Mining for Combined Association Rules on Multiple Datasets *

Yanchang Zhao
Faculty of IT, University of Technology, Sydney, Australia
yczhao@it.uts.edu.au

Huaifeng Zhang
Faculty of IT, University of Technology, Sydney, Australia
hfzhang@it.uts.edu.au

Fernando Figueiredo
Centrelink, Australia
fernando.figueiredo@centrelink.gov.au

Longbing Cao
Faculty of IT, University of Technology, Sydney, Australia
lbcao@it.uts.edu.au

Chengqi Zhang
Faculty of IT, University of Technology, Sydney, Australia
chengqi@it.uts.edu.au

ABSTRACT

Many organisations have their digital information stored in a distributed systems structure scheme, be it in different locations, using vertically and horizontally distributed repositories, which brings about an high level of complexity to data mining. From a classical data mining view, where the algorithms expect a denormalised structure to be able to operate on, heterogeneous data sources, such as static demographic and dynamic transactional data are to be manipulated and integrated to the extent commercial association rules algorithms can be applied. Bearing in mind the usefulness and understandability of the application from a business perspective, combined rules of multiple patterns derived from different repositories, containing historical and point in time data, were used to produce new techniques in association mining applied to debt recovery. Initially debt repayment patterns were discovered using transactional data and class labels defined by domain expertise, then demographic patterns were attached to each of the class labels. After combining the patterns, two type of rules were discovered leading to different results: 1) same demographic pattern with different repayment patterns, and 2) same repayment pattern with different demographic patterns. The rules produced are interesting, valuable, complete and understandable, which shows the applicability and effectiveness of the new method.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining, association rules

*This work was supported by the Australian Research Council (ARC) Discovery Projects DP0449535, DP0667060 & DP0773412 and Linkage Project LP0775041.

2007 ACM SIGKDD Workshop on Domain Driven Data Mining (DDDM2007), August 12, 2007, San Jose, California, USA.

Keywords

Combined association rules, combined patterns

1. INTRODUCTION

Business data is often distributed amongst different databases, relational tables, files, systems and/or geographic locations. Mining this type of data structure, to extract business insight, is difficult and subject to ongoing research, because the existent algorithms work on denormalised file structures, either on a single flat file or table. Algorithm scalability issues concerned with computing time and memory space (e.g. joining tables in memory) can be prohibitory expensive, also privacy and integrity issues play an important role. Heterogeneous data sources, such as demographic and transactional data, are part of everyday business applications and used for data mining research. Traditional data mining algorithms are not applied, directly, to the above data structures. From a business perspective, patterns extracted from a single normalised table or subject file are less interesting or useful than a full set of multiple patterns extracted from different datasets. For example: Which customers, with the same demographic pattern, having different repayment patterns are then classified as quick vs slow payers? Which customers, with the same repayment pattern, having different demographic pattern are then classified as quick vs slow payers? Most traditional data mining techniques focus on extracting a single pattern, either demographic patterns for quick vs fast payers or transactional repayment patterns. Note the difference between mixed data types, consisting of numeric, categorical or ordinal variables and heterogeneous data sources. Heterogeneous data sources means, data sourced from different systems, database types, same database type but different tables, subject tables, historical tables, aggregated data, etc. Transactional, demographic and time series data are examples of heterogeneous data sources, requiring a different data mining technique or algorithm to be applied to extract the necessary knowledge. On the other hand, the data may be distributed, adding another layer of integrity, time and resources complexity, not to mention privacy issues.

A new technique has been designed to discover combined rules on multiple databases and applied to debt recovery in the social security domain. The first step is to discover

and group customer repayment patterns by arrangement types from transactional data. Customers are then classified as quick/moderate/slow payers based on a combination of domain knowledge and group distribution of payoff timeframe. Based on the above classification, demographic data is used to further refine and classify the customer as quick/moderate/slow payer. This method links demographic patterns from static demographic data to arrangement - repayment patterns from transactional data. The last step is to extract association rules based on the above patterns: customers, with the same repayment pattern, having different demographic patterns and customers, with the same demographic pattern, having different repayment patterns. The rules produced are useful, understandable and interesting from a business perspective, which shows the effectiveness of the proposed method.

