Labeling Association Rule Clustering through a Genetic Algorithm Approach

Renan de Padua, Veronica Oliveira de Carvalho, and Adriane Beatriz de Souza Serapião

Instituto de Geociências e Ciências Exatas,
UNESP - Univ Estadual Paulista, Rio Claro, Brazil
paduarenanemail@gmail.com, {veronica,adriane}@rc.unesp.br

Abstract. Among the post-processing association rule approaches, a promising one is clustering. When an association rule set is clustered, the user is provided with an improved presentation of the mined patterns, since he can have a view of the domain to be explored. However, to take advantage of this organization, it is essential that good labels be assigned to the groups, in order to guide the user during the exploration process. Moreover, few works have explored and proposed labeling methods to this context. Therefore, this paper proposes a labeling method, named GLM (*G*enetic *L*abeling *M*ethod), for association rule clustering. The method is a genetic algorithm approach that aims to balance the values of the measures that are used to evaluate labeling methods in this context. In the experiments, GLM presented a good performance and better results than some other methods already explored.

Keywords: Association Rules, Clustering, Labeling Methods, Genetic Algorithm.

1 Introduction

One of the most studied topics in data mining is association mining due to its ability to discover all the frequent relationships that occur among the data set items. The main problem related to this topic is the huge number of patterns that are obtained. Usually, only few of them are of really interesting to the user. To overcome this problem, many approaches have been proposed, being clustering a promising one. In this case, after the rules are obtained, they are grouped in n groups, each one representing a different view of the domain. The idea of the works that use clustering in post-processing is to improve the presentation of the mined patterns, providing the user a view of the domain to be explored, as seen in [1,2,3,4]. However, the grouping becomes useful only if good labels exist, in order to allow an easier browsing of the domain.

Finding good labels is a relevant issue. It is important, for example, that good labels be presented to the user to facilitate exploratory analyses, interesting when the user doesn't have, a priori, an idea where to start. However, few works have explored and proposed labeling methods to the context of association rule

B. Catania et al. (eds.), *New Trends in Databases and Information Systems*,
Advances in Intelligent Systems and Computing 241,
DOI: 10.1007/978-3-319-01863-8_5,

clustering. In [5] a discussion about some methods and their performances, in this context, is done. Since the authors didn't identify any evaluation methodology to check the performance of the methods, two measures were proposed by them. As a result of their study, [5] noticed that none of the methods provide good results in both of the measures, observing that there is a considered difference between their values.

Based on the exposed, this work proposes a labeling method for association rule clustering. The method, named GLM (*G*enetic *L*abeling *M*ethod), is a genetic algorithm approach, since the problem was treated as an optimization one. Its optimization function aims to balance the values of the measures that are used to evaluate labeling methods in this context. In the experiments, GLM presented a good performance and better results than some other methods already explored.

This paper is organized as follows. Section 2 surveys the related works and gives an overview of some labeling methods and evaluation measures. Section 3 describes the proposed method. Sections 4, 5 and 6 present, respectively, the experiments, the results and the conclusions.

2 Background

In this section, works related to the topics covered by this paper are first reviewed. Then, the labeling methods used in this work to compare the performance of GLM and the evaluation measures used in the optimization function are discussed.

Association Clustering. Different clustering strategies have been used for post-processing association rules. In [1] the grouping is done through partitional and hierarchical algorithms using Jaccard, expressed by $\text{J.RT}(r,s) = \frac{|\{t \text{ matched by } r\} \cap \{t \text{ matched by } s\}|}{|\{t \text{ matched by } r\} \cup \{t \text{ matched by } s\}|}$, as the similarity measure – the Jaccard between r and s considers the common transactions (t) the rules match. A rule matches a transaction t if all the rule items are contained in t. Furthermore, the authors select as labels of each group the items that appear in the rule which is more similar to all the other rules in the group. In [2] the grouping is done through hierarchical algorithms also using Jaccard as the similarity measure. However, in their work, the Jaccard between two rules r and s, expressed by $\text{J.RI}(r,s) = \frac{|\{items \text{ in } r\} \cap \{items \text{ in } s\}|}{|\{items \text{ in } r\} \cup \{items \text{ in } s\}|}$, is computed considering the items the rules share. To label the groups, the same strategy of [1] is used. [4] propose a similarity measure based on transactions and uses a density algorithm to carry out the clustering. In this case, the authors don't mention how the labels are found. [3] also proposes a similarity measure based on transactions, although uses a hierarchical algorithm to carry out the clustering. At the end of the process, the author proposes an approach to summarize each cluster by finding the patterns $a \Rightarrow c$ that cover all the rules in the cluster. Works that combine labeling methods and genetic algorithms, in association context, were not found.

