Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 61 (2015) 388 – 394

Procedia Computer Science

Complex Adaptive Systems, Publication 5
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2015-San Jose, CA

Fuzzy Inference Algorithm Based on Quantitative Association Rules

Ling Wang[a,b]*, Ji-Yuan DONG [a,b], Shu-Lin Li

[a]School of Automation & Electrical Engineering, University of Science and Technology Beijing, 100083, China
[b]Key Laboratory of Advanced Control of Iron and Steel Process (Ministry of Education), Beijing 100083, China

Abstract

In order to develop a data mining system to extract the fuzzy inference rules from the data, in this paper a fuzzy inference algorithm based on quantitative association rule (FI-QAR) is proposed. First, a discretization algorithm based on an improved clustering for each dimension data is adopted, and then the quantitative results are represented in the form of a Nominal variables matrix to compute the support and confidence level in the Apriori algorithm for quantitative association rules mining. On the basis of this, the quantitative association fuzzy rules are reconstructed by combing with TS fuzzy model to realize fuzzy inference, which can be applied to predict the output class and precise output. Experiment results demonstrated that the proposed algorithm is feasible and practical.

Keywords: quantitative association rules; apriori algorithm; discretization; TS fuzzy rules;

1. Introduction

Recently, Data Mining to extract fuzzy rules from the huge dataset has attracted a lot of attention to be applied into various fields. Association rules are one of the major data mining techniques. They are used to discover multiple independent itemsets that co-occur frequently and to discover rules that relate to the co-occurred itemsets in a given dataset. Apriori is a classical algorithm[1, 2] for association rule mining, which generates frequent itemsets that meets the predefined support threshold value and generates association rules that have confidence value more than the predefined confidence value. However, it is impossible to directly employ the classical Apriori algorithm to handle the numerical attribute. Moreover, classical Apriori algorithm needs to scan database many times [3]. This extensive scan makes the system consume more time for rule generation. Furthermore, due to the structure of association rule, it is difficult to realize fuzzy inference to predict the accurate value. But the well-known T-S fuzzy model[4] is a popular adopted fuzzy inference model because its consequent part is functional type and it is described by fuzzy(IF-

* Corresponding author, *E-mail address:* lingwang@ustb.edu.cn

doi:10.1016/j.procs.2015.09.166

THEN) rules to inference local input-output relations of a nonlinear system and achieve the actual prediction value. To overcome all above shortcomings, we resort to discretization to deal with the numerical attribute, and succeed mining quantitative association rules based on the improved Apriori algorithm. Finally, the optimized fuzzy rules are discovered according to the quantitative association rules and TS fuzzy inference model. In this way, the semantic integrity between data and the inference of rules are guaranteed.

There are five sections in this paper. In section 2, mining quantitative association rules is explained in detail. The fuzzy rule inference algorithm is proposed in section 3. In section 4, the experiments are presented to verify the effectiveness and the practicability of the proposed algorithm. Finally, section 5 concludes the paper.

2. Quantitative Association Rules Discovery

2.1. Discretization based on clustering algorithm

In order to deal with the numerical attributes in mining association rules, the existing approach that partitioned the numerical attributes into several intervals[5, 6] can't properly deal with the sparse distribution data. The clustering algorithm[7, 8] is effective to be exploited for discretization in the process of mining quantitative association rules. However, it is difficult to formulate appropriate clusters for rule extraction in terms of cluster size and shape. Because the density clustering algorithm does not require one to predefine the number of clusters in the data and can find arbitrarily shaped clusters, we adopt it to deal with the numerical attributes. Nevertheless, since the density clustering algorithm is sensitive to neighborhood parameters, an improved clustering method is adopted in this paper, based on the DFAC (density based fuzzy adaptive clustering algorithm) proposed by Ling Wang [9], to deal with the numerical attributes. Without predefined clustering number and neighborhood parameters, this algorithm adaptively determines the radius of neighborhood to obtain the density of each sample and increases cluster centers based on the density. The major concern in DFAC algorithm is to partition every attribute value to form the discrete intervals until it meets the defined condition. The fuzzy clustering validity index is adopted to assure the correct number of clusters in the DFAC algorithm. The experimental results show that it can provide better clustering performances over other methods. The major concern in DFAC algorithm is to partition every attribute value to form the discrete intervals until it meets the defined condition.

Based on this clustering algorithm, the discretization in the process of mining quantitative association rules is solved. The numerical attributes are partitioned into discrete intervals (without overlapping), which is mapped to sequential integer according to the interval value order. In other words, the numerical attribute is expressed as the sequential integer.

