Class Association Rule Mining with Multiple Imbalanced Attributes

Huaifeng Zhang, Yanchang Zhao, Longbing Cao, and Chengqi Zhang

Faculty of IT, University of Technology, Sydney, Australia
PO Box 123, Broadway, 2007, NSW, Australia
{hfzhang, yczhao, lbcao, chengqi}@it.uts.edu.au

Abstract. In this paper, we propose a novel framework to deal with data imbalance in class association rule mining. In each class association rule, the right-hand is a target class while the left-hand may contain one or more attributes. This framework is focused on the multiple imbalanced attributes on the left-hand. In the proposed framework, the rules with and without imbalanced attributes are processed in parallel. The rules without imbalanced attributes are mined through standard algorithm while the rules with imbalanced attributes are mined based on new defined measurements. Through simple transformation, these measurements can be in a uniform space so that only a few parameters need to be specified by user. In the case study, the proposed algorithm is applied into social security field. Although some attributes are severely imbalanced, the rules with minority of the imbalanced attributes have been mined efficiently.

1 Introduction

Data imbalance is often encountered in data mining especially classification and association rule mining. As integration of classification and association rule, class association rule [3] always suffers from data imbalance problem. In class association rule, the right-hand side is a predefined target class while the left-hand side can be single or multiple attributes. Data imbalance on either side of the class association rule can cause severe problem.

Recently, there are some researchers working on the data imbalance in class association rule mining. In 2003, Gu et al. [2] proposed an algorithm to deal with imbalanced class distribution in association rule mining. They defined a set of criteria to measure the interestingness of the association rules. Arunasalam and Chawla [1] studied the data imbalance in association rule mining. These algorithms are focused on the data imbalance of target class to improve the performance of so-called associative classifier [3].

In 1999, Liu et al. proposed MSApriori algorithm [4] to deal with rare item problem. Rare item problem is essentially the data imbalance problem on transactional dataset. In MSApriori, the authors defined Minimum Item Support (MIS) to apply multiple minimum supports in the algorithm. However, MIS has

M.A. Orgun and J. Thornton (Eds.): AI 2007, LNAI 4830, pp. 827–831, 2007.

to be assigned to every item by user. Moreover, the discovered rules vary depending on the MIS values. Yun et al. [5] proposed an algorithm to mine association rule on significant rare data. In their algorithm, a number of supports also have to be specified by user. The performance of the algorithm heavily depends on the specified supports.

In this paper, we propose a novel algorithm to mine class association rules on the dataset with multiple imbalanced attributes. The rules with and without imbalanced attributes are processed in parallel. With new defined interestingness measurements and simple transformation, the rules can be post-processed in a uniform space. In the case study, this algorithm is applied into social security area. Many more rules with minorities of the imbalanced attributes have been mined, which is very interesting to the business line.

The paper is organized as follows. Section 2 introduces class association rule. Section 3 proposes the outline and the new measurements of proposed algorithm. Section 4 presents a case study. Section 5 is the conclusion. The last section is the acknowledgement.

2 Class Association Rules

Let T be a set of tuples. Each tuple follows the schema $(A_1, A_2, \ldots, A_N, A_C)$, in which $(A_1, A_2, \ldots, A_N)$ are N attributes while A_C is a special attribute, the target class. The attributes may be either categorical or continuous ones. For continuous attributes, the value range is discretized into intervals. For the convenience of description, we call an attribute-value pair an *item*. Suppose itemset $U \subseteq A$, A is the itemset of any items with attributes $(A_1, A_2, \ldots, A_N)$, c is 1-itemset of class attribute, a class association rule can be represented as

$$U \Rightarrow c$$

Here, U may contain a single item or multiple items.

In this paper, we represent the class association rules as

$$X + I \Rightarrow c$$

where, $X \subseteq A$ is the itemset of balanced attributes while $I \subseteq A$ is the itemset of imbalanced attributes.

3 Mining Rules with Imbalanced Attributes

Outline of the Algorithm. In our algorithm, association rule mining is done through two parallel parts. In one part, no imbalanced attributes are involved, and standard Apriori algorithm is used to mine interesting rules. In the other part, the imbalanced attributes are mined on sub-datasets to achieve high efficiency. The following is the detailed procedure of the proposed framework.

1. Standard association rule mining on the balanced attributes. In the original dataset, the imbalanced attributes are excluded and all of the tuples are kept for association rule mining.
2. Filter the original dataset to obtain the tuples containing minority part of one imbalanced attribute. In this step, only a small portion of the dataset is kept.
3. Mine the association rules on the filtered dataset using predefined minimum *confidence* and minimum *conditional support*. For every imbalanced attribute, repeat Step 2 and Step 3.
4. Transform the measurements into a uniform space and put all of the mined rules together. The final class association rule list is selected based on a set of criteria.

Conditional Support. In order to mine the rules including imbalanced attributes, we extend the definition of *support* since the minority of an imbalanced attribute normally occurs in a small portion of the tuples. In this paper, *conditional support* is defined to measure the interestingness of the rules with imbalanced attributes. If a class association rule is $X + I_m \Rightarrow c$ and I_m is the minority part of one imbalanced attribute m, the *conditional support* of this rule is,

$$Supp_c = \frac{P(X \cup I_m \cup c)}{P(I_m)}$$

where $P()$ stands for the probability of the itemsets occurring, X is an itemset of balanced attributes, and I_m is a 1-itemset of imbalanced attribute m, c is a class ID. Note that X, I_m and c are all itemsets rather than tuple sets so that $X \cup I_m \cup c$ means X, I_m and c occur simultaniously.

