Discovering Debtor Patterns of Centrelink Customers

Yanchang Zhao[1], Longbing Cao[1], Yvonne Morrow[2], Yuming Ou[1], Jiarui Ni[1], and Chengqi Zhang[1]

[1] Faculty of Information Technology, University of Technology, Sydney
PO Box 123, Broadway, NSW 2007, Australia

{yczhao, lbcao, yuming, jiarui, chengqi}@it.uts.edu.au

[2] Business Integrity Strategy Branch, Centrelink, Australia
PO Box 312, Sunshine, VIC 3020, Australia

yvonne.y.morrow@centrelink.gov.au

Abstract

Data mining is currently becoming an increasingly hot research field, but a large gap still remains between the research of data mining and its application in real-world business. As one of the largest data users in Australia, Centrelink has huge volume of data in data warehouse and tapes. Based on the available data, Centrelink is seeking to find underlying patterns to be able to intervene earlier to prevent or minimize debt. To discover the debtor patterns of Centrelink customers and bridge the gap between data mining research and application, we have done a project on improving income reporting to discover the patterns of those customers who were or are in debt to Centrelink. Two data models were built respectively for demographic data and activity data, and decision tree and sequence mining were used respectively to discover demographic patterns and activity sequence patterns of debtors. The project produced some potentially interesting results, and paved the way for more data mining applications in Centrelink in near future.

Keywords: Data mining, decision tree, association rule, sequence mining

1 Introduction

Data mining is currently becoming an increasingly hot research field, but a large gap still remains between the research of data mining and its application in real-world business. As one of the largest data users in Australia, Centrelink has huge volume of data in data warehouse and tapes. Centrelink raised over $900 million worth of customer debts, excluding Child Care Benefits and Family Tax Benefits, in the year 2004-05 (Centrelink 2005). Moreover, Customer contact generates massive quantities of activity transactions. For example, Centrelink processed 5.2 billion transactions in the year 2004-2005 (Centrelink 2005). These transactions may contain important information related to both debt prevention and the achievement of Government Social Security objectives.

To discover the debtor patterns of Centrelink customers and bridge the gap between data mining research and application, we have done a project on improving income reporting to discover the patterns of those customers who were or are in debt to Centrelink. Two data models were built respectively for demographic data and activity data, and decision tree and sequence mining were used respectively to discover demographic patterns and activity sequence patterns of debtors.

We used the following analysis methods to discover the demographic characteristics and activity sequence patterns of debtors, which may provide information to help know who the debtors are, why the debts occur, and under what conditions debts has a high probability of occurring.

- decision tree for classification of debtor/non-debtor
- association mining for frequent customer circumstances
- sequence mining for activity sequence patterns

Two softwares, Teradata Warehouse Miner (TWM) (Teradata 2005) and Teradata SQL Assistant, were used for the above analysis. Most of our analytical models for mining debtor patterns were either directly based on the modules of TWM or on the improved SQL codes generated by TWM.

With the help of the above tools, we studied Centrelink data related to debtors, built data models and analytical models, and then produced some potentially interesting results, such as the demographic characteristics and the activity patterns of debtors. Our models and methodologies and the experience acquired in this project paved the way for more data mining applications in Centrelink in near future.

The rest of the paper is organised as follows. The business problem and the data available in this project are discussed in Section 2. Then we present the demographic data model, data mining model and results in Section 3. Section 4 describes the activity sequence data model, data mining model and results. The last section concludes the paper and discusses some future work.

2 Business Problem and Available Data

In this section, the business problem will be introduced and the available data will be described.

2.1 Business Problem

Centrelink is a government agency delivering a range of Commonwealth services to the Australian community. It distributes approximately $63 billion in social security payments on behalf of policy departments. The following are some statistics (Centrelink 2006). Centrelink:

- has 6.4 million customers, or approximately one-third of the Australian population,
- administers more than 140 different products and services for 25 government agencies, and
- pays 9.98 million individual entitlements each year and records 5.2 billion electronic customer transactions each year.

From the above figures, we can see that it has not only a very large population but also very large volume of customer data. Moreover, it has huge volume of transaction data which records its activities, which is shown by the following statistics. It

- has sent 87.2 million letters customers each year,
- receives more than 32.68 million telephone calls each year,
- receives 39.5 million website page views each year,
- grants 2.77 million new claims each year,
- conducts more than 98,700 Field Officer reviews each year, and
- has more than 650,000 booked office appointments each month.

Among the large population, there are some people purposely trying to maximize their payments from Centrelink by reporting less income or inadvertently claiming more benefits than they are entitled, which leads to a large amount of debts. The fraud statistics from Centrelink (Centrelink 2006) show that: between 1 July 2004 and 30 June 2005,

- Centrelink conducted 3.8 million entitlement reviews, which resulted in 525,247 payments being cancelled or reduced.
- Almost $43.2 million a week was saved and debts totaling $390.6 million were raised as a result of this review activity.
- Included in these figures were 55,331 reviews of customers from tip-offs received from the public, resulting in 10,022 payments being cancelled or reduced and debts and savings of $103.1 million.
- There were 3,446 convictions for welfare fraud involving $41.2 million in debts.

