Journal of Southeast University (English Edition) Vol.20 No.1 Mar. 2004 ISSN 1003—7985

Classification methods of association rules with linguistic terms

Lu Jianjiang[1,2,3] Xu Baowen[1,3] Kang Dazhou[1,3]

([1] Department of Computer Science and Engineering, Southeast University, Nanjing 210096, China)
([2] Institute of Science, PLA University of Science and Technology, Nanjing 210007, China)
([3] Jiangsu Institute of Software Quality, Nanjing 210096, China)

Abstract: A partition of intervals method is adopted in current classification based on associations (CBA), but this method cannot reflect the actual distribution of data and still the problem of sharp boundary exists. In this paper, the classification system based on the longest association rules with linguistic terms is first discussed, and the shortcoming of this classification system is analyzed. Then, the classification system based on the short association rules with linguistic terms is presented. The example shows that the accuracy of the classification system based on the association rules with linguistic terms is better than two popular classification methods: C4.5 and CBA.

Key words: data mining; linguistic terms; association rules; classification

Classification and association rules are important research issues and focuses of data mining technology. There are a number of classification methods including decision tree induction[1], Bayesian classification and neural network, etc. The problem of mining Boolean association rules was first introduced in Ref.[2]. There are many known algorithms for mining Boolean association rules, such as Apriori[3], DHP[4], etc. Srikant and Agrawal first introduced the problem of mining quantitative association rules[5]. The algorithm in Ref.[5] found quantitative association rules by partitioning the attribute domain, combining adjacent partitions, and then transforming the problem into binary one. Association rules can also be used to classify. Liu, et al. proposed a classification method based on associations (CBA)[6], which used an iterative method to find the frequent and accurate possible rule set and then used the method of elicitation to build a classification system.

For mining association rules, quantitative attributes are handled by partitioning them into several intervals in CBA. But a partition of intervals method introduces some problems. The first problem is that the equi-depth partition cannot embody the actual distribution of the data. The second problem is caused by the sharp partition boundary. Ref.[7] used a fuzzy set to soften the partition boundary, and presented the concept of fuzzy association rules. Ref.[8] used linguistic clouds to soften the partition boundary. Ref.[9] used the relational fuzzy c-means algorithm to partition the quantitative attributes into several linguistic terms, then the problem of mining association rules with linguistic terms was introduced by combining linguistic terms. The relational fuzzy c-means algorithm can embody the actual distribution of the data. Furthermore, linguistic terms can soften the partition boundary. But combining linguistic terms can obtain excessive association rules with linguistic terms, so the mining algorithm cannot fit for a large database. Ref.[10] improved the algorithm in Ref.[9], and this improved algorithm can fit for a large database. The algorithm of mining fuzzy association rules with weighted items was presented in Ref.[11]. In this paper, classification systems based on association rules with linguistic terms are presented.

The rest of this paper is organized as follows. In section 1, the problem of mining association rules with linguistic terms is introduced. In section 2, the classification system, based on the longest association rules with linguistic terms, is discussed and the shortcomings of this classification system are analyzed. In section 3, the classification system based on short association rules with linguistic terms is presented. The conclusions are briefly noted in section 4.

1 An Algorithm for Mining Association Rules with Linguistic Terms

Let $T = \{t_1, t_2, \cdots, t_n\}$ be a relational database, t_j represents the j-th record in T, let $I = \{i_1, i_2, \cdots, i_m\}$ be the attribute set where i_j denotes a Boolean,

Received 2003-06-17.
Foundation items: Young Scientist's Fund of the National Natural Science Foundation of China (No. 60303024), the State Key Basic Research and Plan Program (973 Program) (No. 2002CB312000), Specialized Research Fund for the Doctoral Program of Higher Education, the Opening Foundation of Jiangsu Key Laboratory of Computer Information Processing Technology in Soochow University.
Biography: Lu Jianjiang (1968—), male, doctor, associate professor, jjlu@seu.edu.cn.

categorical or quantitative attribute, $t_j[i_k]$ represents the value of the j-th record in attribute i_k. Values of the record in attribute need to be partitioned into several linguistic terms for mining association rules with linguistic terms.

Let A_1 and A_2 be two values of the record in Boolean attribute, and then two values can be partitioned into two linguistic terms A_1 and A_2:

$$A_1(x) = \begin{cases} 1 & x = A_1 \\ 0 & x = A_2 \end{cases}$$

$$A_2(x) = \begin{cases} 1 & x = A_2 \\ 0 & x = A_1 \end{cases}$$

A categorical attribute with fewer values can be partitioned into several linguistic terms with the same method.