Let D is a 2-dimension database. An object in D can be written as $\{x_1, x_2\}$. Every attribute is transformed to quantized item as $\{c_{i1}, c_{i2}, \cdots, c_{i,f(i)}\}$, where $c_{i,f(i)}$ denotes the discrete interval of x_i and $f(i)$ denotes the sequential integer of discrete interval of x_i based on clustering algorithm. The numerical value is mapped to the sequential integer.

Example I: Let $\{x_1, x_2\}$ is a 2-dimension object, x_1 and x_2 can be transformed to quantized item as $\{c_{11}, c_{12}\}$ and $\{c_{21}, c_{22}, c_{23}\}$, x_1 has 2 discrete intervals and x_2 has 3 discrete intervals respectively. Suppose the object is $\{4,6\}$, if 4 belongs to the 1-th discrete interval for x_1 and 6 belongs to the 2-th discrete interval for x_2, then the sequential integer item is represented as (1, 2).

2.2. Extraction of Quantitative Association Rules

After discretization, Apriori algorithm can be used to extract the quantitative association rules from the numerical database. But Apriori algorithm had to scan the huge database many times to find the frequent itemsets, more and more time had to spend on the operation. To solve this problem, the nominal vector matrix is proposed to modify the Apriori algorithm and improve its performance. Let the binary variable denotes the cluster state value to the quantized attribute value by the cluster algorithm. So, the nominal vector can be encoded by using these asymmetric binary variables. In the nominal vector matrix, the number of rows is equal to the number of objects, and the number of columns is equal to the sum of the clusters for every attribute.

Example II: Let $\{x_1, x_2\}$ is a 2-dimension object, while corresponding to x_1 and x_2 , the number of discrete interval is 3 and 2 respectively. Let the sequential integer item is (1, 2) for the object $\{x_1, x_2\}$, which can be represented as (10001) in the form of nominal vector. When we apply the Apriori algorithm to the actual huge problems, only the nominal vector matrix needs to be used to calculate the support and confidence. Let X_B denotes the nominal vector matrix, q denotes the nominal vector for the frequent 2-itemsets L_2 . The support of L_2 is calculated by multiplying X_B and q to count for"2". Meanwhile, in order to store the support values for all frequent itemsets, a storage vector is used to record the corresponding support value for every frequent itemsets, in which the i -th data represent the support value of corresponding itemset which can be generated by converting i into binary number . For example, the nominal vector is (10001), which can be converted to a decimal number 17, then the 17-th bit of the storage vector recorded the support value of some frequent 2-itemsets L_2 .

2.3. The algorithm

In this section, we describe our algorithm for mining quantitative association rules. The proposed algorithm is divided into three phases, where phase I is a discretization process based on the DFAC clustering algorithm, phase II is mining quantitative association rules with the improved Apriori algorithm , and phase III is pruning the association rules.

Phase I Discretization process based on the DFAC clustering algorithm

Step1. Data are discreted with an improved clustering method.

Step2. The numerical attributes are transformed into discrete intervals, which are mapped to sequential integers. In other words, the sequential integer is the cluster state value which can be represented as the binary variable.

Phase II Mining quantitative association rules with the improved Apriori algorithm

Step3. After discretization, the numerical data matrix is transformed into the nominal variables matrix of which the nominal vector is encoded using the asymmetric binary variables.

Step4. A row vector is generated by joining the elements of each row of the nominal vector matrix. Each element in the row vector represents the support value for the corresponding itemset which can be generated by converting the sequential number of the element into binary number.

Step5. The frequent 1-itemsets L_1 which meets the predefined support threshold value is generated first. Then, a storage vector is used to record the corresponding support value.

Step6. Set $k = 1$, depending on the 1-itemsets, it generates the 2-itemsets L_2 .

Step7. Set $k = k + 1$, in join step, the union of two frequent $k - 1$ itemsets is J_k , which have first $k - 2$ elements in common are taken. In prune step, all the candidate itemsets of size $k - 1$ in J_k are checked whether they are frequent itemsets by using the nominal vector matrix. If the frequent itemset L_{k-1} is empty, then go to Step8. Otherwise, go to step7.

Step8. The association rules ($A \Rightarrow B$) are decided by calculating the confidence of the rule with the extracted frequent itemsets.

Phase III Pruning Rules

Step9. Calculate the confidence of the rules by the support of frequent itemsets recorded in the storage vector without rescanning the database. Retain the association rules which meet the user predefined confidence value.

Step10. If two rules have same consequents but the antecedents exist inclusion relationship, we will retain the rule with higher confidence.