The paper is organized as follows. Section 2 briefly introduces some related work on distributed data mining and multi-relational data mining. The problem to be addressed and its business background are introduced in Section 3. The proposed idea and framework to mine the combined rules is described in details in Section 4. Section 5 gives experimental results. Some discussion is presented in Section 6 and conclusions are given in Section 7.

2. RELATED WORK

Some related works on multi-relational data mining are [2, 3, 4, 5, 6] and some recent research on distributed data mining are [1, 7, 8, 9]. Park and Kargupta presented an overview of distributed data mining algorithms systems and applications in [8]. Jensen and Soparkar proposed to exploit the inter-table foreign key relationships to obtain decentralized algorithms that execute concurrently on the separate tables and then merge the results [6]. Guo and Viktor proposed an idea of multi-view classification, with which multiple classifiers are constructed in each view and then combined by a meta-learning algorithm [5]. Kargupta et al proposed a framework of collective data mining to conduct distributed data mining from heterogeneous sites [7]. Domingos argues that multi-relational data mining plays a key role in KDD [3].Cristofor and Simovici designed a couple of algorithms to address the problem for mining association rules in databases consisting of multiple tables and designed using the entity-relationship model [2]. Chattratichat et al designed a Kensington software architecture for distributed enterprise data mining, which addresses the problem of data mining on logical and physical distribution of data and heterogeneous computational resources [1]. Provost argued that distributed data mining is "a more natural way to view data mining generally" and "eliminates many difficulties encountered when coalescing already-distributed data for monolithic data mining" [9].

3. BUSINESS PROBLEM, BACKGROUND AND DATA

In this section, the business background and problem will be introduced first, and then the data related to the problem will be described.

3.1 Business Problem and Background

Centrelink, an Australian government agency, delivers a range of government payments and services for retirees, families, carers, parents, people with disabilities, Indigenous Australians and people from diverse cultural and linguistic backgrounds. Centrelink also provides services during times of major change. Centrelink put in practice research activities, for the purpose to analyse and inform strategies for improving (and measuring the improvement in) business integrity outcomes, including strategies aimed at improving overall levels of compliance, reducing fraud, errors, debts, overpayment to customers and improving non-payment outcomes. Customer debt can occur when changes of customer circumstances are not properly advised or processed to Centrelink. There are payability and qualifications giving entitlement to payments from the government over a period of time or until such events preclude the customer from obtaining the benefit. For example, in a carer-caree relationship where the carer receives a payment from the government for the purpose of looking after the caree, the caree passes away and the carer does not advise Centrelink of the event, Centrelink may continue to pay the caree until such time as the event is notified, therefore a debt is raised for the amount equivalent to the time the customer was not entitled to payment. After the debt is raised, the customer is notified of the debt amount and recovery procedures are initiated to recover that debt amount. If the customer cannot repay the total amount in full, a repayment arrangement is worked out between the parties. The purpose of this project is to present management with customers, profiled according to their capacity to pay off their debts in shortened timeframes. This enables management, to target those customers with recovery and amount options suitable to their own circumstances, and increase the frequency and level of repayment. Whether a customer is a quick or slow payer is believed by domain experts to be related to demographic circumstances, arrangements and repayments.

3.2 Data Involved

Three datasets containing current and non-current customers with debts were used: customer demographic data, debt data and repayment data. The first data contains demographic attributes of customers, such as customer ID, gender, age, marital status, number of children, declared wages, location and benefit. The second dataset contains debt related information, such as the date and time when a debt was raised, debt amount, debt reason, benefit or payment type the debt amount is related to , and so on. The repayments dataset contains arrangement types, repaying types, date and time of repayment repayment amount, repayment method (post office, direct debit, withholding payment), etc.

