An Approach to Classification Based on Fuzzy Association Rules

Zuoliang Chen, Guoqing Chen

School of Economics and Management, Tsinghua University, Beijing 100084, P. R. China

Abstract

Classification based on association rules is considered to be effective and advantageous in many cases. However, the "sharp boundary" problem in association rules mining with numerical data may lead to semantics retortion of discovered rules, which may further disturb the understandability, even the accuracy of classification. This paper aims at proposing an associative classification approach, namely Fuzzy Association Rules Classification (FARC), where fuzzy logic is used in partitioning the domains of numerical data items, giving rise to fuzzy association rules for classification. In doing so, two measures, pseudo support and pseudo confidence, as well as the notion of minimal equivalence set (*MESet*), are introduced, along with extensions to the corresponding mining algorithms. The experimental results revealed that FARC generated fewer rules than the traditional CBA approach without loss of accuracy.

Keywords: Associative Classification, Fuzzy Association rules, FARC, Data Mining.

1. Introduction

Association rule mining [1] and classification are two popular methods used in knowledge discovery in different application areas, including finical market, bioinformatics, web analysis, and so on. The goal of association rule mining is to generate certain associative relationships between data items with confidence and support greater than user specified thresholds. Classification is used to find a logical description that results from training datasets with predetermined targets, and could group unlabeled datasets. There are two basic criteria for classification, i.e., accuracy and simplicity. There exist several ways of constructing classifiers, including the ones based on association rules, such as CBA[2], CMAR[3], CPAR[4] and GARC[5]. These classification approaches have received considerable attention due to their accuracy and understandability. Moreover, a number of efforts have put forward to focus on the various aspects of improvements. However, the "sharp boundary" problem in association rules mining with numerical data may lead to semantics retortion of discovered rules, which may further disturb the understandability, even the accuracy of classification.

The purpose of the work proposed in the paper is to demonstrate the usability of a novel classifier called FARC, i.e. Fuzzy Association Rules Classifier, which can deal with the "sharp boundary" problem well. We initialize the population with the fuzzy association rules [6] whose consequents are all class label. At the beginning, all rules will have the equal weight. In the process of the algorithm, the weight of one rule will increase if it classifies one case correctly, and will decrease if not. The rule with the smallest weight would be removed. The process will continue until the training set error stagnates. The remaining rules construct the FARC. Finally, experimental results will show that FARC can generate fewer rules than the traditional CBA approach, with similar accuracy.

This paper is organized as follows. Section 2 introduces the background knowledge required for the proposed architecture, including association rules, and fuzzy association rules. Section 3 describes the definition of fuzzy association rules with new interesting measures, the fuzzy *pseudo support* and the fuzzy *pseudo confidence*. In section 4, the proposed learning algorithm is presented. In section 5, the performance of the proposed algorithm is examined by computer simulation on some data sets. Conclusions are presented in Section 6.

2. The background

2.1. Association rules

In general, an association rule is of the form of $X \Rightarrow Y$, expressing the semantics that "occurrence of X is associated with occurrence of Y", where X and Y are collections of data items. An example of an association rules is "*Apples* & *Bananas* $\Rightarrow$ *Pork*, with the degree of support (*Dsupport*) = 20% and degree of confidence (*Dconfidence*) = 80%" meaning that "20% of all the customers bought *Apples*, *Bananas* and *Pork* simultaneously, and 80% of the customers who bought

Apples and *Bananas* also tended to buy *Pork*". Such association rules called boolean ones, since the association concerned is the correspondence of the states, each being a binary value 0 or 1. *Agrawal et al.*[7] has proposed the Apriori algorithm to quickly find boolean association rules.

Though boolean association rules are meaningful in real world applications, there are usually categorical or quantitative. Usually, quantitative items are represented in a database as attributes whose values are elements of continuous domains such as Real Number Domain *R*. An example is shown in Table 1.

D	*Age*	*Height*	……
ID1	31	170	……
ID2	25	180	……
ID3	16	182	……
ID4	52	165	……
……	……		……

Table 1: Database (*D*) with Continuous Domains.

As we know, the typical Apriori algorithm is incapable of dealing directly with such databases. Therefore, in [8], an algorithm has been proposed to mine quantitative association rules. The algorithm transforms *D* into a binary database D' by partitioning the attribute domains, and then transforming the problem into binary one. For example, D' is a binary database with new attributes as shown in Table 2.

D	*Age(0,27]*	*Age(27,65]*	*Age(60,100]*	……
ID1	0	1	0	……
ID2	1	0	0	……
ID3	1	0	0	……
ID4	0	1	0	……
……	……	……	……	……

Table 2: Database (D') Transformed from *D*.