Step11. All the association rules are ranked as their importance[10] by

$$Importance(A, B) = \log\left(\frac{P(A, B)}{P(A)P(B)}\right) \tag{1}$$

The rule that has the importance value less than 0.5 is excluded.

3. Fuzzy Rules Inference Learning

In order to improve the prediction accuracy of fuzzy models, we reconstruct the fuzzy rules by combining the TS fuzzy model with the extracted quantitative association rules. So, a fuzzy inference algorithm based on quantitative association rules (FI-QAR) is proposed. All the association rules are grouped according to the discrete output interval. In the same output interval, the antecedent variables corresponding to the rule that has the highest importance is determined as the antecedent variables of fuzzy rules. Besides, all data within the output interval are used to fit the consequent function of the fuzzy rules. So, the fuzzy rules is described as

$$\begin{aligned} &R_l : \textit{If}\ x_{1\min}^{(i_1)} < x_1 < x_{1\max}^{(i_1)}, \ldots, x_{p\min}^{(i_p)} < x_p < x_{pax}^{(i_p)} \\ &\textit{Then}\ \ y_{\min}^{(m)} < y^{(l)} < y_{\max}^{(m)}, y^{(l)} = a_0^{(m)} + \sum_{p=1}^{n} a_p^{(m)} x_p \end{aligned} \tag{2}$$

Here, R_l is the l -th rule in the rule base, $x = [x_1, \cdots, x_p]^T$ is the input variable, $x_{p\min}^{(i_p)}$ and $x_{p\max}^{(i_p)}$ denote the minimum and maximum of i_p-th discrete interval of input x_p respectively, $y_{\min}^{(m)}$ and $y_{\max}^{(m)}$ denote the minimum and maximum of m-th discrete interval of output variable y , the consequent parameters for each individual rule are obtained by weighted least square learning.

For the input x , the degree of activation of the l -th rule is calculate by

$$\tau_l = \prod_{p=1}^{n} \mu_p^{(l)} \tag{3}$$

Where $\mu_p^{(l)}$ represents the membership of x_p in the l -th rule, as (4):

$$\mu_p^{(l)} = e^{-\frac{\left(x_p - \bar{x}_p^{(l)}\right)^2}{2(\sigma_p^{(l)})^2}} \tag{4}$$

Where $\bar{x}_p^{(l)}$ and $\sigma_p^{(l)}$ represent the mean and the standard deviation of x_p belong to the l -th rule respectively, which are generated from the clustering results of various attributes. The output y of the model is a weighted sum of rule contributions:

$$y = \frac{\sum_{l=1}^{R} \tau_l y^{(l)}}{\sum_{l=1}^{R} \tau_l}, \quad l = 1, \cdots, R \tag{5}$$

4. Experiments and results analysis

4.1. UCI benchmark datasets

To verify the proposed FI-QAR algorithm, experiments are done on various benchmark datasets (Iris, Wine, Breast Cancer and Seed datasets), which are taken from the UCI repository of machine learning databases [11].

For Iris, as shown in Table 1, 4 numerical attributes are transformed into different discrete intervals, which are mapped to sequential integers as the categories. For example, the first attribute was divided into 12 intervals by discretization, and the maximum and minimum values for each interval are listed in Table 1. Then the intervals are mapped to sequential integers from 1 to 12 as the categories.

Table 1. The discrete results of each dimension of Iris dataset

Attribute	Minimum	Maximum	Width	Category
1	4.3	4.5	0.070711	1
	4.6	4.7	0.05164	2
	……	……	……	……
	7.6	7.9	0.098319	12
2	2	2.7	0.186982	1
	2.8	3.1	0.104878	2
	3.2	3.5	0.112706	3
	3.6	4.4	0.214811	4
3	1	1.4	0.10853	1
	1.5	1.9	0.116697	2
	……	……	……	……
	6	6.9	0.323616	6
4	0.1	0.3	0.056955	1
	……	……	……	……
	2.1	2.5	0.135571	5

On the same benchmark datasets, we evaluate our approach (FI-QAR) about the average running time (in seconds) compared with the classical Apriori algorithm, and the result is illustrated in Figure 1. For instance, the running time of FI-QAR on Iris, Wine and Seed dataset are 0.74s, 0.31s and 0.13s respectively, while the running time of Apriori on these dataset are 1.60s, 1.03s and 3.45s respectively. It is obvious that FI-QAR algorithm decreases the time complexity degree of rule mining, even the running time of the Breast Cancer dataset is longer than that of other dataset but which is still enough shorter comparing with the Apriori algorithm. It can be seen that FI-QAR algorithm greatly improved the efficiency by using the nominal vector matrix to calculate support values of itemsets that recorded in a storage vector without rescanning the database.