The demographic data is relatively static, however, the repayment data is dynamic and transactional. That is why it is difficult to organise both, into a single table for data mining. Also, different data mining techniques may be applied on them, making it difficult to combine the results discovered.

4. MINING COMBINED RULES ON MULTIPLE DATASETS

This section, first introduces the framework of combined patterns mining, followed by mining combined patterns in social security data and finally the detailed procedure is presented.

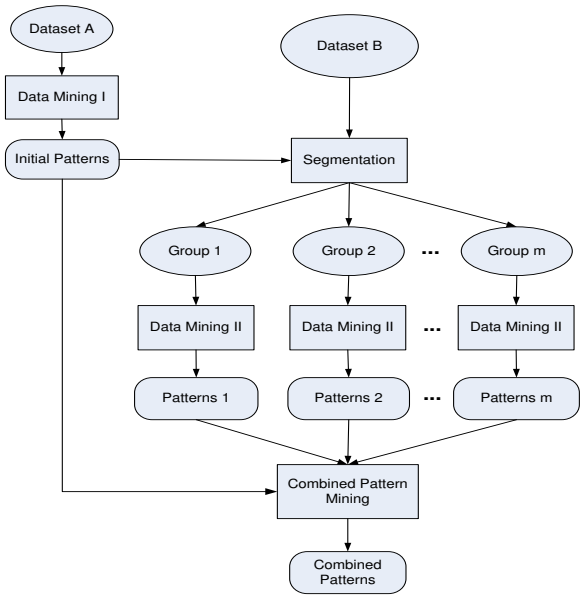

Figure 1: The Framework for Mining Combined Patterns.

4.1 The Framework of Combined Patterns Mining

Our methodology to find combined rules from distributed datasets and combined patterns follows: (see Figure 1).

1. Mining frequent patterns on dataset A. Let $\mathbb{P} = \{P_i\}$ $(i = 1, 2, ..., m)$ be the set of top m frequent patterns discovered. This step is shown as "Data Mining I" in Figure 1.

2. Based on the frequent patterns found, the dataset B is divided into groups. Each group is associated with a frequent pattern and a group of customers.

3. Mining association patterns in every group. Let $\mathbb{Q} = \{Q_j\}$ $(j = 1, 2, ..., n)$ be the set of top n frequent patterns discovered. This step is shown as "Data Mining II" in Figure 1.

4. Based on all the results discovered in the above step, find those rules like: 1) $P_1 + Q_1 \rightarrow R_1$ and $P_1 + Q_2 \rightarrow R_2$, and 2) $P_1 + Q_1 \rightarrow R_1$ and $P_2 + Q_1 \rightarrow R_3$, where R_i $(i = 1, 2, 3)$ stands for result. This step is shown as "Combined Pattern Mining" in Figure 1.

4.2 Mining Combined Patterns in Social Security Data

Quick/moderate/slow payers are defined based on time taken to repay the debt, the forecasted time to repay and the frequency/amount of repayment. There is a also a relationship between the capacity to pay (the arrangement) and total debt amount. The criterion is domain knowledge driven to the extent the expected time frame to repay the debt is not fixed, but has flexibility or fuzzy time limits. Therefore, the customer classification into quick/moderate/slow payer, has been done using domain knowledge, combining debt level amount, arrangement type and time to repay.

Table 1: Rules to Discover

Type of Rules	Arrangement & Repayment Patterns	Demographic Patterns	Class of Customers
Type A	Same	Different	Different
Type B	Different	Same	Different

The idea is first to derive the criterion of quick/slow payers from the data, and then propagate the tags of quick/slow payers to demographic data and to the other data to find frequent patterns and association rules. Since the pay-off timeframe is decided by the arrangement and repayment, customers are partitioned into groups according to their arrangement and repayment type. Second, pay-off timeframe distribution and statistics for each group are presented to domain knowledge experts, who then decide who are quick/slow payers by group. The criterion is applied to the data to tag every customer as quick/slow payer. Third, association rules are generated for quick/slow payers in each single group. Last, the association rules from all groups are organized together to build potentially business-interesting rules.