Labeling Methods. The papers related to association rule clustering have not sorely explored the labeling issue, as noticed by [5]. The four labeling methods

presented in [5], used to this context, are briefly described. These methods were used here to allow a comparative analysis of GLM performance. In the *L*abeling *M*ethod *M*edoid (LM-M), the labels of each cluster are built by the items that appear in the rule of the group that represents the medoid of the group. In the *L*abeling *M*ethod *T*ransaction (LM-T), the labels of each cluster are built by the items that appear in the rule of the group that covers the largest number of transactions. A rule covers a transaction t if all the rule items are contained in t. In the *L*abeling *M*ethod *S*ahar (LM-S), the labels of each cluster are built by the items that appear in the pattern $a \Rightarrow c$ that covers the largest number of rules. A pattern $a \Rightarrow c$ covers a rule $A \Rightarrow C$ if $a \in A$ and $c \in C$. Finally, in the *L*abeling *M*ethod *P*opescul & *U*ngar (LM-PU), the labels of each cluster are built by the N items that are more frequent in their own cluster and infrequent in the other clusters.

Evaluation Measures. [5] propose in their work two measures, Precision and Repetition Frequency, that allow an evaluation of labeling methods in the context of association rule clustering. Since any other measures were found to this context, these are the measures used in the fitness function of GLM. Both of the measures range from 0 to 1. Precision (P), expressed by $P(C) = \frac{\sum_{i=1}^{\#Groups} P(C_i)}{\#Groups}$, $P(C_i) = \frac{\#\{rules\ covered\ in\ C_i\ by\ C_i\ labels\}}{\#\{rules\ in\ C_i\}}$, measures how much the labeling method can generate labels that really represent the rules contained in the clusters. It is expected that a good method must have a high precision. However, it is not enough to be precise if the labels appear repeatedly among the clusters. Therefore, Repetition Frequency (RF), expressed by $RF(C) = 1 - \frac{\#\{distinct\ labels\ that\ repeat\ in\ the\ clusters\}}{\#\{distinct\ labels\ in\ the\ clusters\}}$, measures how much the distinct labels that are present in all the clusters don't repeat. The higher the RF value, the better the method, i.e., less repetitions implies in better performance.

3 GLM: The *G*enetic *L*abeling *M*ethod

GLM is a genetic algorithm approach for labeling association rule clustering. In this proposed labeling method, the labels of the clusters are built by the items that appear in the rules of the groups that ensure a good tradeoff between Precision (P) and Repetition Frequency (RF). Only P and RF were considered since other evaluation measures were not found. Thus, since the problem was treated as an optimization one, the genetic algorithm approach was adopted. The solution was motivated by the fact that none of the methods discussed in [5] provided good results, at the same time, in P and RF. Thereby, a method that yields ways to maximize interesting evaluation measures is a promising one. To understand GLM, the description of the genetic operators and other important aspects are following discussed. For details about the concepts see [6].

Encoding. In GLM, each individual represents a possible solution to the problem, i.e., the labels of each group in an association rule clustering. For that, each individual is composed by n chromosomes, where n represents the number of groups in the clustering given as input. Each chromosome has m genes, where m represents the maximum number of labels to be assigned to each group. m

is a value informed by the user. Although all the chromosomes have the same length m, m is the maximum number of labels a group can have. Thus, in some chromosomes, not all of its genes are filled.

Initialization. Given an association rule clustering, an initial population is generated. A population is composed by PS individuals, where PS (Population Size) is given by the user. To create each individual a looping is done, where each iteration is related to a chromosome (group). The choice of the items to be selected as labels (genes), in each chromosome, is done randomly. However, only the items that appear in the rules of the current group are considered. During this process, it is assured that a group (chromosome) can not contain repeated items in its labels (genes).

Genetic Operators. The genetic operators used in GLM, as well, the fitness function and the termination criterion are described below.

A. Selection. The roulette wheel is used to select two individuals to obtain an offspring. For that, the fitness of each individual is considered as its chance to be selected. The higher the fitness the higher the probability an individual has to be selected.

