Combined Association Rule Mining

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

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

Abstract. This paper proposes an algorithm to discover novel association rules, combined association rules. Compared with conventional association rule, this combined association rule allows users to perform actions directly. Combined association rules are always organized as rule sets, each of which is composed of a number of single combined association rules. These single rules consist of non-actionable attributes, actionable attributes, and class attribute, with the rules in one set sharing the same non-actionable attributes. Thus, for a group of objects having the same non-actionable attributes, the actions corresponding to a preferred class can be performed directly. However, standard association rule mining algorithms encounter many difficulties when applied to combined association rule mining, and hence new algorithms have to be developed for combined association rule mining. In this paper, we will focus on rule generation and interestingness measures in combined association rule mining. In rule generation, the frequent itemsets are discovered among itemset groups to improve efficiency. New interestingness measures are defined to discover more actionable knowledge. In the case study, the proposed algorithm is applied into the field of social security. The combined association rule provides much greater actionable knowledge to business owners and users.

1 Introduction

Association rule mining aims to discover relationships among data in huge database. These relationships may provide some clues for business users to perform actions. In recent years, researchers [6,9] have focused on discovering more actionable knowledge. However, conventional association rules can only provide limited knowledge for potential actions. For example, in the Customer Relationship Management (CRM) field, association rule mining can be used to prevent churning. One possible rule is "$Demo : D \Rightarrow Churning$ ". With this rule, business users may take some actions on the customers with "$Demo : D$" to prevent churning. However, from the mined rule, business users cannot get knowledge on what action should be taken to retain these customers, though there might be many candidate actions.

We have previously defined combined association rule [11] to mine actionable knowledge. However, in [11], all of the attributes are treated equally when finding the frequent itemsets. The algorithm is time-consuming when a large number of

T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 1069–1074, 2008.

attributes are in database. In this paper, we differentiate the attributes and find the frequent itemsets on groups of itemsets. Furthermore, since data imbalance is often encountered in data mining tasks, we will also tackle data imbalance problem in combined association rule mining.

The paper is organized as follows. Section 2 gives the definition of combined association rule and its characteristics. Section 3 proposes the interestingness measures and algorithm outline. Section 4 introduces a case study. Section 5 presents some related work. Section 6 is the summary of this paper.

2 Definition of Combined Association Rule

Let T be a dataset. In this dataset, each tuple is described by a schema $S = (S_{D1}, \ldots, S_{Dm}, S_{A1}, \ldots, S_{An}, S_C)$, in which $S_D = (S_{D1}, S_{D2}, \ldots, S_{Dm})$ are m non-actionable attributes, $S_A = (S_{A1}, S_{A2}, \ldots, S_{An})$ are n actionable attributes, and S_C is a class attribute. Note that the data for combined association rule is not limited to one dataset. In fact, different kinds of attributes are often from multiple datasets [11].

Combined association rule mining is to discover the association among the 'attribute-value' pairs. For the convenience of description, we call an 'attribute-value' pair an '*item*'. Suppose itemset $D \subseteq I_D$, I_D is the itemset of any items with attributes $(S_{D1}, S_{D2}, \ldots, S_{Dm})$, itemset $A \subseteq I_A$, I_A is the itemset of any items with attributes $(S_{A1}, S_{A2}, \ldots, S_{An})$, C is 1-itemset of class attribute, a combined association rule set is represented as

$$\begin{cases} D + A_1 \Rightarrow C_{k1} \\ \quad \vdots \\ D + A_i \Rightarrow C_{ki} \end{cases} \tag{1}$$

Here, "+" means itemsets appearing simutaniouly. Since one action may result in different classes while one class may correspond to different actions, $C_{k1} \ldots C_{ki}$ rather than $C_1 \ldots C_i$ are used in Eq. 1.

