systems 152 (3) (2005) 587–601.
[28] M. Kaya, R. Alhajj, Utilizing genetic algorithms to optimize membership functions for fuzzy weighted association rules mining, Applied Intelligence 24 (1) (2006) 7–152.
[29] D. Martín, A. Rosete, J. Alcalá-Fdez, F. Herrera, A multi-objective evolutionary algorithm for mining quantitative association rules, in: 2011 11th International Conference on Intelligent Systems Design and Applications (ISDA), 2011, pp. 1397–1402.
[30] F. Martínez-Álvarez, A. Troncoso, A. Morales-Esteban, J.C. Riquelme, Computational intelligence techniques for predicting earthquakes, in: Lecture Notes in Artificial Intelligence, 6679 (2), 2011, pp. 287–294.
[31] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, J.C. Riquelme, An evolutionary algorithm to discover quantitative association rules in multidimensional time series, Soft Computing 15 (10) (2011) 2065–2084.
[32] 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 109 (2) (2011) 217–227.
[33] J. Mata, J. Álvarez, J.C. Riquelme, Mining numeric association rules with genetic algorithms, in: Proceedings of the International Conference on Adaptive and Natural Computing Algorithms, 2001, pp. 264–267.
[34] J. Mata, J.L. Álvarez, J.C. Riquelme, Discovering numeric association rules via evolutionary algorithm, in: Lecture Notes in Artificial Intelligence, vol. 2336, 2002, pp. 40–51.
[35] A. Morales-Esteban, F. Martínez-Álvarez, J.L. de Justo, A. Troncoso, C. Rubio-Escudero, Pattern recognition to forecast seismic time series, Expert Systems with Applications 37 (12) (2010) 8333–8342.
[36] A. Orriols-Puig, J. Casillas, 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, 2008, pp. 2031–2038.
[37] V. Pachón Álvarez, J. Mata Vázquez, An evolutionary algorithm to discover quantitative association rules from huge databases without the need for an a priori discretization, Expert Systems with Applications 39 (1) (2012) 585–593.
[38] G. Piatetsky-Shapiro, Discovery, analysis and presentation of strong rules, in: Knowledge Discovery in Databases, 1991, pp. 229–248.
[39] H.R. Qodmanan, M. Nasiri, B. Minaei-Bidgoli, Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence, Expert Systems with Applications 38 (1) (2011) 288–298.
[40] A. Salleb-Aouissi, C. Vrain, C. Nortet, Quantminer: a genetic algorithm for mining quantitative association rules, in: International Joint Conference on Artificial Intelligence (IJCAI), 2007, pp. 1035–1040.
[41] E.H. Shortliffe, B. Buchanan, A model of inexact reasoning in medicine, Mathematical Biosciences 23 (1975) 351–379.
[42] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen P.O. Brown, D. Botstein, B. Futcher, Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization, Molecular Biology of the Cell 9 (1998) 3273–3297.
[43] Q. Tong, B. Yan, Y. Zhou, Mining quantitative association rules on overlapped intervals, in: Lecture Notes in Artificial Intelligence, vol. 3584, 2005, pp. 43–50.
[44] 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.
[45] X. Yan, C. Zhang, S. Zhang, Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support, Expert Systems with Applications 36 (2) (2009) 3066–3076.