B. Crossover. The uniform crossover is used to obtain an offspring. The unique offspring is generated from the parents with the help of a bit mask, which is obtained for each chromosome. The bit mask is a sequence of 0's and 1's, which indicates from which parent the gene has to be copied. When the value is 0, the offspring inherits the gene from parent 1 and when is 1 from parent 2. Thus, the resulting offspring contains a mixture of genes from both parents. The bit mask is randomly obtained as a vector of bits: when one parent has more filled genes (labels) than the other, the bit mask in these not overlapped positions receives the code related to the filled parent. This fact justifies our choice to obtain a unique offspring: if two offspring were generated, they would be very similar to their parents.

C. Mutation. In offspring, the genes of chromosomes occasionally change with a probability MP (Mutation Probability). Only one gene of each chromosome has a chance to be mutated. Thus, for each chromosome, a probability is randomly obtained and compared with MP to check if the mutation will occur in the chromosome. If so, a gene in the chromosome is randomly chosen and the mutation is done.

D. Fitness Function. Since GLM aims to obtain labels that ensure a good tradeoff between Precision (P) and Repetition Frequency (RF), the fitness function of an individual I is defined by $Fitness(I) = (P+RF) - \left(\frac{Max(P,RF)}{Min(P,RF)} * 10^{-5}\right)$. Initially, P and RF are added. However, as 1.0 can be obtained by $P = 0.2$ and $RF = 0.8$ or by $P = 0.5$ and $RF = 0.5$, for example, it is necessary to penalize individuals that present a high variation between the measures to ensure a good tradeoff. The normalized penalization adopted in this work is obtained dividing the measure that has the maximum value (Max) by the one that has the minimum (Min) and, then, normalizing the result with 5 digits of precision (10^{-5}). 10^{-5} represents the ratio $\frac{0.00001}{1.00000}$, in which 0.00001 indicates the minimum value a measure can reach and 1.00000 the maximum. As mentioned before, only P

and RF were considered to define the fitness function, since other evaluation measures were not found to this context. However, as new measures arise, they can also be added to GLM.

E. Termination. GLM stops when the number of iterations, i, is larger than a given number NG (Number of Generations).

The GLM steps are presented in Algorithm 1. The algorithm receives 5 parameters: (i) an association rule clustering (ARC); (ii) the population size (PS); (iii) the mutation probability (MP); (iv) the number of generations (NG); (v) the maximum number of labels to be assigned to a group (m). m gives, in fact, the number of genes the chromosomes will have. At the end of the process, the clustering given as input is outputted to the user with its labels. As seen in Algorithm 1, GLM works as follows: initially, the ARC is loaded to the memory (line 1). After that, a population is created with PS individuals (line 2). Until the stopped criterion is not reached (line 3), the operators of selection (line 4), crossover (line 5) and mutation (line 8) are applied. Before starting a new iteration, the population is updated (line 11), i.e., the offspring is added and the parent with the lowest fitness removed. In the end, the individual with the best fitness represents the solution.

Algorithm 1. The GLM steps.

Input: ARC, PS, MP, NG, m.
Output: A labeled association rule clustering.
1: Read ARC
2: Initialize population with PS individuals
3: **for** 1 to NG **do**
4: Select two individuals I_1 and I_2
5: Crossover I_1 and I_2 to obtain an offspring O
6: **for** each O chromosome **do**
7: **if** ($RN < MP$) **then**
8: Mutate O chromosome, where RN is a $r$andom $n$umber in the range [0,1]
9: **end-if**
10: **end-for**
11: Update population: Add O to population; Remove the parent (I_1;I_2) with the lowest fitness
12: **end-for**

4 Experiments

Some experiments were carried out in order to analyze GLM performance (GLM's quality assessment). Thus, initially, it was necessary to generate some association rule clusterings (ARC). Forty organizations were selected to obtain 40 ARCs for each one of the four data sets used.

The four data sets were Adult (48842;115), Income (6876;50), Groceries (9835;169) and Sup (1716;1939). The numbers in parenthesis indicate, respectively, the number of transactions and the number of distinct items in each data set. The first three are available through the package "arules"[1]. The last one was donated by a supermarket located in São Carlos city, Brazil. All the transactions in Adult and Income contain the same number of items (named here as

[1] `http://cran.r-project.org/web/packages/arules/index.html`.