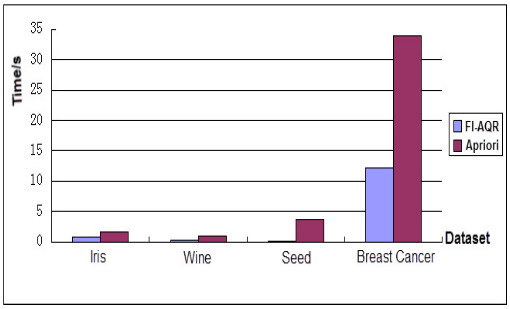

Fig 1. Efficiency of the FI-QAR algorithm.

Combining the quantitative association rules with the TS fuzzy mode, 18 fuzzy rules are extracted from the training data according to the fuzzy inference learning algorithm, where a fuzzy rule is described as:

$$\begin{aligned}&\mathit{If}\;\; 2 \le x_2 \le 2.7, \mathit{and}\; 3 \le x_3 \le 4.3, \mathit{and}\; 1 \le x_4 \le 1.5,\\ &\mathit{Then\ class\ is}\ 2, \mathit{class} = 0.0151 - 0.0900 \times x_2 + 0.1694 \times x_3 + 0.4979 \times x_4\end{aligned} \tag{6}$$

The output class of Iris dataset is known which eliminates the need for discretization. If we need to predict output by using fuzzy inference, the consequent function of fuzzy rules will be adopted. The accuracies for the benchmark datasets with the FI-QAR algorithm are listed in Table 2. It is evident that the prediction accuracy of FI-QAR algorithm is higher than 86.0% for all datasets, especially, which is higher than 90% for Iris and Wine.

Table 2. The correct rate of prediction for each dataset.

dataset	Iris	Wine	Seed	Breast Cancer
The number of the test sample	50	50	50	50
The number of the samples corresponding to correct class	47	45	44	43
Predict accuracy	94%	90%	88%	86%

4.2. Box-Jenkins gas furnace data

The proposed FI-QAR algorithm is applied for online identification using the gas furnace data of Box and Jenkins [12]. The dataset used consists of 296 samples, where $x(t)$ denotes the gas flow at time t, $y(t)$ denotes the concentration of CO_2. The inputs are $\{x(t), x(t-1), x(t-2), y(t-1), y(t-2), y(t-3)\}$, and the output is $y(t)$.

The fitting curve of the concentration of CO_2 with the FI-QAR algorithm is given in Figure 2. From Figure 2, one can observe that FI-QAR algorithm is responsive to changes in the underlying characteristics of the nonlinear system in a timely fashion, where the blue curve represents the actual output value and the red line represents the predictive output value. That is, the FI-QAR algorithm tracks the system characteristics well. Table 4 shows the comparative results of the proposed FI-QAR against other fuzzy models that have been previously applied to the Box-Jenkins gas furnace problem. It can be observed that FI-QAR has better modeling performance than other models, which reports a lower RMSE value while using the same number of rules. Here, The RMSE is

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{7}$$

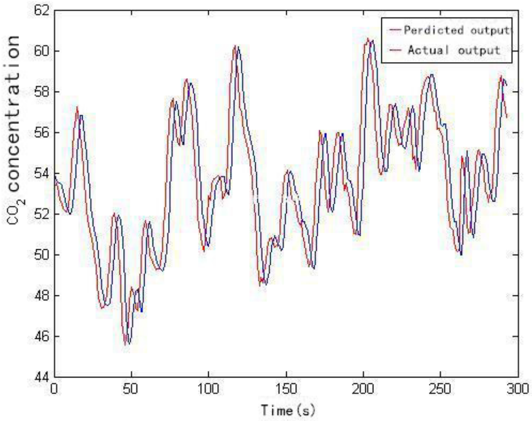

Figure 2. Comparing with the predicted value and actual value of fuzzy rules.

Where n represents the number of data samples, $\hat{y}_i$ and y_i represent predictive output and actual output respectively. With the fuzzy inference, FI-QAR enhances interpretability of its variable relation and the output is easy to comprehend. It can be observed that FI-QAR is able to achieve both improved interpretability and accuracy for the Box and Jenkins data.

