Quantitative Association Rules Applied to Climatological Time Series Forecasting

M. Martínez-Ballesteros[1], F. Martínez-Álvarez[2], A. Troncoso[2], and J.C. Riquelme[1]

[1] Department of Computer Science, University of Seville, Spain
{mariamartinez,riquelme}@us.es
[2] Area of Computer Science, Pablo de Olavide University of Seville, Spain
{fmaralv,ali}@upo.es

Abstract. This work presents the discovering of association rules based on evolutionary techniques in order to obtain relationships among correlated time series. For this purpose, a genetic algorithm has been proposed to determine the intervals that form the rules without discretizing the attributes and allowing the overlapping of the regions covered by the rules. In addition, the algorithm has been tested on real-world climatological time series such as temperature, wind and ozone and results are reported and compared to that of the well-known Apriori algorithm.

Keywords: Time series, forecasting, quantitative association rules.

1 Introduction

The prediction of the temporal evolution of variables –time series forecasting– is typically carried out by means of statistical methods. Despite the good performance and inherent simplicity presented by these methods in synthetic data, when applying to real-world time series the results are not as satisfactory as expected due to the non-linear features that such data exhibit.

The existence of other time series correlated with the one under study is an usual phenomenon. In the field of climatological times series, for instance, it is necessary to evaluate time series such as temperature, humidity or atmospheric pressure in order to forecast if it will rain or not. Thus, the problem faced in this work consists in forecasting the behavior of a time series by obtaining association rules among all the existing correlated time series. Concretely, the time series aimed to be forecasted is the tropospheric ozone, which is an atmospheric constituent classed as pollutant when it exceeds a certain threshold. The variation of the concentration of this agent in the air is under continuous analysis, since it is well known the noxious effects that may cause in both human beings and nature [5].

The goal of the association rules extraction process consists, basically, in discovering the presence of pair conjunctions (attribute (A) – value (v)) that appear in a dataset with a certain frequency in order to formulate the rules that display the existing relationship among the attributes. Formally, an association rule is

E. Corchado and H. Yin (Eds.): IDEAL 2009, LNCS 5788, pp. 284–291, 2009.

a relationship between attributes in a database in the way $C_1 \Rightarrow C_2$, where C_1 and C_2 are pair conjunctions such as $A = v$ if $A \in \mathbb{Z}$ or $A \in [v_1, v_2]$ if $A \in \mathbb{R}$. Generally, the antecedent C_1 is formed by a the conjunction of multiple pairs and the consequent C_2 is usually a single pair.

There exist many efficient algorithms that find these rules. However, many researchers are focused on databases with discrete attributes while most real-world databases comprise essentially continuous attributes, as it happens in time series analysis. Moreover, the majority of the tools said to work in the continuous domain just discretize the attributes using a specific strategy and, then, treat these attributes as if they were discrete [6]. The main motivation of this research is to develop a genetic algorithm (GA) able to find association rules over databases with continuous attributes avoiding the discretization as a previous step of the process.

A revision of the recently published literature reveals that the amount of works that provide metaheuristics and search algorithms related to association rules with continuous attributes is scant. Thus, a classifier was presented in [4] with the aim of extracting quantitative association rules over unlabeled data streams. The main novelty of this approach lied on its adaptability to on-line gathered data. An optimization metaheuristic based on rough particle swarm techniques was presented in [1]. In this case, the singularity was the obtention of the values that determine the intervals for the association rules. They also evaluated and tested several new operators in synthetic data. A multi-objective pareto-based GA was presented in [2]. The fitness function was formed by four different objectives: support, confidence, comprehensibility of the rule (aimed to be maximized) and the amplitude of the intervals that forms the rule (intended to be minimized). The work published in [9] presented a new approach based on three novel algorithms: value-interval clustering, interval-interval clustering and matrix-interval clustering. The application of them was found specially useful when mining complex information. Finally, another GA was used in [8] in order to obtain numeric association rules. However, the unique objective to be optimized in the fitness function was the confidence. To fulfill this goal, the authors avoided the specification of the actual minimum support, which is the main contribution of this work.

