基于模糊聚类的改进的模糊关联规则挖掘算法

李 雷,崔 岩
(南京邮电大学 自动化学院,江苏 南京 210046)

摘 要:文中提出一种新的方法通过使用模糊C均值对原始数据集进行预处理操作,通过这个操作可以把定量属性值转换为二进制值,继而就会得到原始数据集的模糊版本(由模糊记录和模糊属性组成)。另外,文中又提出了一种基于模糊Apriori算法的快速提取规则的算法,这种算法是利用模糊聚类从先前得到的原始数据集的模糊版本中提取模糊频繁项集从而可以得到模糊关联规则。在文章的最后,实验结果显示了提出的新算法在处理大型数据集时在挖掘时间上要优于传统的Apriori算法。对大型数据库来说,该算法在实用性和可用性上面都有很好的发展前景。

关键词:预处理;模糊聚类;模糊C均值;模糊Apriori算法;数据挖掘
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2012)11-0018-04

An Improvement of Fuzzy Association Rules Mining Algorithm Based on Fuzzy Clustering

LI Lei, CUI Yan
(College of Automation, Nanjing University of Post and Telecommunications, Nanjing 210046, China)

Abstract: In this paper, propose a methodology by doing pre-processing the original dataset using FCM which can convert quantitative values of attributes to binary values, and then get a fuzzy version (with fuzzy records and fuzzy attributes) of the original dataset. Moreover, present a fast algorithm based on the fuzzy Apriori algorithm for rule extraction utilizing fuzzy clustering (FAFC) for extracting fuzzy frequent itemsets and fuzzy association rules from the fuzzy version of the original dataset. Eventually, experiments show that the FAFC algorithm outperforms the traditional Apriori algorithm on computing time for huge database. And for huge dataset, the algorithm presented in this paper is found to be promising in terms of practicability and availability.

Key words: pre-processing; fuzzy clustering; FCM; fuzzy Apriori algorithm; data mining

1 Summarization

Association Rules Mining (ARM), an important and hot area of KDD (knowledge discover in database), is to analyze the data in a database to discover potentially interesting association rules[1,2]. Association rules mining is promising for some actual applications in terms of marketing problems, biological knowledge extraction and so on.

An association rule is an expression of $X \rightarrow Y$, where X and Y are sets of items and termed as itemsets, and $X \cap Y = \phi$. Conventional ARM algorithms usually deal with datasets with binary values. However, in the real-life, most data is neither only binary nor only numerical, but a combination of both of them. Whereas, when conventional ARM algorithms are applied to deal with numerical attributes a serious problem called sharp boundaries will be occurred. In data mining approach, the quantitative attributes should be appropriately deal with as well as the Boolean attributes. Fuzzy set theory is used to convert quantitative values of attributes to Boolean attributes, in order to eliminate any loss of information arising due to sharp partitioning, especially at sharp boundaries. In a formal definition of a fuzzy set A, is a fuzzy subset of $U = \{x\}$, in the ordered pairs:

$A = \{(x, \mu_A(x)) \mid x \in U, \mu_A(x) \in [0,1]\}$ where the relation $\mu_A(x)$, is termed as a membership function defining the grade of membership x in A. Fuzzy sets can also be thought of as an extension of the traditional crisp sets, in which each element is either in the set or not in the set[3,4].

In this paper, discuss an improvement of algorithm

收稿日期:2012-03-23;**修回日期**:2012-06-25
基金项目:国家自然科学基金项目(61070234);江苏省高校自然科学基金项目(04KJB110097,08KJB520023);南京邮电大学攀登计划项目(NY207064)
作者简介:李 雷(1958-),男,教授,研究方向为智能信号处理、非线性分析与计算智能。

for extracting fuzzy association rules from database. Firstly, propose a methodology by doing pre-processing for the original dataset based on FCM inspired bydocument[5 ~ 7]. Secondly, propose a basic fuzzy Apriori algorithm based on fuzzy clustering(FAFC) for rule extraction. Experimental results show that get fuzzy membership μ values which can work without Apriori membership function provided by human experts and the FAFC algorithm outperforms the traditional Apriori algorithm on computing time for huge database.

