Tendency Mining in Dynamic Association Rules Based on SVM Classifier

Zhonglin Zhang*, Zongcheng Liu and Chongyu Qiao

School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou, Gansu, 730070, China

Abstract: A method of tendency mining in dynamic association rule based on compatibility feature vector SVM classifier is proposed. Firstly, the class association rule set named CARs is mined by using the method of tendency mining in dynamic association rules. Secondly, the algorithm of SVM is used to construct the classifier based on compatibility feature vector to classify the obtained CARs taking advantage when dealing with high complex data. It uses a method based on judging rules' weight to construct the model. At last, the method is compared with the traditional methods with respect to the mining accuracy. The method can solve the problem of high time complexity and have a higher accuracy than the traditional methods which is helpful to make mining dynamic association rules more accurate and effective. By analyzing the final results, it is proved that the method has lower complexity and higher classification accuracy.

Keywords: Associative classification, classifier, data mining, dynamic association rules, SVM algorithm, tendency mining.

1. INTRODUCTION

Associative classification is an important prediction method in data mining of which the algorithm is to integrate the association rule mining and classification. Since the first classification algorithm named CBA [1] was introduced in 1998, it has been very active to design and apply the algorithm in addition to some new associative classification algorithm in research. In 2001, J. Li proposed the CMAR [2] algorithm based on multiple association rules which used the improved frequent item type growth method named FP-Growth method to mine inertia rules and used multiple strong association rules based on the weight to determine the new instance class label. The method was proved to have higher accuracy than the CBA algorithm. In 2003, Yin proposed the predict type classification algorithm CPAR [3] which used the greedy method to excavate smaller set of rules from the data set. In 2004, Antonio proposed an algorithm based on positive and negative rules [4].

Then in 2005, Wang proposed the HARMONY [5] algorithm which mined the highest confidence rules to cover the sample directly. Adriano Veloso proposed the Lazy classifier [6]. But the method has performance problems when facing large data sets, so it cannot work in large data set to mine all the rules.

Due to the associative classification algorithm considering all the possible relationships between the projects, there may be a large number of redundant rules. On the basis of the above, this paper proposes a method of tendency mining in dynamic association rule based on associative classification. Tendency mining was proposed by introducing a tendency threshold to improve the traditional association rules mining methods to mine those dynamic association rules under a certain trend on the basis of algorithm in support-confidence framework. In associative classification mining, there exist the characteristics of high precision, so by combining the two ways together, it can provide more accurate support for association rule mining method.

2. TENDENCY MINING IN DYNAMIC ASSOCIATION RULES

Tendency mining [7] is a method of dynamic association rule mining according to the characteristics of the rules changing over time on the basis of SV (support vector) and CV (confidence vector). It uses a tendency threshold to eliminate useless rules to reduce the candidate item set and then to generate the tendency rules to find valuable rules to improve the quality of mining. In this paper, it determines the dynamic tendency rules based on the confidence. In order to mine valuable dynamic association rules, the patterns [8, 9] should be known at first:

(1) Stable change. Some patterns of phenomenon do not have obvious changes over time.

(2) Strengthen trend change. Some patterns have obvious rising trend over time.

(3) Weaken trend change. Some patterns have obvious downward trend over time.

(4) Cyclical change. The identical pattern repeats under the time interval.

(5) Random change. Some patterns have no obvious regular change.

Here are some lemma and inference below:

Lemma 1: If an item set $F_c \cos\theta + G\sin\theta \le F_f = (G\cos\theta - F_c \sin\theta) f$ in dataset $F_c = \frac{mu^2}{R}$ satisfies the condition $\sup(X)_0 \ge \min\_\sup$, then

*Address correspondence to this author at the School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou, Gansu, China; Tel: 86-931-4956739; Fax: 86-931-4955743; E-mail: zhangzl@mail.lzjtu.cn there is at least one $i(1 \le i \le n)$ to make $u_{cf\max} \le \sqrt{\frac{gR(B-2h_g \tan\theta)}{2h_g + B\tan\theta}}$.

