Fuzzy Clustering Algorithm Based on Tree for Association Rules

Dechang Pi [1], Xiaolin Qin[2] , and Qiang Wang[3]

College of Information Science and Technology,
Nanjing University of Aeronautics and Astronautics,
Yudao Street 29, Nanjing, Jiangsu, 210016, China
[1] dechang_pi@hotmail.com
[2] qinxcs@nuaa.edu.cn
[3] nuaacs@126.com

Abstract

It is one of the problems in association rules mining that a great many of rules generated from the dataset makes it difficult to analyze and use. An algorithm named FCABTAR for association rules clustering is proposed and applied to association rules managing. Firstly, an example is presented to demonstrate the weakness by the distance clustering. Secondly, the definition of fuzzy simulation degree and simulated matrix for association rules are put forward. Thirdly, a new algorithm based on a dynamic tree is brought forward, which can be used to implement the fuzzy clustering. Experiment with the UCI dataset shows that this algorithm can efficiently cluster the association rules for a user to understand.

Keywords: Fuzzy clustering analysis; Association rule; Data mining; Knowledge management

1. Introduction

Association rule is a focus in the data mining area recently; meanwhile, it is a way to express the knowledge.

Now fuzzy set theory has been applied to many fields including data mining. Fuzzy clustering method is more precise in dealing with data simulation, and the results are easier to be understood and used. Therefore, research into fuzzy clustering method for knowledge is significant not only to theory, but also to application. Many fuzzy clustering methods have already been proposed and used in data mining. Fuzzy clustering with squared Minkowski distances was proposed in [8]. Fuzzy c-means method was used in dynamic data mining [4]. Clustering algorithm based on self-similarity of the dataset was suggested in [3]. Fosca Giannotti et al. cluster transactional data with the standard definition of mathematical distance used in the KMeans algorithm to represent dissimilarity among transactions. Shoji Hirano et al. [13] cluster a numerical dataset using relative proximity. Sang Hyun Oh and Won Suk Lee [12] use clustering method to detect any anomalous behavior in the audit data. Literature [7] gives an overview of cluster analysis techniques from a data-mining point of view. Du [15] presents a method to select some "representative" rules but not a large number of rules for a user to understand. Clustering rules is necessary in merging multi-

sources, which may contain conflicting knowledge and make us illusive currently [2]. From the literatures, we have not found any methods to cluster association rules based on fuzzy simulation for a user to expediently understand.

The rest of the paper is organized as follows. In Section 2,we provide a brief description of association rule and the classical mining algorithm Apriori, and the relative researches on fuzzy clustering. In Section 3, we propose a new algorithm, which we call FCABTAR, and give an example of FCABTAR. In Section 4, we discuss the design of our experiment and the results returned; finally, in Section 5, we present our conclusion and the further work.

2. Background

Association rules are used to discover the relationships, and potential associations, of items or attributes among huge amounts of data. These rules can be effective in uncovering the unknown relationships, providing results that can be the basis of forecast and decision. They have proven to be very useful. The application and development about association rules is a popular area of data mining research [17].

2.1 Review of association rules and the classical mining algorithm

The early data mining method for association rules is the support-confidence framework established by Agrawal et al. [10,11]. They proposed a model to discover meaningful itemsets and construct association rules for market analysis. The following is a formal statement of the problem.

Let I={i1, i2, i3,…,iN} be a set of N distinct literals, called items. In general, a set of items is called an itemset. The number of items in an itemset is the length of an itemset. Itemset of length k is referred to as a k-itemset. Let D be a set of variable length transactions, where each transaction T is a set of items such that $T \subseteq I$. Associated with each transaction is a unique identifier, which shall be referred to as its TID. |D| is the cardinality of database D. A transaction T is said to support an itemset X, where $X \subseteq I$, if it contains all items of X, i.e. $X \subseteq T$. The fraction of the transactions in D that support X is called the support of X, denoted Support(X). An itemset is large if its support is above some user-specified minimum support threshold, denoted MinSup. An association rule is an implication of the form r: $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \varnothing$. The support for rule r is defined as Support($X \cup Y$). A confidence factor defined as support($X \cup Y$)/support(X), is used to evaluate the strength of such association rules.