2 Pre-processing for the original dataset based on FCM

Generally, used sharp partitions to convert quantitative attributes into Boolean attributes, nevertheless, by doing this, introduced a major problem called sharp boundary which may arouse a loss of information of the crisp dataset and also increase the uncertainty in the dataset. Instead of using sharp boundary intervals, fuzzy partitions can eliminate any loss of information whatever the value of any numerical attribute arising due to sharp partitioning, especially at the boundaries of partitions. By using fuzzy partitions, protected the information of the numerical attributes against losing[8].

For generating fuzzy partitions, used fuzzy c-means clustering (FCM) which is a fuzzy extension of the K-means algorithm. FCM algorithm is determined the membership of each quantitative attributes belongs to some cluster to generate fuzzy partitions. Actually, FCM generalizes K-means and the latter is a special case of the former. Thus, the method is an improvement of traditional K-means algorithm[9,10].

2.1 The idea of FCM clustering algorithm

The original dataset $X = \{x_1, x_2, \cdots, x_m\}$ is an n-dimensional in m samples, thereby the matrix of the original sample data can be expressed as the following[11,12]:

$$\boldsymbol{U} = (x_{ik})_{m\times n},\ i = 1, 2, \cdots, m;\ k = 1, 2, \cdots, n$$

The basic idea of FCM algorithm is to find a fuzzy partition matrix and the number of k cluster center $\boldsymbol{C} = \{c_1, c_2, \cdots, c_k\}$. Used Sum of the Squared Error (SSE) as the objective function of FCM clustering algorithm which is defined as:

$$SSE(C_1, C_2, \cdots, C_k) = J_m(U, C) = \sum_{j=1}^{k}\sum_{i=1}^{m}\mu_{ij}^{p} d(x_i, c_j)^2 \quad (1)$$

$$\sum_{j=1}^{k}\mu_{ij} = 1;\ 0 < \sum_{i=1}^{k}\mu_{ij} < m;\ 0 \leqslant \mu_{ij} \leqslant 1 \quad (2)$$

$\boldsymbol{U}$: Matrix of membership degree, $\boldsymbol{U} = \{u_{ij}\}$, $i = 1, 2, \cdots, m$; $j = 1, 2, \cdots k$;

$\boldsymbol{C}$: Matrix of clustering center, $\boldsymbol{C} = \{c_i\}$, $i = 1, 2, \cdots, m$;

μ_{ij}: Membership degree of the j^{th} data point in the i^{th} clustering center;

p: $p \in [1, +\infty]$ is called fuzzy weighted index which is a fuzziness parameter of fuzzified degree of the partitioning. In other words, the higher value of p, the fuzzier is the resulting partitioning.

$d(x_i, c_j)$: is a chosen distance measure (Euclidean distance is chosen in this paper) between a data point x_i and the cluster center c_j which is an indicator of the data points and corresponding cluster centers.

$$d(x_i, c_j) = \| x_i - c_j \| = [\sum_{k=1}^{k}(x_{ik} - c_{jk})^2]^{1/2} \quad (3)$$

c_j: Clustering center of the j^{th} clustering categories, $c_j = \{c_{j1}, c_{j2}, \cdots, c_{jn}\}$, $j = 1, 2, \cdots, k$.

2.2 The steps of FCM algorithm

Step 1 (Initialization): Randomly choose a initiative fuzzy pseudo-partition according to evaluating μ_{ij} of each quantitative attributes;

Step 2: repeat;

Step 3 (Assignment): Calculated for each clustering center of the clustering categories by using the initiative fuzzy pseudo-partition (4).

$$c_j = \frac{\sum_{i=1}^{m}\mu_{ij}^{p} x_i}{\sum_{i=1}^{m}\mu_{ij}^{p}} \quad (4)$$

Step 4 (Updating): Recalculated the fuzzy pseudo-partition according to (5).

$$\mu_{ij} = \left(\frac{1}{d(x_i, c_j)}\right)^{\frac{2}{p-1}} \Big/ \sum_{q=1}^{k}\left(\frac{1}{d(x_i, c_q)}\right)^{\frac{2}{p-1}} \quad (5)$$

Step 5 (Iteration): The iteration continues according to Step 2 ~ Step 4, until $| J_m^l - J_m^{l-1} | < \varepsilon_1$ or $| \mu_{ij}^l - \mu_{ij}^{l-1} | < \varepsilon_2 (\varepsilon_1, \varepsilon_2 > 0)$, ξ_1, ξ_2 are presupposed threshold. In other words, until no changing in the clustering center of the clustering categories.