Inference 1: If an item set $u_{ch\max} \le u_{cf\max}$ meets $\sqrt{\frac{gR(f\cos\theta - \sin\theta)}{f\sin\theta + \cos\theta}} \le \sqrt{\frac{gR(B\cos\theta - 2h_g \sin\theta)}{2h_g \cos\theta + B\sin\theta}}$ for all $f \le \frac{B}{2h_g} \bullet \frac{1+tg^2\theta}{1-tg^2\theta}$, then $\sup(X)_i < s \times d_i$.

So dynamic association rule mining can have the following definitions on the basis above.

Definition 1: If each element in support vector (SV) of rule $A => B$ meets the condition $Sup_{(A\cup B)} > \min\_\sup$, then the rule is called a stability rule based on support. In a similar way, if each element in confidence vector (CV) of rule $A => B$ meets the condition $Conf_{(A\cup B)} > \min\_conf$, then the rule is called a stability rule based on confidence.

Definition 2: If the rule does not meet the definition 1, but the elements in SV satisfy that $Sup_{(A\cup B)_i} < Sup_{(A\cup B)_{i+1}}$, then the rule is called support rising type dynamic association rule. It is similar in CV . If the elements in CV satisfy that $Conf_{(A\cup B)_i} < Conf_{(A\cup B)_{i+1}}$, then the rule is called confidence rising type dynamic association rule.

Definition 3: If the rule does not meet the definition 1, but the elements in SV satisfy that $Sup_{(A\cup B)_i} \ge Sup_{(A\cup B)_{i+1}}$, the rule is called support drop type dynamic association rule. It is similar in CV . If the elements in CV satisfy that $Conf_{(A\cup B)_i} \ge Conf_{(A\cup B)_{i+1}}$, then the rule is called confidence drop type dynamic association rule.

Definition4: If each element in SV does not meet the definition 1 to 3, but at a particular time period $t = \{t_1, t_2, \ldots, t_n\}$, the value of the elements in SV appears alternately. The rule is called support cycle type dynamic association rule. It is similar for CV .

Definition 5: If a rule does not meet the definition 1 to 4, but for an 'n' length of time series, there exists an 'm' length of child sequence $U = \{Sup_{(A\cup B)_k}, \ldots, Sup_{(A\cup B)_p}, \ldots, Sup_{(A\cup B)_q}\}$. If the support of the previous item is less than the after item, then U is called a rising time support vector sequence with the other conditions called drop type sequence. If the length of rising sequence (n) is not more than m, then it is called the biggest rising time support vector sequence. And in the drop type, it is similar to CV .

Definition 6: Setting L as the length of the biggest rising time support vector sequence and S as the length of the biggest drop time support vector sequence in SV , the trend degree based on confidence vector is described as follows.

$SRI = \max\{L, S\} / n$

The trend degree based on support confidence is described as follows.

$CRI = \max\{L, S\} / n$

Definition 7: If a rule meets the definition 1 to 4, it is called a high interesting rule. If its trend degree is above the given DRI threshold, it is called a strong rule.

So the strong rules can be described as follows.

For given data set D , $\min\_\sup, \min\_conf$ and DRI , when a rule meets the following conditions, it is called a strong rule. And s stands for its support, c stands for its confidence, SRI stands for its trend degree based on support vector and CRI stands for its trend degree based on confidence vector.

(1) $s \ge \min\_\sup$;

(2) $c \ge \min\_conf$;

(3) $SRI \ge DRI(CRI \ge DRI)$.

3. METHOD OF ASSOCIATIVE CLASSIFICATION

Mining association rules can generally be divided into two steps. The first step is to find all the frequent and accurate possible rule sets named categorical association rules (CARs) which is the first but most important step of associative classification.

3.1. Mining CARs

The traditional dynamic association rule mining algorithm is improved by introducing a tendency threshold to mine rules under a certain trend on the basis of support and confidence. The rules mined according to the definition 1 to 6 have the following conditions.