From an analysis perspective and addressing the business problem, we need to discover two types of rules (see Table 1). The first type (A), are rules with the same arrangement and repayment pattern but different demographic patterns leading to different customer classes (see Formula 1). For example, young people with a withholding arrangement in place, may be moderate payers, while mature age customers with the same type of arrangement are likely to be slow payers. The second type (B), are rules with the same demographic pattern but different arrangement and repayment pattern leading to different customer classes (see Formula 2). For example, mature age customers are more likely to be quick payers having a cash arrangement in place, but slow payers with a withholding arrangement in place.

$$\text{Type A: } \begin{cases} A_1 + D_1 & \rightarrow & \text{quick payer} \\ A_1 + D_2 & \rightarrow & \text{moderate payer} \\ A_1 + D_3 & \rightarrow & \text{slow payer} \end{cases} \quad (1)$$

$$\text{Type B: } \begin{cases} A_1 + D_1 & \rightarrow & \text{quick payer} \\ A_2 + D_1 & \rightarrow & \text{moderate payer} \\ A_3 + D_1 & \rightarrow & \text{slow payer} \end{cases} \quad (2)$$

where A_i and D_i denotes respectively arrangement patterns and demographic patterns.

4.3 Detailed Procedure

The procedure shown in Figure 2 depicts the interaction and flow of data between the domain experts and datasets. First, Customer ID (CRN), debt raise date and debt amount are extracted from debt data and are propagated to repayment data. Second, the repayment and arrangement patterns are generated from repayment data, and the distribution of pay-off timeframe for each arrangement pattern is derived from the debt database. The distributions are then given to business experts to decide who are the quick/moderate/slow payers based on the distribution and domain knowledge. Last, feedback received from the business experts is applied to the customer data, and then the association rules in customer group for each arrangement pattern are discovered, and combined patterns are gener-

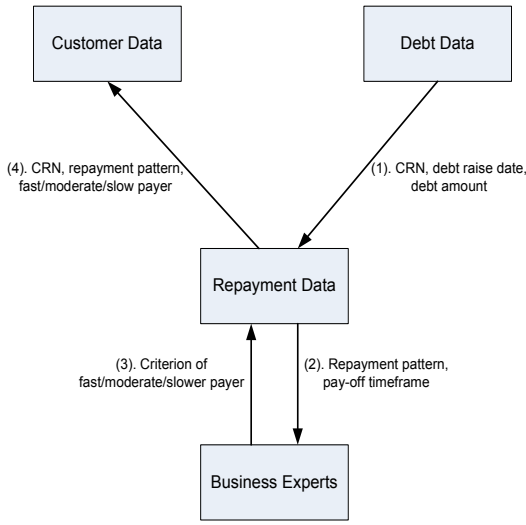

Figure 2: Procedure of Combined Pattern Mining in Social Security Data.

ated.

5. EXPERIMENTAL RESULTS

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

The data used are for debts raised in calendar year 2006 and the corresponding customers and repayments in the same year. Debts raised in calendar year 2006 are first selected, and then the customer data and repayment data in the same year related to the above debt data are extracted. Then the data is cleaned by removing noise and invalid values, for example, repayments with zero or negative amounts.

The cleaned data contains 479,288 customers with demographic attributes and 2,627,348 repayments. An arrangement is an agreement between a customer and Centrelink stating the method, amount and frequency of repayment. However, it may happen that a customer does not abide by or breaks the arrangement. That is, his repayment is not necessarily the same as agreed in his arrangement.

The combination of arrangement and repayment are found first, and the top combination of patterns based on the population are selected. Some examples of the arrangement-repayment patterns are cash, withholding, direct debit, cash plus withholding, etc. Customers associated with each pattern are put in one group, and each arrangement-repayment pattern is associated with a group of customers.