Each quantitative attribute is partitioned into several linguistic terms using the FCM algorithm[12]. These linguistic terms are usually represented with triangular fuzzy numbers for classification. The method of representing linguistic terms in triangular fuzzy numbers is as follows.

Let $\mu_k(x_i)$ be the grade of membership of x_i in the linguistic term with center v_k, let $X^k = \{x_i : \mu_k(x_i) \geqslant \mu_j(x_i),\ \forall j \in \{1, 2, \cdots, c\}\}$. We first find the samples with the minimum grade of membership at both sides of the center v_k in $X^k \cup \{v_k\}$, let the left sample with the minimum grade of membership be x^{l}, its grade of membership be $\mu_k(x^{\mathrm{l}})$, and let the right sample with the minimum grade of membership be x^{r}, its grade of membership be $\mu_k(x^{\mathrm{r}})$, then the expression of triangular fuzzy numbers $f(x)$ or (a, v_k, b) with center v_k is

$$f(x) = \begin{cases} \dfrac{x-a}{v_k-a} & a \leqslant x \leqslant v_k \\ \dfrac{b-x}{b-v_k} & v_k < x \leqslant b \\ 0 & x < a \text{ or } x > b \end{cases}$$

where $a = x^{\mathrm{l}} - \dfrac{\mu_k(x^{\mathrm{l}})(v_k - x^{\mathrm{l}})}{1-\mu_k(x^{\mathrm{l}})}$, $b = x^{\mathrm{r}} + \dfrac{\mu_k(x^{\mathrm{r}})(x^{\mathrm{r}} - v_k)}{1-\mu_k(x^{\mathrm{r}})}$.

In order to mine association rules with linguistic terms, a new database is first built through original database T. In this new database, attributes are linguistic terms, so attributes are called linguistic attributes. An association rule is an implication of the form $X \Rightarrow Y$. Because attributes in X and Y are linguistic attributes, $X \Rightarrow Y$ is called association rule with linguistic terms. The support of the linguistic attribute set and the support and confidence of $X \Rightarrow Y$ are defined in Ref.[10]. Linguistic attribute sets with at least a minimum support are called frequent linguistic attribute sets. The association rules with at least a minimum support and a minimum confidence respectively are called the interesting association rules with linguistic terms. The interesting association rules can be generated from frequent linguistic attribute sets, so discovering all interesting association rules is equal to discovering all frequent linguistic attribute sets. By the definition of the support of the linguistic attribute set, we can know that all subsets of a frequent linguistic attribute set must also be frequent. With the above finding, it is easy to modify the Apriori algorithm[3] to mine association rules with linguistic terms. Further details can be obtained in Ref. [10].

2 Classification System Based on the Longest Association Rules with Linguistic Terms

Let $I = \{i_1, i_2, \cdots, i_m, i\}$ be the attribute set of classification databases. Attribute i is a categorical attribute with values $C_1, C_2, \cdots, C_q$, which are all class labels. Let $y = (y_1, y_2, \cdots, y_m)$ be a sample, where $y_1, y_2, \cdots, y_m$ are the values taken by attributes $i_1, i_2, \cdots, i_m$. In this section, we will discuss how to use association rules with linguistic terms to classify the sample y. We use interesting association rules with linguistic terms to build a classification system. Suppose we use the algorithm in section 1 to discover M interesting association rules with linguistic terms as follows:

R_k: If i_1 is A_1^k and $\cdots$ and i_m is A_m^k, then i is C_j, $k = 1, 2, \cdots, M$

where $A_1^k, A_2^k \cdots, A_m^k$ are linguistic terms of attribute $i_1, i_2, \cdots, i_m$; C_j are class labels. These rules have $m+1$ attributes, so they are the longest association rules with linguistic terms. We use these association rules to build the rule base of classification system. When a sample y is to be classified, we can compute the discriminant function values of each class $g_h(y)$, $h = 1, 2, \cdots, q$, according to the following formula:

$$g_h(y) = \frac{\sum\limits_{1 \leqslant k \leqslant M,\ y = C_h} \prod\limits_{j=1}^{m} A_j^k(y_j)}{\sum\limits_{k=1}^{M} \prod\limits_{j=1}^{m} A_j^k(y_j)}$$

We compare these discriminant function values, and take the class label corresponding to the maximum value as the classification result of the sample y. This inference method considers the information provided by each rule for sample classification. At the same

time, because association rules with linguistic terms are easily understood, the classification system built has better interpretability.