The process of Apriori mining association rule algorithm makes multiple passes over the database D to build candidate itemsets, and then create large itemsets. In the kth level, the algorithm finds all large k-itemsets. Denoting Lk is the set of all large k-itemsets, and Ck is the set of candidate k-itemsets by obtaining from Lk-1, that is, potentially large k-itemsets. For each transaction in D, the candidates in Ck also contained in the transaction are determined and their support is increased by 1/|D|. At the end of scanning, if their supports are greater than, or equal to, the user-specified

minimum support (MinSup), the candidate k-itemsets immediately become the large k-itemsets. Meanwhile, it will generate a large number of the candidate itemsets that need to be contrasted with the whole database level by level in the process of mining the association rules. Therefore, performance is dramatically affected, as the database is repeatedly scanned. The most time-consuming part of the algorithm is to discover large itemsets while the generation of association rules given the large itemsets is straightforward enough. Many researchers have tried to improve its efficiency from different angles, and decreased the number of database scans and the number of candidate itemsets. [9,17].

A common problem in association rule mining is that a large number of rules are generated from the datasets, which makes it difficult for users to analyze and make use of the rules. Solutions have been proposed to overcome this problem, which include constraint-based data mining, post-pruning rules, grouping rules [16], and unexpected patterns based on user's beliefs [1].

2.2 Related research on fuzzy clustering

Currently, most of the commercial clustering systems are based on the Boolean logic model. They assume that a user's requirements can precisely be characterized by the terms. However, this assumption is inappropriate due to the fact that the user's requirements may contain fuzziness. The reason for the fuzziness contained in the user's requirements is that the user may not know much about the subject he/she is clustering or may not be familiar with the clustering system. Since fuzzy set theory can be used to describe imprecise or fuzzy information, many researchers have applied the fuzzy set theory to many systems including clustering [16,14].

The objective of fuzzy clustering methods is to divide a given dataset into a set of clusters based on similarity [6]. In classical cluster analysis each datum must be assigned to exactly one cluster. Fuzzy cluster analysis relaxes this requirement by allowing membership degrees, thus offering the opportunity to deal with data that belong to more than one cluster at the same time. Most fuzzy clustering algorithms are objective function based: they determine an optimal classification by minimizing an objective function. In objective function based clustering usually each cluster is represented by a cluster prototype. This prototype consists of a cluster center and maybe some additional information about the size and the shape of the cluster. The cluster center is an instantiation of the attributes used to describe the domain, just as the data points in the dataset to divide. However, the cluster center is computed by the clustering algorithm and may or may not appear in the dataset. The size and shape parameters determine the extension of the cluster in different directions of the underlying domain.

The degrees of membership to which a given data point belongs to the different clusters are computed from the distances of the data point to the cluster centers. The closer a data point lies to the center of a cluster, the higher is its degree of membership to this cluster. Hence, the problem to divide a dataset $X=\{\vec{x}_1, \ldots, \vec{x}_n\}$ into c clusters can be stated as the task to minimize the distances of the data points to the cluster centers. An iterative algorithm is used to solve the classification problem in

objective function based clustering: since the objective function cannot be minimized directly, the cluster prototypes and the membership degrees are alternately optimized.

2.3 Defects in association rules clustering based on distance

Clustering methods based on distance, such as Euclid distance, Minkowski distance and Hamming distance often have some limitations, and they cannot effectively deal with the association rules clustering.

Usually association rules are denoted as the form: $X \Rightarrow Y$, in which X and Y are often expressed as X_1 and X_2 and … and X_n. As aforementioned, association rule is a way for knowledge expression, and there are some similarities among them. For example, the following two rules are discovered from the zoo dataset in the UCI Machine Learning Database Repository.

Rule 1: if backbone=true and breathes=true and venomous=false then fins=false

Rule 2: if backbone=true and venomous=false and fins=false then breathes=true

Obviously we can find some similarities between the two rules. From the antecedent attributes of these two rules we can cluster them in terms of similarities. The core problem is how we depict the simulation degree between these rules. Suppose that there are four rules as follows:

r1：X_1 and X_2 and X_3 and $X_5 \Rightarrow$ conclusion1

r2：X_2 and X_3 and X_4 and $X_5 \Rightarrow$ conclusion2

r3：X_1 and $X_4 \Rightarrow$ conclusion3

r4：$X_6 \Rightarrow$ conclusion4

In the premise, let X_i appearing denotes 1 and non-appearing denotes 0. Using Euclid distance, the simulation coefficient between two rules is expressed with distance d, as follows:

$d(r_1, r_2)=1.414$

$d(r_1, r_3)=2$

$d(r_1, r_4)=2.23$

$d(r_2, r_3)=2$

$d(r_2, r_4)=2.23$

$d(r_3, r_4)=1.73$

Obviously rule r_1 and rule r_2 should belong to the same class in terms of the distance clustering method, because the distance between them is minimal. After clustering, we denote this class as R_1 and its center of gravity is (0.5,1,1,0.5,1,0). In the same way, we can compute the distances between R_1 and other rules:

$d(R_1, r_3)=1.87$

$d(R_1, r_4)=2.12$

$d(r_3, r_4)=1.73$

Obviously rule r_3 and rule r_4 should be clustered to the same class because the distance between them is minimal. But considering from the attributes appeared in r_3 and r_4, we shouldn't make them belong to the same class because they do not have the same attributes. This result indicates that current clustering methods based on distance are not suitable for rules clustering. Until now, we have not found any literatures to cluster association rules for a user to understand.