(1) $s < \min\_\sup$; It means that the rules' importance is not strong and they need to be deleted.

(2) $s \ge \min\_\sup, c < \min\_conf$; It means that the rules have low accuracy and they need to be deleted.

(3) $s \ge \min\_\sup, c \ge \min\_conf, SRI < DRI$; It means that the rules have no value and they need to be deleted.

(4) $s \ge \min\_\sup, c \ge \min\_conf, 1 \ge SRI \ge DRI$; It means that the rules are valuable and they need to be reserved.

The algorithm of tendency mining in dynamic association rules is described as follows.

$f_{(A\cup B)_{ij}}$ stands for the i th element value of frequent item set l_j . $s_{(A\cup B)_{ij}}$ stands for the i th value of the support vector SV_j and SRI_j stands for the trend degree.

Input: $D, D_1 \sim D_n, \min\_\sup, \min\_conf, DRI$;

Output: The rule set R (CARs).

```
$(L, FV, s)$ =Dynamic-frequent-item-set-algorithm (FP-Growth)
for each frequent-item-set $l_j \in L$ {
gives $s_{(A \cup B)_{ij}}$ from $f_{(A \cup B)_{ij}}$ and then builds $SV_j$;
for i=1 to n
if $s_{(A \cup B)_{ij}} \geq \min\_\sup$ then $SRI_j = 1$;
for i=1 to n
if $s_{(A \cup B)_{ij}} \geq s_{(A \cup B)_{(i+1)j}}$ then $SRI_j = 1$;
for i=1 to n
if $s_{(A \cup B)_{ij}} < s_{(A \cup B)_{(i+1)j}}$ then $SRI_j = 1$;
    Calculating the function $\rho_1, \rho_2, \ldots, \rho_n$ of each element in $SV$;
if $\rho_l$ is close to 1 and $\rho_1, \rho_2, \ldots, \rho_{l-1}, \ldots, \rho_n$ is close to 0 then $SRI_j = 1$;
$SRI_j = \max(M, K) / n$;
$CRI_j = \max(M, K) / n$;
/* $M$ stands for the length of the biggest rise of the child support vector sequence and $K$ stands for the length of the biggest decline of the child support vector sequence in $FV$ */
}
callIntGenRule( $L, SRI, \min\_conf, DRI$ );
// Get the tendency association rules.
Produce callIntGenRule( $L, SRI, \min\_conf, DRI$ ){
$(R, c)$ =rule-generation-sub-algorithm
( $L, SRI, \min\_conf, DRI$ );
for each rule $r_i \in R$ do
Get $c_{(A \cup B)_{ij}}$ from $s_{(A \cup B)_{ij}} / s_{A_i}$ and then build $CV_j$;
Return;}
```

3.2. The SVM Classifier

The SVM classifier [10] based on compatibility feature vector needs to be constructed on the basis of class association rule sets. In the process, it should weight all the classification association rules according to certain strategy and then calculate their compatibility with the original data set to produce a feature vector collection. In this way, a pattern can be represented with a feature vector and a rule in CARs can be corresponded by a feature. The SVM classifier is constructed by a feature vector. This paper adds the tendency threshold to the method of weighting to construct the SVM classifier based on compatibility feature vector to improve classification method. Here are the rules of score metrics which can indicate the importance of a rule. So it can decide the weight of a rule to present its identification capability in the SVM classifier. The rules weighted formula is as follows.

$$W(R_i^C) = (R_i^C.c \times R_i^C.s \times R_i^C.CRI) / d_i^C$$

In the formula, R_i^C stands for the rule whose label is C and d_i^C , its distribution of measures in the training set as is the number of rules whose label is C. In the process of weighting rules, the distribution of data set for each category should be calculated at first and then stored in d. And then according to the rules' weight, each rule with the weight is stored in a new list called w. The next step is to build the compatibility feature vector. In a distributed learning system, the feature vector is a key attribute to describe a data set. A feature vector is like $f_1 = v_1 \cap f_2 = v_2 \cap \cdots \cap f_n = v_n$ in which f_i represents a feature with its value v_i.