3 Combined Association Rule Mining

In order to make the combined association rules in a rule set containing the same non-actionable itemset, it is important to firstly discover frequent non-actionable itemsets. Once these itemsets are discovered, the relationships of frequent non-actionable itemsets with target classes and actionable attributes are mined. In the rule generation step, the conditional support [10] is employed to tackle data imbalance problem.

3.1 Interestingness Measures

For a single combined association rule $D + A_i \Rightarrow C_{ki}$, the conventional interestingness measures are its confidence and lift. However, these two interestingness

measures are not sufficient to mine actionable knowledge from combined association rule. We illustrate this problem using an example. For a discovered frequent pattern $D + A_i \Rightarrow C_{ki}$, suppose $Conf(D + A_i \Rightarrow C_{ki})$ is 60% and the expected confidence of C_{ki} is 30%. So the lift of this frequent pattern is 2, which is high enough in most association rule mining algorithms. However the confidence of $D \Rightarrow C_{ki}$ is 70%, which means objects with non-actionable attribute D have 70% probability to be class C_{ki}. On the other hand, if action A_i happens, objects with non-actionable attribute D only have 60% probability to be class C_{ki}. Obviously action A_i is negatively correlated to class C_{ki} with respect to non-actionable itemset D.

Hence, a new lift named conditional lift is defined as follows to measure the interestingness of a combined association rule.

$$ConLift = \frac{Conf(D + A_i \Rightarrow C_{ki})}{Conf(D \Rightarrow C_{ki})} = \frac{Count(D \cap A_i \cap C_{ki}) \cdot Count(D)}{Count(D \cap A_i) \cdot Count(D \cap C_{ki})} \quad (2)$$

where $ConLift$ stands for the conditional lift of combined association rule $D + A_i \Rightarrow C_{ki}$. $Count(\times)$ is the count of the tuples containing itemset "$\times$". Note that D, A_i,and C_{ki} are all itemsets so that $D \cap A_i \cap C_{ki}$ means D,A_i, and C_{ki} occur simultaneously.

Briefly, Eq. 2 is the lift of $D + A_i \Rightarrow C_{ki}$ with D as a pre-condition, which shows how much is the contribution of A_i in the rule.

3.2 Algorithm Outline

The combined association rule mining procedure in this paper consists of two steps. The first step is to find single rule composed of frequent itemsets. The second step is to extract interesting combined association rule sets. Since itemsets are treated as different groups, the time comlexity of the algorithm is much lower than searching in the whole space of itemsets. In order to calculate the interestingness measures, the support count of each frequent itemset is recorded in the frequent itemset generation step. The outline of combined association rule mining is shown as follows:

1. Discovering frequent non-actionable itemsets I_D and the corresponding support counts C_D;
2. For each frequent non-actionable itemsets I_D
3. Finding frequent itemsets including target class I_{DC};
4. Recording the support count C_{DC} for each I_{DC};
5. Calculating conditional support $ConSup(DC)$;
6. If $(ConSup(DC) > MinSup)$, for each I_{DC}
7. Finding candidate pattern of three kinds of itemsets I_{DCA};
8. Recording the support count C_{DCA} for each I_{DCA};
9. Calculating conditional support: $ConSup(DA)$;
10. Calculating $Conf$, $Lift$ and $ConLift$;
11. If $(Conf \geq min_c$ & $Lift \geq min_l$ & $ConLift \geq min_{cl})$
12. Adding the mined frequent itemsets to the rule set.

4 Case Study

Our proposed technique has been tested with real-world data in Centrelink, which is an Australian Government Service Deliver Agency delivering a range of Commonwealth services to the Australian public.

4.1 Business Background and Problem Statement

When customers receive Commonwealth payments to which they were not entitled, these payments become customer debt that must be recovered. The purpose of data mining in debt recovery is trying to make the customers to repay their debts in a shortened timeframe according to historical debt recovery data and customer demographics. From a technical point of view, the objective is to mine the combined association rule with respect to the demographic attributes and debt information of customers, the arrangement, and the target class. Suppose some customers with similar demographic attributes and debt information belong to different target classes under different arrangements, Centrelink will recommend an arrangement to assist them to pay off a debt in the shortest possible time. Note that an arrangement is an agreement between a customer and Centrelink officer on the method, amount and frequency of repayment.