Based on the research on association rules and fuzzy clustering, we bring forward the FCABTAR algorithm for association rules clustering.

3. Fuzzy Clustering Algorithm Based on Rule Simulation

Fuzzy clustering is partitioning n objects into K subsets, which are called as clustering dollop. The clustering result can be denoted by the membership degree matrix U.

$$U_{s\times n}=(u_{ik})_{s\times n}=\begin{pmatrix} u_{11}, & \ldots, & u_{1n} \\ u_{i1}, & \ldots\ u_{ik}, \ \ldots & u_{in} \\ u_{s1}, & \ldots, & u_{sn} \end{pmatrix}$$

Where $u_{ik} \in [0,\ 1]$, $1 \leqslant i \leqslant s$, $1 \leqslant k \leqslant n$

For clustering methods based on fuzzy similar relation, the first step is to construct the fuzzy simulated matrix, however the key of this step is how to define the simulation coefficient, which can reflect the similarity between two objects. At present, there are many methods that can be used to express simulation coefficient, such as maximum-minimum method. But they cannot describe the simulation between rules, and are not appropriate to rule clustering just as Euclid distance method described in section 2.3.

3.1 Some definitions

We bring forward the concept and definition of rule simulation coefficient in terms of expressions of association rules.

Definition 1: rule r is expressed in this form: X⇒Y, where X and Y denote AND relation between attributes, namely A_i and … and A_k ;

Definition 2: $\|r_u \wedge r_v\|$ is the number of attributes, which appear in the antecedents of rule r_u and r_v at the same time;

Definition 3: $\|r_u \vee r_v\|$ is the number of attributes, which appear in the premise of rule r_u and r_v ;

Definition 4: $c_{uv} = \dfrac{\left\| r_u \wedge r_v \right\|}{\left\| r_u \vee r_v \right\|}$ is the fuzzy simulation coefficient between rule r_u and rule r_v. Apparently, the bigger c_{uv} is, the more similar the rules are, and vice versa.

Definition 5: rule fuzzy simulated matrix $C = \left(c_{uv}\right)_{n\times n}$ denotes simulation degree between rules.

Matrix C expresses the simulation degrees between different rules. The c_{uv} represents the simulation degree of *u*th rule to *v*th rule and it satisfies the follow demands:

1. $c_{uv} \in [0,1]$
2. Reflexivity: $\forall u, v, C_{uv} = 1$, if $u = v$
3. Symmetry: $\forall u, v, C_{uv} = C_{vu}$, if $u \neq v$

3.2 Clustering method based on fuzzy simulation coefficient

Our algorithm FCABTAR combines rule simulation coefficient with dynamic construction tree method. Its general steps can be described as follows:

1. Use association rules to construct simulated matrix.
2. Construct a tree;
3. Prune the tree according to the user-specified threshold, and then get the clustering results.

Let $N = \left(V, \{E\}\right)$ be the storage form for the simulated matrix, where V is a set of the rules, $\{E\}$ is a set of the branches among rules. Let U be a non-empty subset of V, and TE be a set of the branches about N. Clustering algorithm based on the simulation coefficient is described as follows:

$U = \{u_0\}$, *where* $u_0 \in V$
TE = ∅;
WHILE $U \neq V$ DO
BEGIN
  Find a maximal $c_{u_0 v_0}$, where $u_0 \in U, v_0 \in V - U$
  Execute $TE = TE \cup \{(u_0, v_0)\}$ and $U = U \cup \{v_0\}$
END
FOR each branch b∈TE DO
  IF b.value < the user-specified threshold λ THEN
    Cut down this connection
  ENDIF
ENDFOR

We still use these rules r_1~r_4 given in section 2.3 as example to indicate that our method has an advantage over the distance method. The fuzzy simulation coefficients between two rules can be expressed with simulation degree sd, as follows:

sd (r1, r2)=0.6
sd (r1, r3)=0.2
sd (r1, r4)=0
sd (r2, r3)=0.2
sd (r2, r4)=0
sd (r3, r4)=0

and the simulated matrix *C* for these four rules is:

$$C = \begin{pmatrix} 1 & & & \\ 0.6 & 1 & & \\ 0.2 & 0.2 & 1 & \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The connection among these rules is described in Fig.1, and the numbers on the line are the fuzzy simulation coefficients.