The compatibility of distribution of the class association rules in the original training set can be obtained by building new feature vectors to get the compatibility measurements between rules and patterns. A feature in the vector can describe the compatibility between a pattern vector and a class association rule. The number of features equal to the number of rules in w. In the given model, a pattern compatible with a rule must meet the following conditions.

(1) Their class labels are the same;

(2) The compatibility between them is above 0.

Setting the pattern is a continuous attributes space with the dimension of n and it includes m patterns with the same label in the training set in which c stands for the number of labels. The compatibility between the pattern $x_p = (x_{p1}, x_{p2}, \ldots, x_{pn})$ and the rule R_i in w can be calculated by the following formula.

$$\mu_i(x_p) = \mu_{i1}(x_{p1}) \times \cdots \times \mu_{in}(x_{pn}); p = 1, 2, \ldots, m$$

In this, $x_{pn} (1 \leq p \leq m)$ stands for the continuous attributes in the original database and $\mu_{ik}(x_{px})$ stands for the membership of the rules R_i with x_{pn}. So the new way of calculating the $< f_i, v_i >$ can be described as follows.

When a rule and a pattern are compatible, then $f_p^{R_i} = w_i \times \mu_i(x_p)$ with under other conditions, $f_p^{R_i} = 0$. $f_p^{R_i}$ stands for the characteristic value of the pattern x_p of rules and w_i stands for the weight of rules. So the $< f_i, v_i >$ can be rewritten as $< x_p^{R_i}, f_p^{R_i} >$.

For each pattern x_p included in D, first its compatibility with the rules in w is calculated. Then the new characteristics are added to the feature vector to create a feature vector FV_p with its description as follows.

$$FV_p = classC_p :< x_p^{R_1}, f_p^{R_1} >< x_p^{R_2}, f_p^{R_2} > \cdots < x_p^{R_n}, f_p^{R_n} >$$

At last, it uses the feature vectors to form a feature vector set $FV_s = \{FV_p \mid p = 1, 2, \ldots, m\}$.

When the feature vectors have been built, the next step is to construct the SVM classifier. In the traditional classification algorithm, the purpose is to reduce the number of rules as far as possible. But it affects classification accuracy. The SVM algorithm has advantages when dealing with problem with high complexity. So it can be helpful to use as many rules to generate a classifier without effects on the results. While in reality, the labels of actual instance are unknown. The classifier is useful to predict the test cases with known labels.

4. EXPERIMENTAL RESULTS

To test and verify the classification results of the proposed method, the paper selected six data sets in UCI as the experimental objects. The algorithm carries out classification through a 10-cross validation method [11]. The relevant attribute information of the six data sets is shown in the Table **1** below. The experiment is run in a computer with the windows XP operating system and the programming language is C#. The FP-Growth algorithm is used in generating CARs.

Table 1. Data sets.

DS	Records	Attributes	Labels
Glass	214	11	7
Ionosphere	351	35	2
Iris	150	5	3
papeBlocks	5473	11	5
Pima	768	9	2
waveform	5000	22	3

In the case of the same experiment environment and data, the method is compared with the classic associative classification algorithm (CBA and CMAR). Firstly, it used the weighted rules to pruning rules and the results are shown in Table **2**. It set the confidence, support and the trend degrees with the values 1%, 50% and 50%.

Table 2. Pruning results.

DS	CBA	CMAR	TMCAR-SVM
Glass	64	59	9
Ionosphere	204	38	29
Iris	6	8	15
papeBlolcks	580	657	146
Pima	206	177	41
waveform	975	1040	350

As is shown in Table **3**, the method is better than the two algorithms. Due to that, the CBA algorithm sorts the rules based on confidence and its classifier is based on a single rule matching. The method sorts the rules both based on confidence and support. So it is more efficient than the CBA. It establishes a pruning strategy more effective than the CMAR by considering the compatibility between rules and patterns. So the results are more accurate than the CMAR.

Table 3. Accuracy.