An example of quantitative association rules may be "*Age(0,27)* $\Rightarrow$ *Height(170,172)*". The Support and Confidence, as the interesting measures of quantitative association rules, are defined similarly.

2.2. Fuzzy association rules

Although the above quantitative association rule mining algorithms can solve some problems introduced by quantitative attributes, they introduce some other problems. The first problem is caused by the sharp boundary between intervals. The algorithms either ignore or over-emphasize the elements near the boundary of the intervals in the mining process. The use of sharp boundary intervals is not intuitive with respect to human perception. To cope with the problem, researchers discover association rules with fuzzy sets [9]. Such sets are usually expressed in forms of labels or linguistic terms. For example, for attribute *Age*, some fuzzy sets may be defined on its domain U_{Age} such as *Young*, *Middle* and *Old*. In this way, these new attributes (e.g. *Young-Age*, *Middle-Age* and *Old-Age* in place of *Age*) will be used to constitute a new database with partial belongings of original attribute values to each of the new attributes. Table 3 illustrates an example of the new database obtained from the original database, given fuzzy sets *Young* (*Y*), *Middle* (*M*) and *Old* (*O*) as characterized by membership functions shown in Figure 1.

D'	*Young-Age*	*Middle-Age*	*Old-Age*	……
ID1	0.8	0.4	0.1	……
ID2	0.9	0.3	0	……
ID3	1	0	0	……
ID4	0.1	0.2	0.8	……
……	……	……	……	……

Table 3: Database (D'') with Fuzzy Items.

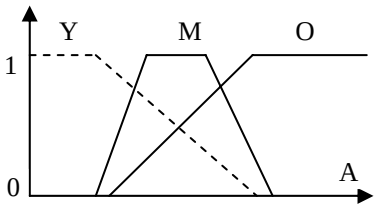

Figure 1: Fuzzy Sets *Young (Y), Middle (M) and Old (O).*

Generally, for original database *D* with attributes $A = \{I_1, I_2, \ldots, I_m\}$, each I_k ($1 \le k \le m$) can be associated with q_k fuzzy set defined on the domain of I_k, and usually labeled as q_k new attributes. We use $F_k = \left\{ I_k^1, I_k^2, ..., I_k^{q_k} \right\}$ to represent the set of fuzzy sets associated with I_k. That is, the new database D'' is with respect to schema *R(A')* where $A' = \left\{ I_1^1, ..., I_1^{q_1}, ..., I_k^1, ..., I_k^{q_k}, ..., I_m^1, ..., I_m^{q_m}, ... \right\}$.

3. Fuzzy association rules in classification

To make associations suitable for classification task, the consequents of the rules should be limited to class label values only. Thus, an example of fuzzy association rules in classification is in the following form:

$$F \Rightarrow C.$$

In the above rule, $F = \{f_1, f_2, \ldots, f_p\}$ is the subset of *A'*, and $\left| F \cap F_k \right| \le 1$, where $1 \le k \le m$ and $\left| X \right|$ represents the number of elements in set *X*. *C* is a possible class, and it is crisp.

According to the semantics of the rule, we can imply that the class label of the case *d* is *C* with the occurrence of *F*. If the degree that a case *d* belongs to

f_l is $\mu_l(d)$ $(1 \le l \le p)$, we can define the degree that the case d belongs to F as

$$\mu_F(d) = \min\{\mu_1(d), \mu_2(d), ..., \mu_p(d)\}.$$

The definitions of *Dsupport* and *Dconfidence* can be adapted to the fuzzy association rules in classification [10]. According to [10], the *Dsupport* of the rules can be computed as

$$S(F \Rightarrow C) = \frac{\sum_{d[I_c]=C} \mu_F(d)}{|D|},$$

and the *Dconfidence* can be computed as

$$C(F \Rightarrow C) = \frac{\sum_{d[I_c]=C} \mu_F(d)}{\sum_{d \in D} \mu_F(d)},$$

where $d[I_c]$ represent the class label of the case d.

In this paper, we propose two new measures named *Pseudo Support* and *Pseudo Confidence*. Assume *MS* is a user-specified number belonging to [0, 1). The *Pseudo Support* of a fuzzy rule can be computed as

$$PS(F \Rightarrow C) = \frac{\sum_{d[I_c]=C} \mu_F(d) \lceil \max(0, \mu_F(d) - MS) \rceil}{|D|},$$

and the Pseudo Dconfidence can be computed as

$$PC(F \Rightarrow C) = \frac{\sum_{d[I_c]=C} \mu_F(d) \lceil \max(0, \mu_F(d) - MS) \rceil}{\sum_{d \in D} \mu_F(d) \lceil \max(0, \mu_F(d) - MS) \rceil}.$$

The calculating of Pseudo *Support* and *Pseudo Confidence* are similar to the typical ones, respectively, except that these degrees smaller than *MS* are ignored. Tables 4-6 illustrate the ideas.