The goal is to minimize the objective function (1), thereby fuzzy partitioning is generated through an iterative optimization of the objective function (1), with the update of membership μ_{ij} (5) and the cluster centers c_j (4).

FCM clustering algorithm is used to transform the original dataset to the corresponding fuzzy dataset, and here

C represents the number of fuzzy partitions with quantitative attribute. In the following example, use $C=3$ (High, Middle, Low) for the FCM clustering process. The initiative numeric were shown in Table 1, and then the fuzzy version of the original dataset after clustering also shown in Table 2.

Table 1 Initiative numeric table of the original dataset

ID	Edu-level	Skill-level	Income
1	11	2	2500
2	15	4	5000
3	18	3	6500
4	8	7	3500

Table 2 Fuzzified table of the original dataset (fuzzy version)

ID	Edu-level	Skill-level	Income
1	(H:0.3;M:0.4;L:0.5)	(H:0.1;M:0.3;L:0.6)	(H:0.1;M:0.2;L:0.7)
2	(H:0.4;M:0.6;L:0.3)	(H:0.4;M:0.5;L:0.3)	(H:0.6;M:0.3;L:0.2)
3	(H:0.7;M:0.2;L:0.2)	(H:0.3;M:0.6;L:0.4)	(H:0.8;M:0.2;L:0.1)
4	(H:0.2;M:0.3;L:0.7)	(H:0.8;M:0.2;L:0.1)	(H:0.2;M:0.7;L:0.3)

3 Improvement of Apriori Algorithm Based on Fuzzy Clustering

3.1 The Apriori Algorithm

The Apriori algorithm is one of the most influential and representative technology for mining Boolean association rules which has been brought in 1993 by Agrawal. The principle of Apriori algorithm is to find the valuable association rules whose support and confidence must satisfy the user predefined minimum support and confidence. The basic idea of Apriori algorithm is to identify all the frequent itemsets whose support is not less than predefined minimum support at least.

The Apriori algorithm has an important property to improve the efficiency of the level-wise generation of frequent itemsets.

Apriori property: All nonempty subsets of a frequent itemset must also be frequent.

Used this property in the two key steps of the algorithm: connecting step and pruning step.

· Connecting step: To find L_k, a set of candidate k-itemsets is generated by L_{k-1} and its connecting. This set of candidates is denoted C_k.

· Pruning step: C_k is a superset of L_k whose members may or may not be frequent, but all of the frequent k-itemsets are included in C_k. A scan of the database to determine the count of each candidate in C_k would result in the determination of L_k (all frequent candidates having a count no less than the minimum support count, and therefore belong to L_k). However, C_k can be large, and so this could involve heavy computation. For compression of C_k, the Apriori property is used: any non-frequent $(k-1)$-itemset cannot be subsets of frequent k-itemset. Consequently, if any $(k-1)$-subset of a C_k is not in L_{k-1}, then the candidate cannot be frequent either and so can be deleted from C_k. This test of subsets can be done quickly by maintaining a hash tree of all frequent itemsets.

But traditional Apriori algorithm has two shortcomings that affect the efficiency of the algorithm. One is too many scans of the transaction database when seeking frequent itemsets that requires lots of I/O load and computing time. The other is large amount of useless candidate itemsets generated and need to repeat to scan database. For the shortcomings of the above Apriori algorithm, an improvement of Apriori algorithm based on fuzzy clustering (FAFC) is presented in this paper. For large database, this approach is found to be promising in terms of computational time and availability. Moreover, each membership function μ can be constructed manually by an expert in traditional fuzzy Apriori algorithm. But using an expert-driven approach is very cumbersome and not feasible in real-life datasets. Thus, it is humanly impossible for an expert to create fuzzy partitions for each attribute. Instead, fuzzy clustering can be used to automate the creation of the fuzzy partitions.