In order to check the accuracy of our classification system, this paper discusses the Diabetes dataset from UCI Machine Learning Repository. Each quantitative attribute is partitioned into three linguistic terms represented with triangular fuzzy numbers by the FCM algorithm. In the experiment, ten-fold cross-validation method is applied to estimate the classification accuracy. The dataset is randomly divided into ten disjointed subsets, with each containing approximately the same number of records. Sampling is stratified by class labels to ensure that the subset class proportions are roughly the same as those in the whole dataset. For each subset, a classifier is built using the records not in it. The classifier is then tested on the withheld subset to obtain a cross-validation estimate of its accuracy. Then ten cross-validation estimates are averaged to provide an estimate for the classifier built from all the data. The cross-validation estimate in each subset is obtained as follows.

The number of interesting association rules with linguistic terms decides the complexity and accuracy of the classification system. A perfect classification system has a few rules and good accuracy. In order to control the complexity of the classification system, we first mine 1 000 interesting association rules on the withheld subset and rank these rules by their support. Then some rules that have high support are selected to evaluate the accuracy. In order to save computing time, we select the number of rules at a multiple of 50, such as 50, 100, 150, etc. We select 20 times and select the number of rules with the best accuracy. Tab. 1 shows the experimental results with the classification system based on the longest association rules with linguistic terms (LARLT).

Tab.1 Experimental results

Folds	Rule number	Training accuracy	Test accuracy
1	800	0.855 491	0.763 158
2	700	0.852 388	0.779 221
3	100	0.759 768	0.753 247
4	100	0.742 402	0.766 234
5	450	0.829 233	0.792 208
6	600	0.843 705	0.779 221
7	650	0.843 705	0.818 182
8	150	0.771 346	0.766 234
9	450	0.824 891	0.766 234
10	100	0.754 335	0.736 842
Average	410	0.807 726	0.772 078

We compare LARLT with two popular classification methods: C4.5[1] and CBA. In the experiment, the algorithm of C4.5 is downloaded from http://www.cse.unsw.edu.au/~quanlian, and the algorithm of CBA is downloaded from http://www.comp.nus.edu.sg/~dm2. Where the minimum support is set to 1%, the minimum confidence is set to 50%, and other parameters are unchanged. The accuracy of C4.5 is 74.2%, CBA is 74.5%, and LARLT is 77.207 8%. It is obvious that the accuracy of LARLT is better than CBA and C4.5 in the Diabetes dataset.

3 Classification System Based on Short Association Rules with Linguistic Terms

When a database has many quantitative attributes, the longest association rules with linguistic terms have very small support. But it costs too much time for mining association rules with very small support. In this section, we discuss how to use short association rules with linguistic terms to build a classification system.

3.1 Building the classification system

Suppose we use the algorithm in section 1 to discover M interesting short association rules with linguistic terms as follows:

R_k: If $X(1, k)$ is $B(1, k)$ and $\cdots$ and $X(l_k, k)$ is $B(l_k, k)$, then i is C_k, $k = 1, 2, \cdots, M$

where $X(1, k), X(2, k), \cdots, X(l_k, k) \in \{i_1, i_2, \cdots, i_m\}$, $B(l_k, k)$ are linguistic terms of attribute $X(l_k, k)$, $C_k \in \{C_1, C_2, \cdots, C_q\}$. Because l_k, $k = 1, 2, \cdots, M$ can be different from m, so these association rules are short association rules with linguistic terms.

We use these association rules with linguistic terms to build the rule base of the classification system. When a sample $y = (y_1, y_2, \cdots, y_m)$ is to be classified, compute the discriminant function values of each class $g_h(y)$, $h = 1, 2, \cdots, q$, according to the following formula:

$$g_h(y) = \frac{\sum_{1 \leqslant k \leqslant M,\ i = C_h} \prod_{j=1}^{l_k} B(j, k)[X(j, k)(y)]}{\sum_{k=1}^{M} \prod_{j=1}^{l_k} B(j, k)[X(j, k)(y)]}$$

where $X(j, k)(y)$ is the value taken by attribute $X(j, k)$ in the sample y, $B(j, k)[X(j, k)(y)]$ is the grade of membership of $X(j, k)(y)$ in linguistic terms $B(j, k)$, and $\prod_{j=1}^{l_k} B(j, k)[X(j, k)(y)]$ is the activated degree of sample y to the association rule R_k. We compare these discriminant function values, and take the class label corresponding to the maximum value as the classification result of the sample y.