The rest of the paper is divided as follows. Section 2 provides the methodology used in this work. The results of the approach are discussed in Section 3. Finally, Section 4 describes the achieved conclusions.

2 Description of the Search of Rules

In a continuous domain, it is necessary to group certain sets of values that share same features and, as a consequence, it is required to be able to express the membership of the values to each group. No fixed ranges but intervals of confidence have been chosen to represent the membership of such values in this work. The search of the most appropriate intervals is carried out by means of a GA. Thus, the intervals are adjusted to find the association rules with high values for both support and confidence, together with other measures used in order to quantify the quality of the rule.

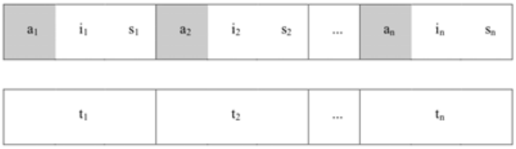

Fig. 1. Representation of an individual of the population

In the population, each individual constitutes a rule. These rules are then subjected to an evolutionary process in which both mutation and crossover operators are applied and, at the end of the process, the individual that presents the best fitness is designated as the best rule. Moreover, the fitness function has been provided with a set of parameters in order to the user can drive the process of search depending on the desired rules. The punishment of the covered instances allows the subsequent rules found with the GA to try to cover those instances that were still uncovered, by means of an Iterative Rule Learning (IRL) [7].

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

2.1 Codification of the Individuals

Each gene of an individual represents the upper and lower limit of the intervals of each attribute. The individuals are represented by a real codification since the values of the attributes are continuous. Each individual is formed by a variable number of attributes, which has to be lower than n, where n is the number of attributes belonging to the database.

Two structures are available for the representation of an individual, as it is shown in Fig. 1. Note that all the attributes included in the database are depicted in the upper structure. The limits of the intervals of each attribute are stored in this structure, where i_i is the inferior limit of the interval and s_i the superior one.

Nevertheless, not all the attributes will be present in the rules that describe an individual. A second structure indicating the type of each attribute, shown in the lower part of the Fig. 1, has been developed with the aim of improving the efficiency. Note that t_i can have three different values: 0 when the attribute does not belong to any individual, 1 when the attribute belongs to the antecedent and 2 when it belongs to the consequent. Therefore, if an attribute is wanted to be retrieved for a specific rule, it can be done by modifying the value equal to 0 of the type by a value equal to 1 or 2.

2.2 Generation of the Initial Population

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

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

2.3 Genetic Operators

The genetic operators used in the proposed algorithm are: selection, crossover and mutation

1. *Selection.* An elitist strategy is used replicating thus the individual with the best fitness and a roulette selection-based method for the remaining individuals rewarding the best individuals according to their fitness.
2. *Crossover.* Two parent individuals, chosen by means of the roulette selection, are combined to generate a new individual. First, all the attributes associated to each parent are analyzed in order to discover their type. Then, if the same attribute in both parents belonged to the same type of attribute, this type of attribute would be assigned to the descendent and the interval is obtained generating two random numbers among the limits of the intervals of both parents. Thus, the lower interval is generated by means of a random number that belongs to the interval formed by both lower intervals of the parents; the upper interval is analogously calculated. Otherwise, one of the two types would be randomly chosen between both parents, without modifying the intervals of such attribute.
3. *Mutation.* It consists in varying one gene of the individuals. The mutation can be focused on the type of the attribute (antecedent to consequent, consequent to antecedent or antecedent or consequent to null) or on the intervals, in which three different cases are possible: equiprobable mutation of the upper limit, of the lower limit or of both limits of the interval. For this aim, a random value between 0 and the maximum amplitude is generated and it will be added or subtracted to the limit of the interval which is randomly selected.

2.4 The Fitness Function

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

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

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

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

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

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

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

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

3 Results

The proposed algorithm has been applied to discover association rules between temperature, wind and ozone time series from June 2003 to September 2003. Note that for the prediction task, the temperature and wind are forced to be in the antecedent and the ozone in the consequent, obtaining thus an approximate prediction on the basis of these rules.