Table 3. Comparison of prediction performance with different models

Model name	Input	RMSE
Box and Jenkins[12]	6	0.2020
A . Habbi [13]	2	0.1570

Hao Li [14]	6	0.0620
GA[15]	6	0.1474
VABC-FCM[16]	2	0.1256
FI-QAR	6	0.0560

5. Conclusion

The FI-QAR algorithm is proposed in this paper, which combines TS fuzzy modelling approach and the quantitative association rules mining approach coupled with discretization based on clustering approach to achieve improved interpretability-accuracy representation for linguistic fuzzy modelling. The FI-QAR model employs the clustering approach for discretization in the process of mining quantitative association rules, thus allowing it to solve the real world problems with quantitative characteristics. Besides, in order to improve the efficiency, quantitative association rules based on the improved Apriori algorithm are mined and succeed optimizing the fuzzy rules combined with the TS fuzzy inference model. The learning robustness and modelling versatility of FI-QAR are demonstrated through UCI database and Box-Jenkin gas furnace datasets, and the results are encouraging.

6. Acknowledgements

This paper is supported by the Fundamental Research Funds for the Central Universities (No.FRF-SD-12-009B) and the State Scholarship Fund.

References

1. Agrawal R. Fast algorithms for mining association rules[C]// Proc. of 20th Intl. Conf. on VLDB, 1994:487-499.
2. Li N, Zeng L, He Q, et al. Parallel Implementation of Apriori Algorithm Based on MapReduce[C]// 2012 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed ComputingIEEE Computer Society, 2012:236-241.
3. Varde A S, Takahashi M, Rundensteiner E A, et al. Apriori algorithm and game-of-life for predictive analysis in materials science[J]. International Journal of Knowledge-based and Intelligent Engineering Systems, 2004, 8(4):228.
4. Takagi T, Sugeno M. Fuzzy identification of systems and its applications to modeling and control[J]. Systems, Man and Cybernetics, IEEE Transactions on, 1985 (1): 116-132.
5. Srikant R, Agrawal R. Mining generalized association rules[J]. Future Generation Computer Systems, 1995, 13(12):407-419.
6. Srikant R, Agrawal R. Mining Quantitative Association Rules in Large Relational Tables[J]. Acm Sigmod Record, 1996, 25:1-12.
7. Peng Yan, Guoqing Chen, Fuzzy Quantitative Association Rules and Its Applications, Fuzzy Applications in Industrial Engineering, vol. 201, 2006, pp 573-587.
8. Hannes Verlinde, Martine De Cock, and Raymond Boute, Fuzzy Versus Quantitative Association Rules: A Fair Data-Driven Comparison, IEEE Transactions on system, Man, and Cybernetics-Part B: Cybernetics, vol. 36, no. 3, 2007, pp 679-684.
9. Wang Ling, Wu Lu Lu, Fu Dong MEi, A density-based fuzzy adaptive clustering algorithm, Journal of University of Science and Technology Beijing, 2014,36(11):1560-1565
10. Chih-Fong Tsai, Mao-Yuan Chen. Variable Selection By Association Rules For Customer Churn Prediction Of Multimedia On Demand[J]. Expert Systems with Applications, 2010, 37(3):2006-2015.
11. C. L. Blake and C. J. Merz, UCI Repository of machine learning databases, http://archive.ics.uci.edu/ml/datasets/Glass+Identification. Irvine, CA: University of California, Department of Information and Computer Science, 1998.
12. M .Willmann, M. Marschal, F. Hölzl , K . Schröppel , IB. Autenrieth , S. Peter . Time series analysis as a tool to predict the impact of antimicrobial restriction in antibiotic stewardship programs using the example of multi-drug-resistant Pseudomonas aeruginosa. Antimicrob Agents Chemother,2013,57(4):1797-1803.
13. A. Habbi, Y. Boudouaoui. Hybrid Artificial Bee Colony and Least Squares Method for Rule-Based Systems Learning. International Scholarly and Scientific Re-search & Innovation,2014,8(12):1938-1941.
14. Hao Li, Xiaohong HUANG, Zhiwei SHI. A Hybrid Approach Based on QPSEA and RLS for Fuzzy Modeling. Journal of Computational Information Systems, 2014,10(19):8343-8353
15. Zhao L, Qian F, Yang Y, et al. Automatically extracting T-S fuzzy models using cooperative random learning particle swarm optimization[J]. Applied Soft Computing, 2010, 10(3):938-944.
16. Su Z, Wang P, Shen J, et al. Convenient T-S fuzzy model with enhanced performance using a novel swarm intelligent fuzzy clustering technique[J]. Journal of Process Control, 2012, 22(1): 108-124.