3.2 The Improved Apriori Algorithm

The pseudo code of improved Apriori algorithm as follow:

Input: numerical database D, minimum support threshold (min_sup), minimum confidence (min_conf);

Output: fuzzy frequent itemsets(FL), fuzzy association rules (Rule);

(1) for each attribute A_i

(2) U_i = cluster(A_i)

(3) $U=\cup_i U_i$

(4) FL_1 = Get_frequent_1-Itemset(U);

(5) for($k=2; FL_{k-1}\neq\emptyset; k++$){

(6) FC_k = fuzzy_Ap-gen(FL_{k-1});

(7) for each fuzzy transaction $X\in FC_k$ do{

(8) compute sup(X);

(9) $FL_k=\{X\in FC_k \mid \text{sup}(X)>=\text{min\_sup}\}$; }

(10) fuzzy_Ap-genrules($FL_k R_m$, min_conf); }

(11) return $FL = \cup_k FL_k$, R_m;

procedurefuzzy_Ap-gen :

(1) for each fuzzy itemset $l_1 \in FL_{k-1}$

(2) for each fuzzy itemset $l_2 \in FL_{k-1}$

(3) if $((l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge \cdots (l_1[k-2] = l_2[k-2]) \wedge (l_1[k-1] < l_2[k-1]))$

(4) then { $X = l_1 \infty l_2$

(5) for each ($k-1$)-subset c of X

(6) if $c \notin FL_{k-1}$ then delete X

(7) else add X to FC_{k+1};

(8) return FC_k;

procedurefuzzy_Ap-genrules (FL_k, R_m, min_conf) :

(1) $k = | FL_k |$; $m = | R_m |$

(2) if $k > m+1$ then

(3) R_{m+1} = fuzzy_Ap-gen(R_m);

(4) for each subsets $r_{m+1} \in R_{m+1}$ do{

(5) If ($\text{conf} = \dfrac{\sup(FL_k)}{\sup(FL_k - r_{m+1})} \geqslant \text{min\_conf}$)

(6) Rule: ($FL_k - r_{m+1}$) $\rightarrow r_{m+1}$

(7) else delete r_{m+1} }

4 Experiment Conclusion

All the experiments are performed on a 2.2GHz Intel (R) Core(TM) 2 Duo PC with 2048MB memory, running on the Windows XP ,program language is C.

In order to describe FCM how to generate fuzzy partitions, have used the FAM95 dataset. Use numeric attributeIncome to illustrate the processing of FCM and creation of fuzzy partitions. For the attribute Income, use $C = 4$ for the FCM clustering process, and then get four different fuzzy partitions, namely "Around 2.5", "Around 3.5", "Around 5", "Around 7.5" (the units are thousand). The resultant four fuzzy partitions are shown in fig. 1.

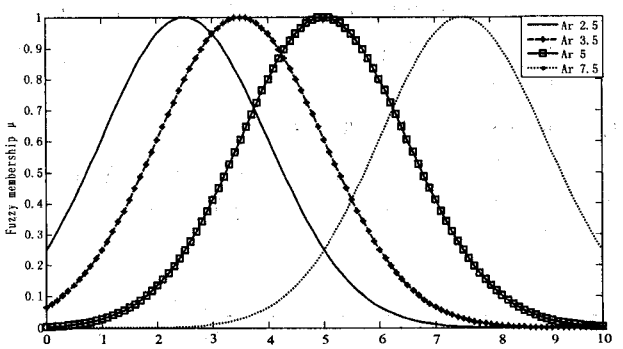

Fig 1　Fuzzy partitions generated by applying FCM

Through this numerical experiment, get fuzzy membership μ values which can work without Apriori membership function provided by human experts. Therefore, deal with real-life huge datasets which contain many quantitative attributes.