*s*tandardized *d*ata *s*ets (SDS)), different from Groceries and Sup (named here as *n*on-*s*tandardized *d*ata *s*ets (NSDS)), whereupon each transaction contains a distinct number of items. Thus, the experiments considered different data types. The rules, in each data set, were mined using an *Apriori* implementation[2] with a minimum of 2 and a maximum of 5 items per rule. With the Adult set 6508 rules were extracted using a minimum support (min-sup) of 10% and a minimum confidence (min-conf) of 50%; Income 3714 rules with min-sup=17%, min-conf=50%; Groceries 2050 rules with min-sup=0.5%, min-conf=0.5%; Sup 7588 rules with min-sup=0.7%, min-conf=0.5%.

To cluster these four rule sets, forty organizations were selected, which one obtained by the combination of an algorithm, a similarity measure and a value of k (number of groups to be obtained). For that, two algorithm (PAM; Ward), two similarity measures (J.RI; J.RT (Section 2)) and ten values of k were considered (5 to 50, steps of 5) (40=2*2*10). An organization provides a different way to organize the extracted patterns. In fact, these forty organizations can be grouped in four sets, each one related to a combination of an algorithm and a similarity measure (2*2). Although it is necessary to set k, to obtain an organization, this value can be used to analyze the combinations of algorithms and similarity measures on different views. This was the idea used to do the analysis of the results (Section 5). Most of the experiments choices were done based on [5] work.

Before definitely executing GLM and the methods described in Section 2 (LM-M, LM-T, LM-S, LM-PU), in order to do a comparative analysis of its performance, it was necessary to find out the most suitable parameters to set GLM. For that, many experiments were executed to adjust the parameters PS, NG and MP. In each experiment, GLM was executed on all the 40 *ARC*s, in each data set. Being a genetic approach, GLM doesn't obtain the same results for the same parameters every run. Thus, each one of the experiments was executed 10 times in order to obtain an average performance. In the end, the following values were selected: $PS = 50.000$, $NG = 50.000$ and $MP = 0.75$. Regarding the value of m, the parameter was set to 5 (N in LM-PU was set to 5 too). To allow the comparative analysis, the methods LM-M, LM-T, LM-S and LM-PU were also executed on all the 40 *ARC*s, in each data set.

5 Results and Discussion

Since the GLM optimization function aims a good tradeoff between P and RF, all the results are shown and discussed over these measures (only the *ARC*s results related to the GLM selected parameters were considered). Table 1 presents the averages of P and RF in GLM and in the methods used for comparison. Each average was obtained from the results related to the presented configuration. The value $P = 0.740$ in GLM at SDS:PAM:J.RI, for example, was obtained from the average of the P values in GLM at Adult:PAM:J.RI and Income:PAM:J.RI over the ks. Thus, notice that the forty organizations, related to each data set, were grouped in four sets, considering that k can be used to analyze the combinations

[2] `http://www.borgelt.net/apriori.html` [Christian Borgelt's Web Page].

Table 1. Performance of the labeling methods, measured through P and RF, in the different data types

Data type	Alg.	Sim. M.	LM-M		LM-T		LM-S		LM-PU		GLM	
			P	RF	P	RF	P	RF	P	RF	P	RF
SDS [Adult/ Income]	PAM	J.RI	0.999▲	0.153✓	0.961	0.260	0.965	0.272	0.999▲	0.170✓	0.740✓	0.503▲
		J.RT	0.995	0.355	0.934	0.455	0.965	0.427	0.998▲	0.403✓	0.878✓	0.613▲
	Ward	J.RI	0.996▲	0.338✓	0.915	0.437	0.963	0.423	0.993	0.369	0.847✓	0.557▲
		J.RT	0.988	0.350	0.929	0.535	0.963	0.401	0.995▲	0.412✓	0.912✓	0.616▲
NSDS [Groce- ries/Sup]	PAM	J.RI	0.979	0.511	0.852	0.482	0.913	0.398	0.986▲	0.523✓	0.478✓	0.744▲
		J.RT	0.911	0.611	0.743	0.633	0.818	0.646	0.935▲	0.671✓	0.452✓	0.769▲
	Ward	J.RI	0.955	0.770	0.905✓	0.855▲	0.931	0.787	0.966▲	0.572✓	0.616	0.687
		J.RT	0.899	0.616	0.773	0.690	0.832	0.672	0.929▲	0.645✓	0.698✓	0.704▲

of algorithms and similarity measures on different views (Section 4). Therefore, each presented configuration represents the average of twenty results (10 related to each data set of the same data type).

