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

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.
$Conf(A \Longrightarrow B) > Sup(B)$

$$CF(A \Longrightarrow B) = \frac{P(A \mid B) - P(B)}{1 - P(B)} = \frac{conf(A \Longrightarrow B) - sup(B)}{1 - sup(B)} \quad (7)$$

$Conf(A \Longrightarrow B) <= Sup(B)$

$$CF(A \Longrightarrow B) = \frac{P(A \mid B) - P(B)}{P(B)} = \frac{conf(A \Longrightarrow B) - sup(B)}{sup(B)} \quad (8)$$

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).

$$lev(A \Longrightarrow B) = P(A \cap B) - P(A)P(B) = sup(A \Longrightarrow B) - sup(A)sup(B) \quad (9)$$

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 multi-objective 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)