In fig. 2, the improvement of Apriori algorithm based on fuzzy clustering (FAFC) made a comparison with traditional Apriori algorithm on computing time for huge database. Use UCI datasets as the testing data. Set up min_sup=0.2, min_conf=0.3. From the results, it can be seen that much waste computation is fulfilled by traditional Apriori algorithm and FAFC algorithm can deal with huge datasets effectively.

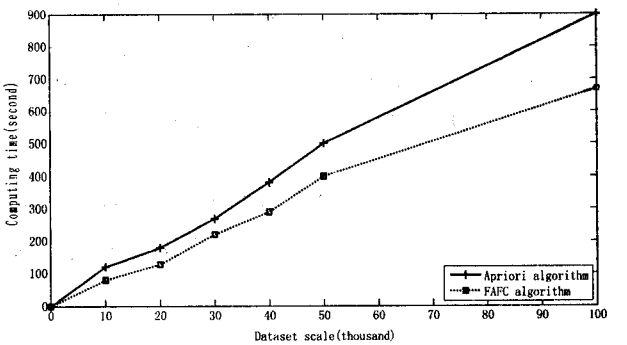

Fig 2　Computing time relative to dataset attribute Income

5 Conclusions

In this paper, a methodology by doing pre-processing for the original dataset based on FCM is proposed first. By doing this, get fuzzy membership μ values and fuzzy partitions which can work without Apriori membership function provided by human experts. And then an improvement of Apriori algorithm based on fuzzy clustering (FAFC) is presented. Experimental results show that improved Apriori algorithm (FAFC) outperforms the traditional Apriori algorithm on computing time for huge database.

参考文献:

[1] 毛国君,段立娟,王　实,等.数据挖掘原理与算法[M].北京:清华大学出版社,2005.

[2] 韩家炜,坎　伯.数据挖掘概念与技术[M].范　明,孟小峰译.第2版.北京:机械工业出版社,2007.

[3] Yue S, Tsang E, Yeung D, et al. Mining fuzzy association rules with weighted items[C]//Proceedings of the IEEE international conference on systems, man and cybernetics. [s. l.]: [s. n.], 2000:1906-1911.

[4] Watanabe T. An Improvement of Fuzzy Association Rules Mining Algorithm Based on Redundancy of Rules[C]//2010 2nd International Symposium on Aware Computing. [s. l.]: [s. n.], 2010:68-73.

[5] 孙晓霞,刘晓霞.模糊C均值聚类算法的实现[J].计算机应用与软件,2008,25(13):48-50.

(下转第26页)

和 ASR 已经达到了 Full 状态,即已经完成了数据库同步的工作,33686018 和 16843009 分别为 ASR 和 GSR 的 router-id。

```
                    接口配置信息
[root@localhost idmp-ripidd]# idconfig
    lo ID->  1                                  prefix 128
  eth0 ID->  10000-8000~2                       prefix 64
  eth0 ID->  65152~549-46079-65108-47411        prefix 64
                    路由配置信息
GSR# show running-config
Current configuration:
hostname GSR
!
idmp route 10201-300~/64 eth0
idmp route 11111~/64 eth0
!
interface eth0
 idmp ospfid interface type is route
!
router ospfid
 router-id 16843009
 interface eth0 area 0
!
                    路由状态信息
GSR# show idmp ospfid neighbor
RouterID  State/Duration  DR        BDR        I/F[State]
33686018  Full/00:01:43   33686018  16843009   eth0[BDR]
```