4.2 Data Involved

The dataset used for the combined association rule mining is composed of customer demographic data, debt data and repayment data. The customer demographic data includes customer ID, gender, age, marital status, salary, and so on. The debt data includes the debt related information. The repayment data includes the debt recovery arrangement, debt repayment amount and debt repayment date. The class ID of each customer is defined by business experts based on the information in debt data and arrangement data.

4.3 Experimental Results

In our experiment, the frequent patterns of the demographic itemsets were first mined using standard Apriori algorithm [1] on demographic data. The $Conf$, $Lift$ and $ConLift$ can be calculated on each frequent itemset. In the experiments, we set $minconf = 0.45$, $minlift = 1.2$, and $minconlift = 1.2$. Using these parameters and the calculated interestingness measures, the interesting combined association rule sets are selected. In this case study, 28 rule sets are discovered, which include 111 single rules algother. Selected results are shown in Table 1. For privacy reason, the benefit type, arrangement pattern and class ID are recoded in the experiments.

With the mined combined association rules, much actionable knowledge can be obtained. For example, suppose the priority of target class in this experiment is $C_2 > C_1 > C_3$, if a customer is with demographic attributes $MARITAL : SIN$ & $Age : 26y - 50y$ & $Earnings$: [\$200, \$400), the arrangement A_2 will be recommended to him/her with the greatest priority. If A_2 is impossible, A_1 will be recommended. The arrangement A_{10} is recommended with the least priority.

Table 1. Selected results of combined association rules

Demographics	Arg	Class	$Conf$	$Lift$	$ConLift$	$IsRule$
BENType:AAA & MARITAL:MAR & Age:65y+	A_1	C_1	0.46	1.31	1.70	Yes
BENType:AAA & MARITAL:MAR & Age:65y+	A_2	C_2	0.92	2.01	1.61	Yes
BENType:AAA & MARITAL:MAR & Age:65y+	A_3	C_2	0.91	1.97	1.58	Yes
MARITAL:SIN & Age:26y-50y & Earnings:[$200, $400)	A_1	C_1	0.78	2.20	1.83	Yes
MARITAL:SIN & Age:26y-50y & Earnings:[$200, $400)	A_4	C_1	0.29	0.83	0.69	No
MARITAL:SIN & Age:26y-50y & Earnings:[$200, $400)	A_{11}	C_1	0.50	1.42	1.18	No
MARITAL:SIN & Age:26y-50y & Earnings:[$200, $400)	A_2	C_2	0.79	1.72	2.15	Yes
MARITAL:SIN & Age:26y-50y & Earnings:[$200, $400)	A_7	C_3	0.42	2.27	2.04	No
MARITAL:SIN & Age:26y-50y & Earnings:[$200, $400)	A_{10}	C_3	0.46	2.47	2.23	Yes
BENType:BBB & Earnings:0 & Children:0	A_5	C_1	0.77	1.67	2.97	Yes
BENType:BBB & Earnings:0 & Children:0	A_7	C_3	0.64	3.44	1.47	Yes
BENType:BBB & Earnings:0 & Children:0	A_8	C_3	0.50	2.70	1.16	No

5 Related Work

The work in this paper is obviously different from any previous association rule mining algorithms. Hilderman et al. [4] extended simple association rule to mine characterized itemsets. Employing the concept of "share measures", their algorithm may present more information in terms of financial analysis. Different from Hilderman et al.'s algorithm, each single rule in this paper is associated with a target class to provide ordered action list.