A comparative analysis was done to evaluate the performance of GLM, in relation to the other methods usually used, based on the average of each measure (P; RF), considering the different data types, apart from the data set used. For that, in Table 1, the highest averages, regarding each one of the measures (P; RF), are marked with ▲ in each considered configuration. For the SDS:PAM:J.RI configuration, for example, the best average for RF is the one related to GLM (0.503). In the table, for each ▲, there exist a ✓ on the other measure of the pair P/RF to indicate the method is, in theory, suitable. The measure marked with ▲, in the ▲/✓ pair, indicates the one that leads to the selection of the method – RF in GLM (0.503), for example. Thereby, it is possible to observe, in each considered configuration, the method that presents the best performance. Finally, since the results related to LM-M, LM-T, LM-S and LM-PU are deterministic and the differences among the 10 GLM executions were too small, no statistical test was done. It can be noticed that:

Configurations related to SDS. In all the SDS configurations, the method that presents the best result in RF is GLM and in P LM-M (SDS:PAM:J.RI; SDS:Ward:J.RI) and/or LM-PU (SDS:PAM:J.RI; SDS:PAM:J.RT; SDS:Ward:-J.RT). However, it can be observed, in the selected methods (▲/✓ pairs), that P presents good results, different from RF in LM-M and/or LM-PU, where lower values are obtained. Therefore, GLM is a suitable method to be used when seeking for a balance between P and RF, since it improves RF while maintains P.

Configurations related to NSDS. In most of the NSDS configurations, the method that presents the best result in P is LM-PU and in RF GLM (NSDS:PAM:J.RI; NSDS:PAM:J.RT; NSDS:Ward:J.RT) (exception to NSDS:-Ward:J.RI with the selection of LM-T for RF). However, it can be observed, in the selected methods (▲/✓ pairs), that P presents better results in LM-PU with a reasonable RF compared to GLM P and RF. Therefore, LM-PU is a suitable method to be used when seeking for a balance between P and RF, since it presents a good P while maintains a reasonable RF. In relation to NSDS:Ward:J.RI, while P presents good results both in LM-PU and LM-T, the

same doesn't occur in LM-PU RF, where a lower value is obtained. Therefore, in this last case, LM-T is the one to be chosen.

Based on the obtained results, it can be noticed that while the method that seems to be more suitable for SDS regarding association rule clustering is the proposed one, i.e., GLM, for NSDS, in almost all the cases, is LM-PU, although GLM presented reasonable results. Thus, GLM seems to be useful in some circumstances and good and useful if a tradeoff between P and RF is essential (it can be seen, from Table 1, that GLM presents good results in almost all the considered configurations).

6 Conclusions

This paper proposed a labeling method for association rule clustering, named GLM, based on a genetic algorithm approach. This is an essential issue, since good labels must be assigned to the groups in order to guide the user during the exploration process. GLM was modeled to balance the values of P and RF, two measures that are used to evaluate labeling methods in this context. In the experiments, GLM presented a good performance and better results than some other methods already explored in the literature, mainly when applied in SDS. As future works, other genetic operators can be tested, as other ways to iterate the population during the process, aiming to refine GLM performance.

Acknowledgments. We wish to thank Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (process number 2010/07879-0) for the financial support.

References

1. Reynolds, A.P., Richards, G., de la Iglesia, B., Rayward-Smith, V.J.: Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. Journal of Mathematical Modelling and Algorithms 5(4), 475–504 (2006)
2. Jorge, A.: Hierarchical clustering for thematic browsing and summarization of large sets of association rules. In: 4th SIAM International Conference on Data Mining, pp. 178–187 (2004)
3. Sahar, S.: Exploring interestingness through clustering: A framework. In: IEEE International Conference on Data Mining, pp. 677–680 (2002)
4. Toivonen, H., Klemettinen, M., Ronkainen, P., Hätönen, K., Mannila, H.: Pruning and grouping discovered association rules. In: Workshop Notes of the ECML Workshop on Statistics, Machine Learning, and Knowledge Discovery in Databases, pp. 47–52 (1995)
5. Carvalho, V.O., Biondi, D.S., Santos, F.F., Rezende, S.O.: Labeling methods for association rule clustering. In: 14th International Conference on Enterprise Information Systems, pp. 105–111 (2012)
6. Sivanandam, S.N., Deepa, S.N.: Introduction to genetic algorithms (2008)