For each group of customers, the association rules of demographics and quick/moderate/slow payers are discovered using association mining algorithms. Some selected results are shown in Table 2. Note that the real benefit type codes are replaced with AAA, BBB or CCC in all tables in this paper for privacy preserving. Most of the rules show that "*Cash and Post Office repayments*" are associated with quick payer, "*Withholding plus Cash or Post Office repayments*" and "*Withholding plus Direct Debit repayments*" are associated with moderate payer, and "*Agent Recovery repayment*" is associated with slow payer. These rules show and agree with business knowledge acquired over time. However, there are some interesting rules like "*Benefit=BBB, Arrangement=Withholding and Irregular, Repayment= Withholding → Quick Payer*" with confidence of 64.9%, "*Weekly Income =[200,400), MARITAL= Single, Arrangement = Withholding and Irregular, Repayment =Withholding → Moderate Payer*" with confidence of 49.1%, "*Age=65y+, Arrangement =Withholding and Irregular, Repayment= Withholding → Slow Payer*" with confidence of 63.3%, and "*Weekly Income=0, Children Number=0, Arrangement =Withholding and Irregular, Repayment =Withholding → Slow Payer*" with confidence of 50.0%. The above rules shows that arrangement "*Arrangement=Withholding and Irregular, Repayment=Withholding*" can be applied to customers with BBB benefit, rather than to mature age people or those without any income or children.

Finally, the above discovered patterns are put together to find combined association rules. Selected combined association rules are given in Table 3 and 4. Table 3 shows some examples of rules with the same demographic characteristics. For those customers, different arrangements lead to different results. The table shows that male customers with CCC benefit repay their debts fastest with "*Arrangement=Cash, Repayment=Agent recovery*", while slowest with "*Arrangement=Withholding and Voluntary Deduction, Repayment= Withholding and Direct Debit*" or "*Arrangement=Cash and Irregular, Repayment=Cash or Post Office*". Therefore, for a male customer with a new debt, if his benefit type is CCC, Centrelink may try to encourage him to repay under "*Arrangement=Cash, Repayment=Agent recovery*", while try to persuade him not to pay under "*Arrangement=Withholding and Voluntary Deduction, Repayment=Withholding and Direct Debit*" or "*Arrangement =Cash and Irregular, Repayment=Cash or Post Office*", for the debt, will probably be repaid quickly.

Table 4 shows some examples of rules with the same arrangements but different demographic characteristics. The tables indicates that "*Arrangement=Withholding and Irregular, Repayment=Withholding*" arrangement is more appropriate for customers with BBB benefit, while they are not suitable for mature age customers, or those with no income or children. For young customers with a AAA benefit or single, it is not a bad choice suggesting to them, to repay their debts under "*Arrangement=Withholding and Irregular, Repayment=Withholding*".

6. DISCUSSION

The proposed framework can be used in similar data mining applications where multiple heterogeneous data are involved and combined patterns are preferred. The proposed method changes the complexity and difficulty of organising multiple heterogeneous datasets together, and generates combined rules which are more interesting and useful to business. It can also be used in applications where heterogeneous data requires different data mining techniques but the comprehensive results from all data are needed. For example, the repayment data is transactional and should be mined with a sequence mining technique, but the customer demographic data should be mined using decision trees or association rule mining techniques. Note that the results can be different depending on the order of mined datasets. For instance, the lifts vary with the order of datasets used in data mining. However, the confidence and count are always