DS	Accuracy (%)		
	CBA	CMAR	TMCAR-SVM
Glass	70.2	72.4	95.8
Ionosphere	42.0	89.3	91.6
Iris	95.7	94.6	92.0
pape Blocks	89.4	88.0	95.5
Pima	73.2	76.9	94.6
waveform	80.5	79.2	99.3

In the process of experiments, it is found that the change of the support frequency is too random to get effective decision-making information. It can generate fewer rules in uniform distribution data set. But the result is not ideal when the data set is unbalanced while it is better for large data sets like papeBlokcks and waveform. The limitation of DRI and confidence threshold would affect the results a lot. But by improving the confidence threshold, the accuracy of the results is not improved. It may produce invaluable decision rules due to high degree of confidence. So the next study should consider both the confidence and support threshold to obtain a better classifier.

CONCLUSION

This paper presents a method of tendency mining in dynamic association rules based on SVM classifier based on compatible feature vector. First of all, it classifies the dynamic association rules according to the tendency mining method. And secondly, by improving the method of weighting rules to build the classifier, it describes the relevant algorithm. By comparing with the traditional mining algorithm with a sample application, it has been proven that the method can effectively improve the quality of dynamic association rule mining and the mining efficiency and accuracy. The algorithm of classification accuracy is affected by the classification frequency and the confidence. So in the future, the need is to improve the algorithm under the action of the confidence and support at the same time. But in fact, it is always pruning away too many valuable rules for users due to the limitation of tendency threshold. So it is necessary to consider to lighten the weight of trend to build the classifier.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (No.61163010), the Scientific and Technological support projects in Gansu Province

(1104GKCA016) and Natural Science Foundation of Gansu Province (1308RJZA194).

REFERENCES

[1] B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining", *Proceeding of the 4th International Conference on Knowledge Discovery and Data Mining*. New York: AAAI Press, 1998, pp.80-86.

[2] L. Wenmin, H. Jiawei, and P. Jian. "CMAR: Accurate and efficient classification based on multiple class-association rule". ICDM 2001. San Jose, California: IEEE Computer Society, 2001, pp. 369-376.

[3] Q. Yang, and X. Wu, "10 Challenging problems in data mining research", *International Journal of Information Technology & Decision Making*, vol. 5, pp. 597-604, 2006.

[4] A. Maria-Luiza, O. R. Zaiane, "Associative classifier based on positive and negative rules", Paris, France: *Proceeding of 9th ACM SIGMOD Workshop on Research Issues in Data Minging and Knowledge Discovery*, 2004, pp. 64-69.

[5] J. Wang, K. G. Harmony, Efficient mining the best rules for classification [EB/OL]. [2010-05-10]. Available from: http://www.siam.org/proceed ings/datamining/2005/dm05_19wangj.pdf

[6] G. P. Baralis, "A lazy approach to pruning classification rules", *Proceeding of IEEE 2002 International Conference on Data Mining*. Washington, DC: IEEE Press, 2002, pp.35-42.

[7] Z. Zhonglin, Z. Qinfei, and X. Fan, "Method of tendency mining in dynamic association rules", *Computer Application*, vol. 32, pp. 196-198, 2012.

[8] Z. Shanwen, L. Yingjie, and F. Youjie. "Matlab in the application of time series analysis". Xian: Xian University of Electronic Science and Technology Press, 2007, pp.4-13.

[9] J.-W. Han, and M. Kamber. Data mining technology and the concept. Fan Ming, Meng Xiaofeng, translation. Beijing: China Machine Press, 2007, pp. 321-325.

[10] L. Qin, and Z. Xindong, "Research on frenquent pattern list based associative classifier construction algorithm", *Computer Applications and Software*, vol. 28, pp. 39-42, 2011.

[11] C. Jian, L. Qiang, and L. Yong, "Method of SVM classifier generation based on fuzzy classification association rule", *Computer Applications*, vol. 31, pp. 1348-135, 2011.

Received: July 25, 2014 Revised: August 4, 2014 Accepted: August 4, 2014