Ras et al. [7,8] proposed to mine action rules. They divided the attributes in a database into two groups: stable ones and flexible ones. The action rules are extracted from a decision table given preference to flexible attributes. However, in their algorithm, only flexible attributes appear in the mined rules.

In combined association rule mining, each single combined association rule is similar to class association rule (CAR), which was proposed by Liu et al. [5] in 1998. However, in [5], the class association rules are mined to build associative classifier while the combined association rule sets are mined for direct actions rather than prediction.

Data imbalance problem has attracted much research attention in recent years. Arunasalam and Chawla [2] studied the data imbalance in association rule mining. Their algorithm is focused on the imbalanced distribution of one attribute, the target class. In our algorithm, the data imbalance problem occurs not only on target class but also actionable attributes.

6 Summary

This paper proposes an efficient algorithm to mine combined association rules on imbalanced datasets. Unlike conventional association rules, our combined association rules are organized as a number of rule sets. In each rule set, single combined association rules consist of different kinds of attributes. A novel frequent pattern generation algorithm is proposed to discover the complex inter-rule and intra-rule relationships. Data imbalance problem is also tackled in this paper.

The proposed algorithm is tested in a real world application. The experimental results show the effectiveness of algorithm.

Acknowledgments

We would like to thank Mr. Hans Bohlscheid, Business Manager and Project Manager of Centrelink Business Integrity and Information Division, for his ongoing support, and Mr. Fernando Figueiredo and Mr. Peter Newbigin for their assistance in extracting Centrelink data.

This work was supported by Discovery Projects DP0449535, DP0667060, DP0773412, Linkage Project LP0775041 from Australian Research Council (ARC) and Early Career Rsearch Grants from University of Technology, Sydney (UTS).

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: VLDB 1994, Santiago de Chile, Chile, pp. 487–499 (1994)
2. Arunasalam, B., Chawla, S.: Cccs: a top-down associative classifier for imbalanced class distribution. In: KDD 2006, Philadelphia, PA, USA, pp. 517–522 (2006)
3. 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)
4. Hilderman, R.J., Carter, C.L., Hamilton, H.J., Cercone, N.: Mining market basket data using share measures and characterized itemsets. In: Wu, X., Kotagiri, R., Korb, K.B. (eds.) PAKDD 1998. LNCS, vol. 1394, pp. 159–170. Springer, Heidelberg (1998)
5. Liu, B., Hsu, W., Ma, Y.: Integrating classification and association rule mining. In: KDD 1998, New York, NY, USA, pp. 80–86 (1998)
6. Liu, B., Hsu, W., Ma, Y.: Identifying non-actionable association rules. In: KDD 2001, San Francisco, CA, USA, pp. 329–334 (2001)
7. Ras, Z.W., Alicja, W.: Action-rules: How to increase profit of a company. In: Zighed, A.D.A., Komorowski, J., Żytkow, J.M. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 587–592. Springer, Heidelberg (2000)
8. Ras, Z.W., Ras, Z.W., Tzacheva, A.A., Tsay, L.-S., Giirdal, O.: Mining for interesting action rules. In: Tzacheva, A.A. (ed.) IAT 2005, pp. 187–193 (2005)
9. Yang, Q., Yin, J., Ling, C., Pan, R.: Extracting actionable knowledge from decision trees. IEEE TKDE 19(1), 43–56 (2007)
10. Zhang, H., Zhao, Y., Cao, L., Zhang, C.: Class association rule mining with multiple imbalanced attributes. In: Orgun, M.A., Thornton, J. (eds.) AI 2007. LNCS (LNAI), vol. 4830, pp. 582–587. Springer, Heidelberg (2007)
11. Zhao, Y., Zhang, H., Figueiredo, F., Cao, L., Zhang, C.: Mining for combined association rules on multiple datasets. In: Proceedings of the KDD 2007 Workshop on Domain Driven Data Mining, San Jose, CA, USA, pp. 18–23 (2007)