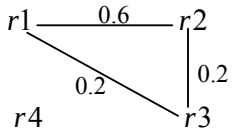

Fig. 1. Fuzzy simulation degree between rules r_1~r_4 which are described in section 2.3

If user-specified threshold be 0.5, obviously these four rules can be divided into three classes, {r1, r2}, {r3} and {r4}. If the threshold be 0.2, we can get two classes, {r1, r2, r3} and {r4}. Apparently, by employing c_{uv}, we can avoid the weakness of distance clustering, and at the same time, we can control the clustering by adjusting the threshold of c_{uv}.

4 Experiment

As C++ STL is powerful, this fuzzy clustering algorithm is implemented with C++ STL, compiled with Microsoft Visual C++ 6.0. We use Apriori algorithm to mine association rules from the zoo dataset with different support and confidence. This dataset can be obtained from UCI Machine Learning Database Repository at http://www.ics.uci.edu/~mlearn/ MLRepository.html, and it contains 101 records and 18 attributes. Table 1 shows the numbers of association rules generated from the dataset Zoo with different support and confidence.

Table 1. Numbers of association rules mined from the Zoo dataset. This dataset can be obtained from UCI Machine Learning Database Repository at http://www.ics.uci.edu/~mlearn/MLRepository.html, and it contains 101 records and 18 attributes

MinConf \ MinSup	0.5	0.6	0.7
0.9	64	19	6
0.8	120	34	10
0.7	165	50	12
0.6	203	56	12

If the thresholds for support and confidence are 0.5 and 0.9 respectively, 64 rules are found. The following four rules are examples, and obviously they should be clustered into one class.

1. if feathers=no and airborne=true and toothed=true then backbone=true and venomous=no
   support=0.544554 confidence=0.932203
2. if feathers=no and toothed=true and backbone=true then airborne=no and venomous=no
   support=0.544554 confidence=0.901639
3. if airborne=no and toothed=true then backbone=true and feathers=no and venomous=no
   support=0.544554 confidence=0.932203
4. if toothed=true then backbone=true and venomous=no and feathers=no and airborne=no
   support=0.544554 confidence=0.901639

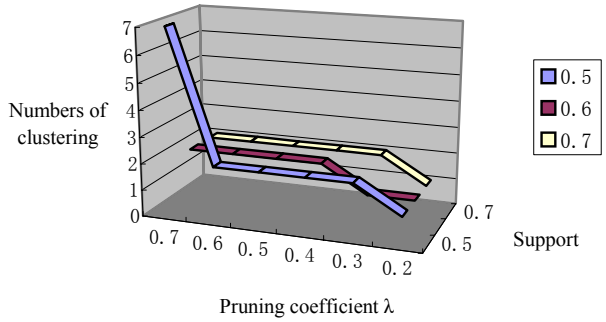

Fig. 2. Clustering the association rules discovered from the Zoo dataset

We use this algorithm to cluster the above 12 rules sets as shown in Table 1 as well as to prune the simulation tree with different pruning coefficient λ. Let the confidence threshold be 0.9, the range of λ be [0.2,0.7] and the range of support be [0.5,0.7], then we get the clustering results as shown in Fig. 2. With the pruning coefficient λ being 0.7, the numbers of the association rules clustering are 7,2 and 2. If we change λ into 0.2, the clustering number will be 1.

5 Conclusions and Further Work

Usually the number of association rules found by data mining algorithms is large. Although these rules have high supports and confidences, they are often pertained to some aspects and have higher similarity. Using this method we can categorize the association rules; hence they can be easily understood and used. The core of this paper can be summarized as follows:

1. Propose the definition c_{uv}, which denotes the fuzzy close degree between rule r_u and rule r_v;
2. Define the simulated matrix;
3. Put forward a clustering algorithm based on the simulated matrix, by which we can obtain different clustering results with different pruning coefficients.

We use Apriori algorithm to mine association rules from the Zoo dataset provided by the UCI, and employ our algorithm to cluster these found rules. Experiment shows that this algorithm is efficient in clustering the association rules.

What we need to explain is, if the premise is out of the AND expression, we must translate it into the normal form. If the premise contains the NOT operator, our algorithm cannot run smoothly. We are studying the new method to resolve this problem.