Several experiments have been carried out in order to validate the behavior of the proposed operators. The parameters of the algorithm are initially set with default values although a more exhaustive analysis should be performed to establish the optimum set of values. The main parameters of the GA are as follows: 100 for the size of the population, 100 for the number of generations; 20 for the number of rules to be obtained and 0.8 for the mutation probability. The weights of the fitness function are: 2 for w_s, 0.5 for w_c, 1 for w_r, 0.2 for w_n and 1.2 for w_a.

The reason for assigning a high value to the weight w_s is to cover the maximum number of examples by the obtained rules. However, the weight associated to the confidence is lower since it is impossible to obtain rules with a great confidence

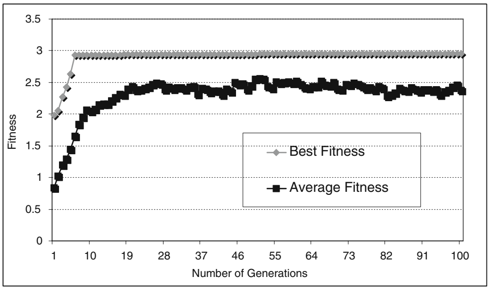

Fig. 2. Evolution of the best rule and the average population

Table 1. Description of the rules found by the proposed GA

Rules	Description
R1	temperature $\in$ [28.5,32.2] $\Longrightarrow$ ozone $\in$ [112.7,139.3]
R2	temperature $\in$ [31.1,34.8] $\Longrightarrow$ ozone $\in$ [119.0,145.8]
R3	temperature $\in$ [25.3,29.0] $\Longrightarrow$ ozone $\in$ [97.7,124.0]
R4	temperature $\in$ [22.6,26.3] $\Longrightarrow$ ozone $\in$ [103.0,128.7]
R5	temperature $\in$ [20.4,23.0] and wind $\in$ [13.0,15.5] $\Longrightarrow$ ozone $\in$ [91.5,115.5]

due to the existence of a lot of noise in the dataset. The weight associated to the instances covered by other rules as well as the amplitude of the intervals are moderately high in order to penalize the rules whose intervals are too large and are also covering samples already covered by other rules (remind that the goal is to cover all the dataset). The weight associated to the number of attributes has been set with a small value in order to allow the rule to comprise as many attributes as necessary.

Figure 2 shows the evolution of the fittest individual rule and the average of the population throughout the evolutionary process for 10 runs. It can be noticed that the initial set of rules improves its quality all over the generations.

Table 1 shows the five rules selected among the twenty rules found by the GA. It can be noticed that four of them have just two attributes, which are the temperature (in the antecedent) and the ozone (in the consequent). This fact reveals that the temperature provides more information about the ozone than the wind. The fifth selected rule has two attributes in the antecedent –the temperature and the wind– and the ozone in the consequent. Equally remarkable is the possibility of finding rules that have overlapping but covering the whole domain of the consequent, to which the majority of instances belong to. On the other hand, the amplitude of the intervals is similar for all the discovered rules, showing the stability of the proposed algorithm.

Table 2 presents three measures for each rule shown in Table 1. The *Confidence* column indicates the percentage of samples covered by the rule among those samples that only cover the antecedent. The second column, *Covered*, shows the number of samples covered by each rule which is directly related to the support. The *Amplitude* column presents the average amplitude of the intervals for each rule. As it can be observed, the confidence in most cases, despite the small associated weight, is greater than 50% (and even greater than 70% in some cases),

Table 2. Measures for the rules obtained using the GA

Rules	Confidence (%)	Covered	Amplitude
R1	50.8	159	15.1
R2	47.8	143	15.2
R3	54.8	135	15.0
R4	56.0	84	14.7
R5	72.7	8	9.7

Table 3. Description of the rules found by the Apriori algorithm

Rules	Description
R1	temperature $\in$ [24.49,27.42] $\Longrightarrow$ ozone $\in$ [100.76,121.54]
R2	temperature $\in$ [30.35,33.28] $\Longrightarrow$ ozone $\in$ [121.54,142.32]
R3	temperature $\in$ [27.42,30.35] $\Longrightarrow$ ozone $\in$ [100.76,121.54]
R4	wind $\in$ [11.36,14.2] $\Longrightarrow$ ozone $\in$ [121.54,142.32]
R5	wind $\in$ [11.36,14.2] $\Longrightarrow$ ozone $\in$ [100.76,121.54]