D	*Age*	*Class*	……
ID1	50	A	……
ID2	50	A	……
……	……	……	……
ID10	50	A	……
ID11	20	……	……
……	……	……	……
ID100	20	……	……

Table 4: Database (*D*) with Continuous Domains.

D	*Age(27,65]*	*Class*	……
ID1	1	A	……
ID2	1	A	……
……	……	……	……
ID10	1	A	……
ID11	0	……	……
……	……	……	……
ID100	0	……	……

Table 5: Database (*D'*) Transferred from *D*.

From Table 4, we may obtain Table 5 by partitioning the domain of *Age*, and Table 6 by fuzzy extensions. A rule "*Age(27, 65]* $\Rightarrow$ *A* could be obtained from Table 5 with its *Dconfidence*=100%, but a fuzzy rule like "*Middle-Age* $\Rightarrow$ *A* could not from Table 6 because the *Dconfidence* may be only 60%, which should be smaller than the user-specified threshold. That is to say, the interesting measures may ignore some interesting rules.

D	*Middle-Age*	*Class*	……
ID1	1	A	……
ID2	1	A	……
……	……	……	……
ID10	1	A	……
ID11	0.1	……	……
……	……	……	……
ID100	0.1	……	……

Table 6: Database (D'') with Fuzzy Items.

The *Pseudo Support* and *Pseudo Confidence* can deal with this problem with the right *MS*, and we set it to 10% in the following experiment. Furthermore, the corresponding mining method can be developed upon these measures, which is also an Apriori-type extension.

4. Classification based on fuzzy associations

Several publications have managed to mine fuzzy association rules [11][12][13][14], we use the algorithm in [11] to generate fuzzy rules with the new interesting measures. Before presenting the FARC algorithm, let us introduce some definitions on fuzzy association rules.

Definition 1: Given two fuzzy association rules, r_i and r_j, $r_i > r_j$ (also called r_i precedes r_j) if
1. the *Pseudo Confidence* of r_i is greater than that of r_j, or
2. their *Pseudo Confidences* are the same, but the *Pseudo Support* of r_i is greater than that of r_j, or
3. both the *Pseudo Confidences* and *Pseudo Supports* of r_i and r_j are the same, but r_i is generated earlier than r_j;

Definition 2: Given two fuzzy association rules in classification, r_i and r_j, $r_i \succ r_j$ (also called r_j is inferior to r_i) if
The antecedent part of r_i is the subset of that of r_j, and $r_i > r_j$.

Definition 3: Given a set of fuzzy association rules in classification R and its subset R', R' is the minimum equivalence set (*MESet*) of R if
$\forall r \in R - R'$, $\exists r' \in R'$, where $r' \succ r$, and
$\forall r \in R'$, if r' $r' \succ r$, then $r' \notin R'$.

Proposition 1: There exists one and only one *MESet* of any rule set *R*.

Proof: Firstly, the algorithm to obtain the *MESet* of *R* is shown in Figure 2. It has three steps:

- Step 1 (line 1): Sort the set of generated rules *R* according to the relation ">".
- Step 2 (line 2-8): Delete the rules that are inferior to others.

Thus, we get the *MESet* and prove its existence.

Secondly, we suppose that R' and R'' are both the *MESet* of *R*. If there exists a rule $r \in R'' - R'$, so $\exists r' \in R'$ and $r' \succ r$ because R'' is a *MESet*. Obviously, $r' \notin R''$, so $\exists r'' \in R''$ and $r'' \succ r'$. That is to say, $r'' \succ r$ and $r'', r \in R''$, which contradicts the definition of *MESet*. So there exists no rule $r \in R'' - R'$, which means $R'' - R' = \phi$. We can draw $R' - R'' = \phi$ in the same way. So $R'' = R'$. That is to say, there exists only one *MESet* of *R*. □

```
1    R' =sort(R)
2    for each rule r ∈ R' in sequence do
3        for each rule r' ∈ R' and r > r' in sequence do
4            if r ≻ r' then
5                delete r' from R'
6        end
7    end
8    return R' (the MESet of R)
```

Figure 2: Algorithm of *MESet*.