Table 2: Selected Results

Arrangement	Repayment	Demographic Pattern	Result	Expected Con↓dence(%)	Con↓dence (%)	Support (%)	Lift	Count
Cash	Agent recovery	Marital:single & Gender:F & Bene↓t:AAA	Slow Payer	51.0	60.0	6.4	1.2	33
Cash & Irregular	Cash or Post O¡ce	Bene↓t:BBB	Quick Payer	40.8	67.0	4.9	1.6	61
Withholding & Irregular	Cash or Post O¡ce	Weekly:0 & Age:65y+	Quick Payer	72.4	88.7	8.9	1.2	110
Withholding & Irregular	Cash or Post O¡ce	Bene↓t:BBB	Quick Payer	72.4	88.4	8.7	1.2	107
Withholding & Irregular	Cash or Post O¡ce	Weekly:0 & Gender:M	Quick Payer	72.4	86.0	9.0	1.2	111
Withholding & Irregular	Cash or Post O¡ce	Age:65y+	Quick Payer	72.4	85.7	10.7	1.2	132
Withholding & Irregular	Cash or Post O¡ce	Children:0 & Bene↓t:AGE	Quick Payer	72.4	84.8	10.4	1.2	128
Withholding & Irregular	Cash or Post O¡ce	Weekly:0 & Gender:F & Children:0	Quick Payer	72.4	84.6	12.0	1.2	148
Withholding & Irregular	Withholding & Cash or Post O¡ce	Weekly:[$200, $400) & Marital:single & Bene↓t:AAA	Moderate Payer	60.4	82.5	4.4	1.4	104
Withholding & Irregular	Withholding & Cash or Post O¡ce	Marital:single & Gender:F & Children:0 & Bene↓t:AAA	Moderate Payer	60.4	81.6	4.9	1.4	115
Withholding & Irregular	Withholding & Cash or Post O¡ce	Marital:single & Bene↓t:AAA & Age:26y-50y	Moderate Payer	60.4	78.5	5.4	1.3	128
Withholding & Irregular	Withholding & Cash or Post O¡ce	Children:0 & Bene↓t:AAA & Age:26y-50y	Moderate Payer	60.4	78.2	9.3	1.3	219
Withholding & Irregular	Withholding & Cash or Post O¡ce	Gender:M & Bene↓t:AAA & Age:26y-50y	Moderate Payer	60.4	78.0	5.9	1.3	138
Withholding & Irregular	Withholding & Cash or Post O¡ce	Weekly:[$200, $400) & Children:0 & Bene↓t:AAA	Moderate Payer	60.4	77.8	7.4	1.3	175
Withholding & Irregular	Withholding & Cash or Post O¡ce	Weekly:[$200, $400) & Children:0 &Age:26y-50y	Moderate Payer	60.4	77.6	5.4	1.3	128
Withholding & Irregular	Withholding & Cash or Post O¡ce	Gender:F & Children:0 & Bene↓t:AAA	Moderate Payer	60.4	77.3	7.9	1.3	187
Withholding & Irregular	Withholding	Bene↓t:BBB	Quick Payer	35.4	64.9	6.4	1.8	50
Withholding & Irregular	Withholding	Age:65y+	Slow Payer	25.6	63.3	6.4	2.5	50
Withholding & Irregular	Withholding	Weekly:[$200, $400) & Marital:single	Moderate Payer	39.0	49.1	7.3	1.3	57
Withholding & Irregular	Withholding	Weekly:0 & Children:0	Slow Payer	25.6	50.0	11.4	1.9	89
Withholding & Voluntary Deduction	Withholding & Direct Debit	Weekly:[$200, $400) & Gender:M & Age:22y-25y	Moderate Payer	56.6	67.1	2.7	1.2	106
Withholding & Voluntary Deduction	Withholding & Direct Debit	Weekly:[$400, $600) & Marital:SEP & Age:26y-50y	Moderate Payer	56.6	65.4	2.6	1.2	100
Withholding & Voluntary Deduction	Withholding & Direct Debit	Gender:M & Bene↓t:AAA & Age:22y-25y	Moderate Payer	56.6	65.1	4.5	1.1	175

Table 3: Selected Results with the Same Demographic Patterns

Arrangement	Repayment	Demographic Pattern	Result	Confidence(%)	Count
Cash	Agent recovery	Gender:M & Benefit:CCC	Quick Payer	37.9	25
Withholding & Irregular	Withholding & Cash or Post Office	Gender:M & Benefit:CCC	Moderate Payer	75.2	100
Withholding & Voluntary Deduction	Withholding & Direct Debit	Gender:M & Benefit:CCC	Slow Payer	36.7	149
Cash & Irregular	Cash or Post Office	Gender:M & Benefit:CCC	Slow Payer	43.9	68
Withholding & Irregular	Cash or Post Office	Age:65y+	Quick Payer	85.7	132
Withholding & Irregular	Withholding & Cash or Post Office	Age:65y+	Moderate Payer	44.1	213
Withholding & Irregular	Withholding	Age:65y+	Slow Payer	63.3	50