Table 4. Measures for the rules obtained using the Apriori algorithm

Rules	Confidence (%)	Covered	Amplitude
R1	42	71	11.8
R2	41	93	11.8
R3	39	88	11.8
R4	29	59	11.8
R5	27	55	11.8

which means that the reached error by the rules can be considered satisfactory. The number of covered samples is much greater with two-attributes rules (more than 100 samples in most cases) than with those with three attributes. Moreover, the average amplitude of the intervals is approximately 14, which is a good result when predicting ozone.

The Apriori algorithm [3] implemented in WEKA has been applied in order to obtain association rules with the purpose of establishing a comparison between the results of the proposed algorithm and that of the Apriori algorithm. Before applying the Apriori algorithm, the temperature, wind and ozone datasets have been discretized because this algorithm only can handle categorical attributes. The rules obtained by this algorithm are shown in Table 3. Note that all the generated rules comprise only two attributes. It can be observed that there are different rules with the same prediction interval for the ozone, e. g., $R1$, R_3 and R_5, and R_2 and R_4. Finally, it is worth noting that these rules do not cover the interval from 90 to 100 in which the dataset has many instances, while the proposed algorithm does.

Table 4 is equivalent to Table 2 but when applying the Apriori algorithm. With regard to the confidence, no rules have values greater than 50% which implies that the rules provide a prediction error greater than that of the proposed algorithm in most cases. The number of instances covered by the rules provided by the proposed approach is greater than that of the Apriori algorithm, obtaining rules with better support. With reference to the average amplitude of the intervals, both algorithms have a similar behavior. Finally, no rules with three attributes have been found using the Apriori algorithm.

4 Conclusions

A new GA has been proposed in this work in order to discover association rules among correlated real-world time series. This algorithm has determined the intervals that form the rules without discretizing the attributes and allowing the overlapping of the regions covered by the rules. When predicting the ozone time series with the new approach, the obtained error is lower than the one provided by the well-known Apriori algorithm, since the confidence of the rules generated by the GA is greater than that of the Apriori algorithm.

Acknowledgments

The financial support from the Spanish Ministry of Science and Technology, project TIN2007-68084-C-00, and from the Junta de Andalucía, project P07-TIC-02611, is acknowledged.

References

1. Alatas, B., Akin, E.: Rough particle swarm optimization and its applications in data mining. Soft Computing 12(12), 1205–1218 (2008)
2. Alatas, B., Akin, E., Karci, A.: MODENAR: Multi-objective differential evolution algorithm for mining numeric association rules. Applied Soft Computing 8(1), 646–656 (2008)
3. Kotsiantis, S., Kanellopoulos, D.: Association rules mining: A recent overview. GESTS International Transactions on Computer Science and Engineering 32(1), 71–82 (2006)
4. Orriols-Puig, A., Casillas, J., Bernadó-Mansilla, E.: First approach toward on-line evolution of association rules with learning classifier systems. In: Proceedings of the 2008 GECCO Genetic and Evolutionary Computation Conference, pp. 2031–2038 (2008)
5. Sahua, S.K., Yipc, S., Hollandb, D.M.: Improved space-time forecasting of next day ozone concentrations in the eastern US. Atmospheric Environment 43(3), 494–501 (2009)
6. Vannucci, M., Colla, V.: Meaningful discretization of continuous features for association rules mining by means of a som. In: Proceedings of the European Symposium on Artificial Neural Networks, pp. 489–494 (2004)
7. Venturini, G.: SIA: a Supervised Inductive Algorithm with genetic search for learning attribute based concepts. In: Brazdil, P.B. (ed.) ECML 1993. LNCS, vol. 667, pp. 280–296. Springer, Heidelberg (1993)
8. Yan, X., Zhang, C., Zhang, S.: Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. Expert Systems with Applications: An International Journal 36(2), 3066–3076 (2009)
9. Yin, Y., Zhong, Z., Wang, Y.: Mining quantitative association rules by interval clustering. Journal of Computational Information Systems 4(2), 609–616 (2008)