Definition 4: Given a fuzzy rule *r* with the form of "$F \Rightarrow C$" and a case *d* belonging to *D*, the confidence of classifying *d* with *r* is

$$DF = \mu_F(d) * PC(r) * w_r ,$$

where w_r is the weight of the rule. *DF* contents three factors: the degree that the case *d* belongs to *F*, the *Pseudo Confidence* of the rule and the weight of the rule. Each factor influences the rule *r* in classifying the case *d*, so we multiply them together as the confidence of the classifying.

Let *R* be the MESet of generated rules, and D'' the training data set with fuzzy items. The basic idea of FARC algorithm is to choose a set of rules with high weight from *R* and to select the "best" rule to cover each d'' in D''. Our classifier is of the following format:

$$<r_1, r_2, \ldots, r_n, default\_class>,$$

where *default_class* is the default class. In classifying an unseen case, sort *R* by *DF* and the first rule will be chosen to classify it. If there is no such rule that applies to the case, it takes on the default class. Our algorithm for building such a classifier has three stages:

1. Training

   The algorithm for training the rule set *R* is shown in Figure 3. The training process has 3 steps:

   - Step 1 (line 2-8): calculate the *DF* of each rule and sort the rules by *DF*.
   - Step 2 (line 9-22): select the rule in sequence until the case is classified correctly. For each rule *r*, *RightN* and *WrongN* are used to record the number of cases it has classified right and wrong, respectively.
   - Step3 (line 1): iterate step 1 and 2 for each case, and then compute the weight w_k of each rule whose initial value is 1.

```
1.  for each d'' ∈ D'' do
2.      for each r ∈ R do
3.          r.DF = μ_F(d'') * PC(r) * w_r
4.          r.use=0
5.          r.RightN=0
6.          r.WrongN=0
7.      end
8.      sort R by r.DF desc
9.      whe=0
10.     while(whe=0) do
11.         r=fisrt rule of R
12.         if r.use=1 then
13.             Break
14.         r.use=1
15.         if clas(r, d'') then
16.             r.RightN++
17.             whe=1
18.             break
19.         else
20.             r.WrongN++
21.             move r to the bottom of R
22.     end
23. end
24. for each r ∈ R do
25.     r.w_k= r.RightN/(r.RightN+r.WrongN)
26. end
```

Figure 3: Algorithm of Training.

2. Delete the "worst" rule from *R* and compute the error rate.

   We discard those rules that do not improve the accuracy of the classifier, including those whose *RightN* is zero, and the one with the smallest w_k. We then compute and record the total number of errors that are made by the current classifier and the default class. This is the sum of the number of errors that have been made by all the selected rules in the classifier and the number of errors to be made by the default class in the training data.
3. Iterate stage 1 and 2 until the error rate on training set increases.

5. Experimental Results

To evaluate the effectiveness of our approach, we now compare the classifiers produced by algorithm FARC

with those produced by CBA. We use some datasets from UCI ML Repository [15] for the purpose.

In our experiments reported below, we set the threshold of the Pseudo *Support* (*minpsup* for short) to 10%. For the *Pseudo Confidence*, its threshold (*minpconf*) is set to be 85%. For the *MS*, it is more complex. *MS* has a strong effect on the quality of the classifier produced. If *MS* is set too low, the interesting measures perform the same with the typical ones, and some useful rules may not be included. If *MS* is set too high, the rules generated will not be as "confident" as they appear, and some interesting rules with high *pseudo confidence* may be not generated because of their low *Pseudo Support*. Thus, the accuracy of the classifier may suffer in both situations. In the experiments, we set *MS* to 10%.

Discretization of continuous attributes is done using the Entropy method [16], and then triangular membership functions are specified by the discretized points for fuzzy items. In the experiments, all CBA parameters had their default values. The basic information of the dataset is listed in Table 7. The experimental results are shown in Table 8 and 9.

	Dataset	Attr.	Num. of attr.	Num. of training data	Num. of testing data
1	Australian	Dis., con.	14	460	230
2	Breast	Con.	10	466	233
3	Heart	Con.	13	180	90
4	Wine	Con.	13	118	60

Table 7: Basic information of the datasets.

	Dataset	CAB%	FARC%
1	Australian	87.39	87.83
2	Breast	96.57	95.71
3	Heart	83.33	85.55
4	Wine	86.67	93.33
	Mean	88.49	90.61
	Standard deviation	5.67	4.71

Table 8: Algorithms' accuracy on CBA and FARC.