In order to check the accuracy of our classification system, this paper discusses Wine dataset from UCI Machine Learning Repository, which has 13 quantitative attributes and 1 categorical attribute. Each quantitative attribute is partitioned into three linguistic terms represented with triangular fuzzy numbers by FCM algorithm. Ten-fold cross-validation method is applied to estimate the classification accuracy. Tab.2 shows the experimental results with the classification system based on short association rules with linguistic terms (SARLT). We compare SARLT with two popular classification methods: C4.5 and CBA. The accuracy of C4.5 is 92.7%, CBA is 91.6%, SARLT is 97.15686%. It is obvious that the accuracy of SARLT is better than that of CBA and C4.5 in the Wine dataset.

Tab.2 Experimental results

Folds	Rule number	Training accuracy	Test accuracy
1	100	0.956 522	1
2	250	0.962 5	1
3	250	0.937 5	1
4	150	0.975	0.944 444
5	100	0.968 75	0.888 889
6	150	0.975	1
7	50	0.962 5	1
8	50	0.962 5	1
9	50	0.968 75	1
10	50	0.981 366	0.882 353
Average	**120**	**0.965 038 8**	**0.971 568 6**

3.2 Simplifying the classification system

There are two issues that must be addressed in the classification system based on association rules with linguistic terms. The first is that a huge number of rules can contain noisy information. The second is that a huge set of rules would extend the classification time. This can be an important problem in applications where fast responses are required. So association rules with linguistic terms should be pruned. The pruning techniques that we employ are as follows.

Definition 1 Let $r: X \Rightarrow C$ be an association rule with linguistic terms, the lift of a rule r is defined as

$$\text{lift}(r) = \frac{\text{conf}(X \Rightarrow C)}{\text{conf}(C)}$$

where conf (C) is the expectation confidence of consequent C without arbitrary conditions.

If lift(r) is greater than 1, then rule r is positively correlated, meaning X encourages C. If lift (r) is less than 1, then rule r is negatively correlated, meaning X discourages C. If lift(r) is equal to 1, then X and C are independent. Association rules with lifts less than 1 or equal to 1 should be pruned. It is insufficient to prune the association rules with their lifts. The difference of minimum confidence (minconf_dif) is introduced next to prune the association rules farther.

Definition 2 Given two association rules with linguistic terms $r_1: X \Rightarrow C$ and $r_2: X' \Rightarrow C$, we say that the rule r_2 is a sub-rule of the rule r_1 if $X' \subseteq X$.

The improvement of an association rule with linguistic terms can be defined as the minimum difference between its confidence and the confidence of any proper sub-rule with the same consequent.

Definition 3 Given an association rule with linguistic terms $X \Rightarrow C$, the improvement of $X \Rightarrow C$ (imp$(X \Rightarrow C)$) is defined as

$$\min(\forall X' \subset X, \text{conf}(X \Rightarrow C) - \text{conf}(X' \Rightarrow C))$$

Given a minconf_dif, if imp$(X \Rightarrow C)$ is greater than minconf_dif, then the rule $X \Rightarrow C$ contains new information. Otherwise, we consider that the sub-rule of association rule $X \Rightarrow C$ contains the information provided by $X \Rightarrow C$. So association rule $X \Rightarrow C$ should be pruned. For example, given two association rules with linguistic terms:

Rule 1 If blood pressure is high and dextrose is normal, then he has an illness (sup = 10%, conf = 41%).

Rule 2 If blood pressure is high, then he has an illness (sup = 12%, conf = 40%).

Suppose minconf_dif is set to 5%. It is obvious that rule 2 is a sub-rule of rule 1. The difference between the confidence of rule 1 and the confidence of rule 2 is 1%, which is less than 5%. So rule 2 cannot provide new information and will be pruned.

To Wine dataset, Fig.1 shows the average number of the association rule pruned with different minconf_dif in the ten-fold cross-validation method. In addition, we can notice from Fig. 1 that association rules can be pruned effectively with the minconf_dif increasing. Fig.2 shows the average test accuracy with different minconf_dif. In addition, we can notice from Fig. 2 that some useful association rules may be pruned when the minconf_dif increases, this will make the test accuracy descend. A perfect classification system has a few rules and a good accuracy. Let minconf_dif be 0.10, the average number of the association rule is 52, the average test accuracy is 96.60%. Comparing with the classification system in section 3.1, the average number of the association rule descends to 68. The average test accuracy only descends 0.556 86%. So we can claim that the simplified classification system is a perfect classification system.