Acknowledgements

This research is Supported by the Aeronautical Science Foundation of China under Grant No. 02F52033 and the Hi-Tech Research Project of Jiangsu province under Grant No. BG2004005.

Many thanks to Professor Junhai Lin for his valuable advise. We would like to thank Mr. Wangfeng Gu for his help in implementing the algorithm presented in this paper. Gratitude is also extended to anonymous referees for their constructive comments.

References

1. Balaji Padmanabhan, Alexander Tuzhilin. *Unexpectedness as a measure of interestingness in knowledge discovery*. Decision Support Systems, 27, 1999, pp. 303–318

2. Christopher W. Zobel, Loren Paul Rees. *Automated merging of conflicting knowledge bases, using a consistent, majority-rule approach with knowledge-form maintenance*. Computers & Operations Research 32 (2005) 1809–1829
3. DANIEL BARBARA, PING CHEN. *Using Self-Similarity to Cluster Large Data Sets. Data Mining and Knowledge Discovery*, 7, 123–152, 2003
4. Fernando Crespoa, Richard Weberb. *A methodology for dynamic data mining based on fuzzy clustering*. Fuzzy Sets and Systems 150 (2005) 267–284
5. Fosca Giannotti, Cristian Gozzi, Giuseppe Manco. *Clustering Transactional Data*. PKDD, LNAI 2431, pp. 175–187, 2002.Springer-Verlag Berlin Heidelberg 2002
6. Heiko Timm, Christian Doring and Rudolf Kruse. *Different approaches to fuzzy clustering of incomplete datasets*. International Journal of Approximate Reasoning 35,2004, pp. 239–249
7. JOHANNES GRABMEIER, AND REAS RUDOLPH. *Techniques of Cluster Algorithms in Data Mining*. Data Mining and Knowledge Discovery, 6, 303–360, 2002
8. P. J.F.Groenen, K.Jajuga. *Fuzzy clustering with squared Minkowski distances*. Fuzzy Sets and Systems. 120 (2001) 227-237.
9. PI De-chang, Qin Xiao-lin and Wang Ning-sheng. *Mining association rules based on dynamical pruning*. MINI-MICRO SYSTEMS. Vol. 25, No. 10 (2004), pp. 1850-1852.
10. R. Agrawal, R. Srikant, *Fast algorithm for mining association rules in large databases*, Proceedings of 1994 International Conference on VLDB, 1994 pp. 487–499.
11. R. Agrawal, T. Imilienski, A. Swami, *Mining association rules between sets of items in large databases*, Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, DC, 1993 pp. 207–216.
12. Sang Hyun Oh, Won Suk Lee. *Optimized Clustering for Anomaly Intrusion Detection*. PAKDD 2003, LNAI 2637, pp. 576–581, 2003.Springer-Verlag Berlin Heidelberg 2003
13. Shoji Hirano, Shusaku Tsumoto. *Dealing with Relative Similarity in Clustering: An Indiscernibility Based Approach*. PAKDD 2003, LNAI 2637, pp. 513–518, 2003.Springer-Verlag Berlin Heidelberg 2003
14. T.V.Ravi, K.Chidananda Gowda. *An ISODATA clustering procedure for symbolic objects using a distributed genetic algorithm*. Pattern Recognition Letters 20(1999).659-666
15. Xiaoyong Du, Sachiko Suzuki, and Naohiro Ishii. *A Distance-Based Clustering and Selection of Association Rules on Numeric Attributes*. RSFDGrC'99, LNAI 1711, pp.423–433, 1999. Springer-Verlag Berlin Heidelberg 1999
16. Yih-Jen Horng, Shyi-Ming Chen, Yu-Chuan Chang, and Chia-Hoang Lee. *A New Method for Fuzzy Information Retrieval Based on Fuzzy Hierarchical Clustering and Fuzzy Inference Techniques*. IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL.13, NO.2, 2005 pp. 216-227.
17. Yuh-Jiuan Tsay, Jiunn-Yann Chiang. *CBAR: an efficient method for mining association rules*. Knowledge-Based Systems 18 (2005) 99–105

Dechang Pi received his MS degree in computer software and PhD degree in data mining from Nanjing University of Aeronautics and Astronautics in China, in 1997 and 2002, respectively. He is now an associate professor. His main research interests include data mining, database system, data warehouse, knowledge management and fuzzy systems. He is currently in the Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics.

Xiaolin Qin received the MS degree in computer software from Nanjing University of Aeronautics and Astronautics(NUAA) in China, in 1987. He is now a professor at Department of Computer Science and Engineering in NUAA. His main research interests include database systems, geographic information systems, and information security.