	Dataset	CAB	FARC
1	Australian	110	21
2	Breast	31	13
3	Heart	31	22
4	Wine	13	13
	Mean	46.25	17.25

Table 9: Number of rules generated by CBA and FARC.

Accuracy is one of the basic performance measures for classification algorithms. Table 8 show the accuracy results compared with CBA, which indicate that the classification accuracy of FARC is satisfactory. On average, the accuracy of FARC seemed similar to that of CBA. Moreover the FARC appeared the same stable as CBA in terms of standard deviations of accuracy. These findings could be further justified by statistical significance tests. The testing results revealed that on average the accuracy of FARC was not significantly different from that of CBA.

Table 10 tabulates the number of rules generated by CBA and FARC. Clearly, FARC generated fewer rules than CBA, providing better understandability, even more stable performance. The main reason of fewer rules is that FARC uses fuzzy association rules for classification. Because of smooth boundary, a fuzzy rule can cover more cases than a crisp one with the same original attributes and discretized points. That is to say, FARC need fewer rules than CBA to cover all of the cases. In addition, the number of rules in FARC could also be reduced by using pruning/resolution strategies such that certain conflicting and redundant rules could be dropped.

6. Conclusion

This paper proposed a framework to integrate classification and fuzzy association rule mining. New interesting measures have been presented to generate all fuzzy association rules in classification, and an algorithm, FARC, has also been proposed to build an accurate classifier. Compared with CBA, the new approach has better understandability because of the terms of number of rules and the smooth boundary. In our future work, we will focus on improving efficiency of the classifier and on proposing novel techniques for discretization.

Acknowledgement

The work was partly supported by the National Natural Science Foundation of China (70231010/70621061) and the Research Center for Contemporary Management, Tsinghua University.

References

[1] R. Agrawal, T. Imielinski and A.Swami, Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference Management of Date*, pp. 207-216, Washington, 1993.

[2] B. Liu, W. Hsu and Y. Ma, Integrating classification and association rule mining. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining* (SIGKDD98), pp. 80-86, New York, 1998.

[3] W. Li, J. W. Han and J. Pei, CMAR: Accurate and efficient classification based on multiple classification rules. In *Proceedings 2001 IEEE*

International Conference on Data Mining (ICDM 2001), pp. 369-376, California, 2001.

[4] X. Yin and J. Han, CPAR: Classification based on predictive association rules. In *Proceeding of 3rd SIAM International Conference on Data Mining* (SDM'03), San Francisco, CA, 2003.

[5] G. Chen, H. Liu, L,Yu, Q. Wei and X. Zhang, A New Approach to Classification Based on Association Rule Mining. *Decision Support System*, 42: 674-689, 2006.

[6] H. Ishibuchi, T. Nakashima and T. Yamamoto, Fuzzy Association Rules for Handling Continuous Attributes. In *Proceedings of the IEEE International Symposium on Industrial Electronics*, pp. 118-121, Korea, 2001.

[7] R. Agrawal and R. Srikant, Fast algorithms for mining association rules. In *Proceedings of the international conference on very large databases*, 1994.

[8] R. Srikant and R. Agrawal, Mining Quantitative Association Rules in Large Relational Tables. In *Proceedings of the ACMSIGMOD Conference on Management of Data*, pp. 1-12, Montreal, Canada, 1996.

[9] L.A. Zadeh, Fuzzy sets. *Information Control*, 8: 338-353, 1965.

[10] T. Hong, C. Kuo, S. Chi and S. Wang, Mining Fuzzy Rules from Quantitative Data Based on the AprioriTid Algorithm. In *Proceedings of the ACM SAC 2000, Fuzzy Application and Soft Computing Track*, pp. 534-536, Italy, 2000.

[11] G. Chen and Q. Wei, Fuzzy Association Rules and the Extended Mining Algorithms. *Information Sciences*, 147: 201-228, 2002.

[12] W. Au and K. C. C. Chan, An effective algorithm for discovering fuzzy rules in relational databases. In *Proceeding IEEE International Conference Fuzzy Systems* (FUZZ IEEE 98), pp. 1314-1319, 1998.

[13] R. B. V. Subramanyam and A. Goswami, Mining fuzzy quantitative association rules. *Expert Systems* 23: 212-225, 2006.

[14] T. Hong, K. Lin and S. Wang, Fuzzy data mining for interesting generalized association rules. *Fuzzy Sets and Systems* 138: 255-269, 2003.

[15] C. Merz and P. Murphy, UCI repository of machine learning databases. http://www.cs.uci.edu/~mlearn/MLRepository.html, 1996.

[16] U. M. Fayyad and K. B. Irani, Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pp. 1022-1027, 1993.