图5 GSR 的接口配置以及路由配置信息

图6使用 show idmp route 命令展示了 GSR、AR 上的路由表条目,由于在 ASR 进行了分离机制使得接入网络信息不能扩散到核心网络,即接入网的路由信息 13333-7000~/64 信息不会泛洪到 GSR 上,核心网路由信息 10000-8000~/64 以及2条外部路由不会泛洪到 AR 上。通过这种机制减少了核心网的路由条目,能够实现快速查询以及转发。

```
              GSR上的路由信息
GSR# show idmp route
Codes: K - kernel route, C - connected, S - static,
       R - RIPid, O - OSPFid, B - BGP, * - FIB route.

C>* 1/128 is directly connected, lo
O   10000-8000~/64 [110/0] is directly connected, eth0
C>* 10000-8000~/64 is directly connected, eth0
S>* 10201-300~/64 [1/0] is directly connected, eth0
S>* 11111~/64 [1/0] is directly connected, eth0
C * 65152~/64 is directly connected, eth0
              AR上的路由信息
AR# show idmp route
Codes: K - kernel route, C - connected, S - static,
       R - RIPid, O - OSPFid, B - BGP, * - FIB route.

C>* 1/128 is directly connected, lo
O   13333-7000~/64 [110/0] is directly connected, eth0
C>* 13333-7000~/64 is directly connected, eth0
C>* 65152~/64 is directly connected, eth0
```

图6 三台路由器的路由信息

5 结束语

通过对一体化网络下 OSPF 路由协议的设计和实现,完善了一体化网络的路由功能。测试结果表明 OSPFID 已经能够支持标识协议栈,并且能够正常地完成路由功能,具有良好的稳定性及可靠性,能够适合大规模网络的部署。

参考文献:

[1] 张宏科,苏 伟.新网络体系基础研究--一体化网络与普适服务[J].电子学报,2007,35(4):593-598.

[2] 杨 冬,周华春,张宏科.基于一体化网络的普适服务研究[J].电子学报,2007,35(4):607-613.

[3] 李 玮.一体化网络标识网络协议栈设计[D].北京:北京交通大学,2009.

[4] 姚 苏.接入网标识路由协议用户层的设计与实现[D].北京:北京交通大学,2010.

[5] 张宏科,苏 伟.路由器原理与技术[M].北京:高等教育出版社,2010.

[6] 张宏科,苏 伟.IPv6 路由协议栈原理与技术[M].北京:邮电大学出版社,2006.

[7] Coltun R,Ferguson D. OSPF for IPv6[S]. RFC 5340,2008.

[8] Deering S,Hinden R. Internet Protocol Version 6(IPv6) Address Architecture[S]. RFC3513,2006.

[9] Moy J. OSPF Version 2[S]. RFC 2328,1998.

[10] Nan Yao. Design and Implementation of Routing Protocol Extensions Supporting Separation of the Core and Access Network[J]. Journal of Internet Technology,2008,9(5):355-360.

[11] 李建昊,王志克,张春青,等. OSPFv3 详细设计[M].北京:北京交通大学,2003.

[12] 王江林. OSPFv3 概要设计[M].北京:北京交通大学,2002.

(上接第21页)

[6] 李 雷,罗红旗,丁亚丽.一种改进的模糊C均值聚类算法[J].计算机技术与发展,2009,19(12):71-73.

[7] Mangalampalli A,Pudi V. Fuzzy Association Rule Mining Algorithm for Fast and Efficient Performance on Very Large Datasets[C]//IEEE International Conference on Fuzzy Syetem.[s.l.]:[s.n.],2009:20-24.

[8] Lee Y C,Hong T P,Wang T C. Mining Fuzzy Multiple-level Association Rules under Multiple Minimum Supports[C]// Proc. of the 2006 IEEE International Conference on Systems, Man and Cybernetics.[s.l.]:[s.n.],2006:4112-4117.

[9] 罗军生,李永忠.基于模糊C-均值聚类算法的入侵检测[J].计算机技术与发展,2008,18(1):178-180.

[10] 吴 瑛,王秋生.模糊C均值聚类算法在web使用挖掘上的应用研究[J].计算机技术与发展,2008,18(6):32-35.

[11] Wu Zhenglong,Xiong Fanlun,Teng Minggui. Mining Fuzzy Association Rules for Numerical Attributes Based on Fuzzy Clustering[J]. MINI-MICRO SYSTEMS,2004,25(7):1295-1297.

[12] Gyenesei A. A Fuzzy Approach for Mining Quantitative Association Rules[R].[s.l.]:[s.n.],2001.