Table 4: Selected Results with the Same Arrangement-Repayment Patterns

Arrangement	Repayment	Demographic Pattern	Result	Expected Confidence(%)	Confidence (%)	Support (%)	Lift	Count
Withholding & Irregular	Withholding	Age:17y-21y	Moderate Payer	39.0	48.6	6.7	1.2	52
Withholding & Irregular	Withholding	Age:65y+	Slow Payer	25.6	63.3	6.4	2.5	50
Withholding & Irregular	Withholding	Benefit:BBB	Quick Payer	35.4	64.9	6.4	1.8	50
Withholding & Irregular	Withholding	Benefit:AAA	Moderate Payer	39.0	49.8	16.3	1.3	127
Withholding & Irregular	Withholding	Marital:married & Children:0	Slow Payer	25.6	46.9	7.8	1.8	61
Withholding & Irregular	Withholding	Weekly:0 & Children:0	Slow Payer	25.6	49.7	11.4	1.9	89
Withholding & Irregular	Withholding	Marital:single	Moderate Payer	39.0	45.7	18.8	1.2	147

the same, which can be used to generate and find interesting combined patterns.

7. CONCLUSIONS

This paper shows a framework for mining combined association rules from multiple datasets, with heterogeneous datasets requiring different data mining techniques capable of producing comprehensive and useful rules. The above framework has been tested with real-world social security data and the results are interesting and help business to classify customers as quick/moderate/slow payers and their repayment patterns.

8. ACKNOWLEDGMENTS

We would like to thank Mrs. Leigh Galbrath, Mr. Shannon Marsh, Mr. Peter Newbigin and Mr. David Weldon from Centrelink Australia for their support of domain knowledge, and thank Mrs. Carol Ey, Ms. Michelle Holden and Mrs. Yvonne Morrow from the same organisation for their helpful comments and suggestions.

9. REFERENCES

[1] Jaturon Chattratichat, John Darlington, et al. An Architecture for Distributed Enterprise Data Mining. Proceedings of the 7th International Conference on High-Performance Computing and Networking, 1999.

[2] L. Cristofor and D. Simovici. Mining association rules in entity-relationship modeled databases. Technical report, University of Massachusetts Boston, 2001.

[3] P. Domingos. Prospects and challenges for multi-relational data mining. SIGKDD Explorations Newsletter, Volume 5, Issue 1, July 2003.

[4] Sašo Džeroski. Multi-relational data mining: an introduction. ACM SIGKDD Explorations Newsletter, Volume 5, Issue 1, July 2003.

[5] H. Guo and H. L. Viktor. 2005. Mining relational databases with multi-view learning. In Proceedings of the 4th international Workshop on Multi-Relational Mining (Chicago, Illinois, August 21 - 21, 2005). MRDM '05. ACM Press, New York, NY, 15-24.

[6] Viviane Crestana Jensen and Nandit Soparkar. Frequent Itemset Counting Across Multiple Tables. Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2000.

[7] H. Kargupta, B. Park, D. Hershbereger and E. Johnson. Collective data mining: A new perspective toward distributed data mining. Accepted in the Advances in Distributed Data Mining, Eds: Hillol Kargupta and Philip Chan, AAAI/MIT Press (1999).

[8] B. Park and H. Kargupta. Distributed Data Mining: Algorithms, Systems, and Applications. Data Mining Handbook, N. Ye, Ed., 2002.

[9] Foster Provost. Distributed data mining: Scaling up and beyond. In Advances in Distributed and Parallel Knowledge Discovery. MIT Press, 2000.