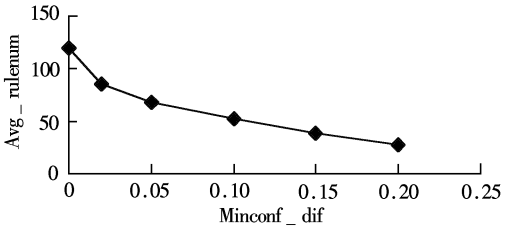

Fig.1 Average rule number

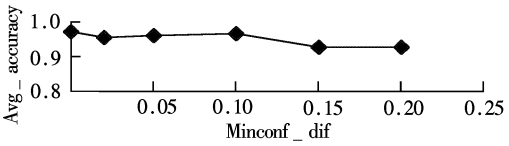

Fig.2 Average precision

4 Conclusion

In this paper, the classification system based on the longest association rules with linguistic terms is first discussed, and the shortcomings of this classification system are analyzed. Then the classification system based on the short association rules with linguistic terms is presented. Because fuzzy *c*-means algorithm can embody the actual distribution of the data and linguistic terms can soften the partition boundary, the classification system based on the association rules with linguistic terms can obtain better classification accuracy than two popular classification methods: C4.5 and CBA.

References

[1] Quinlan J R. *C*4.5: *programs for machine learning* [M]. San Mateo, CA: Morgan Kaufmann, 1993. 28 - 38.

[2] Agrawal R, Imieliski T,Swami A. Mining association rules between sets of items in large databases[A]. In: *Proceedings of ACM SIGMOD Conference on Management of Data* [C]. Washington DC, 1993. 207 - 216.

[3] Agrawal R, Srikant R. Fast algorithms for mining association rules [A]. In: *Proceedings of the International Conference on Very Large Databases* [C]. Santiago, Chile, 1994. 487 - 499.

[4] Park J S, Chen M S, Yu P S. An effective hash-based algorithm for mining association rules [A]. In: *Proceedings of the* 1995 *ACM-SIGMOD International Conference on Management of Data* [C]. San Jose, CA, 1995. 175 - 186.

[5] Srikant R, Agrawal R. Mining quantitative association rules in large relational tables [A]. In: *Proceedings of the ACM-SIGMOD Conference on Management of Data* [C]. Montreal, Canada, 1996. 1 - 12.

[6] Liu B, Hsu W, Ma Y. Integrating classification and association rule mining [A]. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining*[C]. New York, 1998. 80 - 86.

[7] Chan M K, Ada F, Man H W. Mining fuzzy association rules in database [A]. In: *Proceedings of the ACM Sixth International Conference on Information and Knowledge Management* [C]. Las Vegas, Neveda, 1997.10 - 14.

[8] Lu Jianjiang, Qian Zuoping, Song Ziling. Application of normal cloud association rules on prediction[J]. *Journal of Computer Research and Development*, **2000, 37**(11): 1317 - 1320. (in Chinese)

[9] Lu Jianjiang, Qian Zuoping, Song Ziling. Mining linguistic valued association rules [J]. *Journal of Software*, 2001, **12**(4): 607 - 611. (in Chinese)

[10] Zou Xiaofeng, Lu Jianjiang, Song Ziling. Mining linguistic valued association rules[J]. *Journal of System Simulation*, **2002, 14**(9): 1130 - 1132. (in Chinese)

[11] Lu Jianjiang. Research on algorithms of mining association rules with weighted items [J]. *Journal of Computer Research and Development,* **2002, 39**(10): 1281 - 1286. (in Chinese)

[12] Hathaway R J, Davenport J W, Bezdek J C. Relational **dual of the c**-means algorithms [J].*Pattern Recognition,* **1989, 22**(2): 205 - 212.

语言值关联规则分类方法

陆建江[1,2,3] 徐宝文[1,3] 康达周[1,3]

([1] 东南大学计算机科学与工程系, 南京 210096)
([2] 解放军理工大学理学院, 南京 210007)
([3] 江苏省软件质量研究所, 南京 210096)

摘要: 目前采用的区间划分的关联分类法不能有效地体现出数据的实际分布情况, 并存在划分边界过硬的缺点.文中首先讨论了通过挖掘最长的语言值关联规则构建分类系统的方法并分析了其不足, 然后给出了通过挖掘短的语言值关联规则构建分类系统的方法.实验表明, 基于语言值关联规则的分类系统能在精度上优于 2 种流行的分类方法: C4.5 和关联分类法.

关键词: 数据挖掘; 语言值; 关联规则; 分类

中图分类号: TP311.13