Confidence and Lift. Suppose the original dataset is represented as T. The subset T_m consists of the tuples containing the minority of imbalanced attribute m. If an association rule is $X + I_m \Rightarrow c$, it is not difficult to prove that the confidence on original dataset T and subset T_m are same. The confidience of this rule is

$$Conf = Conf' = \frac{P(X \cup I_m \cup c)}{P(X \cup I_m)} \tag{1}$$

However, on T and T_m, the expected confidences are different, which are

$$Conf_E = P(c) \tag{2}$$

$$Conf'_E = P'(c) = \frac{P(I_m \cup c)}{P(I_m)} \tag{3}$$

Hence the lifts obtained on T and T_m are different. In order to use uniform criteria to select the rules, the lift on T_m has to be transformed. On the original

dataset T, the expected confidence with respect to c is known, which is $P(c)$. On T_m, the confidence can also be obtained. So we may transform the lift obtained from subset T_m.

$$L_{new} = \frac{Conf'}{Conf_E} = \frac{Conf'}{P(c)} \tag{4}$$

So far the confidence, lift and the conditional support of all the rules are on the same base. We can use a uniform criteria to select the final rules. In this paper, minimum confidence, minimum lift and minimum conditional support are used to select the final class rule list.

4 Case Study

Our proposed algorithm has been tested with real-world data in Centrelink, Australia, which is an Australian government agency delivering a range of Commonwealth services to Australian community.

4.1 Problem Statement and Datasets

When a customer has received public monies to which he or she was not entitled, those funds may become a debt to be recovered. The purpose of data mining in debt recovery is to profile the customers according to the speed at which they pay off such debts. From a technical point of view, the objective is to mine the association rule with respect to the demographic attributes and debt information of a customer, the arrangement, and the target classes. Note that an arrangement is an agreement between a customer and Centrelink officer on the method, amount and frequency of repayments.

There are three kinds of datasets used for the association rule mining task: customer demographic data, debt data and repayment data. The class IDs, which are defined by business experts, are included in the repayment dataset. In the involved three datasets, there are three attributes having imbalanced distributions. For privacy reason, the three imbalanced attributes are denoted as "A", "B" and "C" respectively. The majorities of these attributes are all more than 90%.

4.2 Experimental Results

In the experiments, all the customers are grouped based on the arrangement patterns. Thus each arrangement is associated with a group of customers. Since conditional support rather than conventional support is used to generate the frequent itemset, the association rules including minorities of imbalanced attributes are mined without exhaustive searching. In Table 1, "A:1", "A:2", "B:1", "B:2" and "C:1" are all minorities of imbalanced attributes.

Table 1. Selected Rules with Imbalanced Attributes

Arrangement	Demographic Pattern	Class	$Conf_E(\%)$	Conf(%)	$Supp_c(\%)$	L_{new}	Count
W_W	Weekly:[$200, $400) & A:1 & GENDER:F	Class 1	39.0	48.6	6.7	1.2	52
C_A	MARITAL:SEP & A:1	Class 2	25.6	63.3	6.4	2.5	50
CI_A	Weekly:[$400, $600) & A:2	Class 3	35.4	64.9	6.4	1.8	50
WV_W	Children:0 & A:2 & MARITAL:SEP	Class 2	39.0	49.8	16.3	1.3	127
V_V	Weekly:0 & B:1 MARITAL:MAR &	Class 1	25.6	46.9	7.8	1.8	61
WV_WV	B:2 & GENDER:F	Class 3	25.6	49.7	11.4	1.9	89
WI_CW	Weekly:[$200, $400) & C:1 & GENDER:F	Class 3	39.0	45.7	18.8	1.2	147

5 Conclusions

This paper proposes an efficient algorithm to mine class association rules on the dataset with multiple imbalanced attributes. Unlike previous algorithms dealing with class imbalance, our algorithm processes the data imbalance on multiple attributes. Also different from the algorithms dealing with rare item problem, our algorithm employs a uniform selection criteria to discover the final combined association rule, which makes the algorithm more robust. The experimental results show the effectiveness of our proposed algorithm.

Acknowledgments

We would like to thank Mr. Fernando Figueiredo, Mrs. Leigh Galbrath, Mr. Shannon Marsh, Mr. Peter Newbigin, Mr. David Weldon, Ms. Michelle Holden and Mrs. Yvonne Morrow from Centrelink Australia for their support of domain knowledge and helpful comments.This work was supported by the Australian Research Council (ARC) Discovery Projects DP0449535, DP0667060 & DP0773412 and Linkage Project LP0775041.

References

1. Arunasalam, B., Chawla, S.: Cccs: a top-down associative classifier for imbalanced class distribution. In: KDD 2006, New York, NY, USA, pp. 517–522 (2006)
2. Gu, L., Li, J., He, H., Williams, G., Hawkins, S., Kelman, C.: Association rule discovery with unbalanced class distributions. In: Gedeon, T.D., Fung, L.C.C. (eds.) AI 2003. LNCS (LNAI), vol. 2903, pp. 221–232. Springer, Heidelberg (2003)
3. Liu, B., Hsu, W., Ma, Y.: Integrating classification and association rule mining. In: KDD 1998, New York, pp. 80–86 (August 27-31, 1998)
4. Liu, B., Hsu, W., Ma, Y.: Mining association rules with multiple minimum supports. In: KDD 1999, New York, USA, pp. 337–341 (1999)
5. Yun, H., Ha, D., Hwang, B., Ryu, K.H.: Mining association rules on significant rare data using relative support. Journal of Systems and Software 67(3), 181–191 (2003)