All the above figures show that debt prevention is a very important task for Centrelink. From the above statistics, we can see that about 14% out of all entitlement reviews resulted in debts, with a lot of money saved. However, 86% reviews resulted in nil debt. Therefore, much effort can be saved if we can target those customers who are debtors or are of high probability to have debts and then conduct reviews only on those customers. From the perspective of activity, if we can detect some activity sequence patterns are associated with debts, then something can be done in advance to prevent or reduce debts. Based on the above idea, we conducted a project to discover demographic characteristics and activity sequence patterns of debtors, expecting that the results may help to target specific customer groups or activity sequence patterns associated with high probability of debts. On the basis of the discovered patterns, more data mining work can be done on debt detection and then a debt prevention and/or reduction system can be built in near future.

2.2 Data Available

The Centrelink corporate database contains information about Centrelink customers and activities used to support them as well as activities to manage the government business. Centrelink delivers help and payments to customers on behalf of the Government and at the same time maintains tight controls over those payments to keep the integrity of the system accountable and transparent.

Newstart Allowance (NSA) is an allowance for those unemployed residents who are aged 21 or over and under Age Pension age. The population of this project is those customers who benefit from NSA. NSA is composed of up to four parts: Basic rate, RA (Rent assistance), PHA (Pharmaceutical allowance) and RAA (Remote area allowance). NSA customers can be classified into three groups: those with income debt, those with other debts and those without debt. This project concentrates on those NSA customers with income debt and analyses their characteristics and activities.

There are three kinds of data which relates to the above problem: customer data from Newsnap database, debt data from Debt database and activity data from AMHS database.

- Newsnap database: It keeps summary information of Newstart Sytems (NSS) cluster population, including personal information about the customer, such as gender, age, marital status, and indigenous indicator as well as address movements, income and migrant information.
- Debt database: It is used by debt management and contains debt details, co-debtors, recoverable debts and overpayments.
- Transaction database: It contains data about transactions and activity management before they are applied to the database. For the activities, what's available in mainframe is those activities after the beginning of 2006. Those data before that are stored in tapes. Therefore, only the activities from 1/1/2006 to 31/3/2006 are used in this project because of the time limit.

The debts under consideration are from 1/7/2004 to 28/2/2006 for debt file. The data from 1/1/2006 to 31/3/2006 in transaction database are considered. The customer data is a snapshot of Newsnap file on 30/6/2005. To compute the history summary, eg., address change times, the summary of the previous financial year is used.

All the three databases are internal databases stored in IBM mainframe, and can be accessed via SAS. The data related to NSA will be extracted from the above three databases and then loaded into Teradata warehouse.

Personal ID will be used to link the three databases together. From Newsnap, those whose benefit type is NSA

Table 1. Demographic data model

Fields	Notes
Customer current circumstances	These fields are from the current customer circumstances in customer data, which are personal id, indigenous code, medical condition, sex, age, birth country, migration status, education level, postcode, language, rent type, method of payment, etc.
Aggregation of debts	These fields are derived from debt data by aggregating the data in the past financial year (from 1/7/2004 to 30/06/2005), which are debt indicator, the number of debts, the sum of debt amount, the sum of debt duration, the percentage of a certain kind of debt reason, etc.
Aggregation of history circumstances	These fields are derived from customer data by aggregating the data in the past financial year (from 1/7/2004 to 30/06/2005), which are the number of address changes, the number of marital status changes, the sum of income, etc.

Table 2. Summary of demographic data

Time frame	# of NSA customers	# of NSA debtors*
Snapshot on 30/06/2005	498,597	189,663

* Number of NSA customers on 30/06/2005 who had NSA debts during the previous financial year (from 1/7/2004 to 30/06/2005)

will be extracted. With those personal IDs from NSA data, those related transactions and debt records will be extracted from transaction database and debt database, respectively.

3 Discovering Demographic Patterns of Debtors

This section will present the construction of a customer demographic data model based on customer data and debt data, the data mining model and the results.

3.1 Building Demographic Data Model

This data model is to organise customer circumstances data and debt information into one table (see **Table 1**), based on which the characteristics of debtors and non-debtors will be discovered. In this data model, each customer has one record, which shows its latest or aggregated information of customer circumstances and debt. There are three kinds of attributes in this data model: customer current circumstances, the aggregation of debts, and the aggregation of customer history circumstances (say, the number of address changes), which are shown in **Table 1**. Debt indicator is defined as a binary attribute which indicates whether a customer had any debts in the previous financial year. The summary of the built demographic data model is given in **Table 2**.

3.2 Feature Selection

There are over 80 features in the constructed demographic data model, which is too much for available data mining software due to the huge search space. The following methods were used to select features.

- Correlation: the correlation between variables and debt indicator (see **Table 3**),
- Chi-square: the contingency difference of variables to debt indicators (see **Table 4**),
- Data exploration based on statistics: the impact difference of a variable on debtors and non-debtors.

Chi-square analysis is used to find the relationship between debt indicator and customer circumstances. It is implemented in module "Test based on Contingency Tables" with TWM. In those modules, debt indicator is set as first column, while customer circumstances variables are set as second columns. The statistical test style is set to "Chi Square".

Based on correlation, chi-square test and data exploration, 15 features, such as ADDRESS_CHANGE_TIMES, RENT_AMOUNT, RENT_TYPE, CUSTOMER_SERVICE_CENTRE_CHANGE_TIMES and AGE, are selected as input for the following decision tree and association rule analysis.

3.3 Decision Tree Mining on Demographic Data

Based on the above feature selection, decision tree is used to build a classification model for debtors/non-debtors. It is implemented in TWM module "Decision Tree". In those modules, debt indicator is set to dependent column, while customer circumstances variables are set as independent columns. The best result we got is a tree of 676 nodes, and its accuracy is shown in **Table 5**, where "0" and "1" stand for "no debt" and "debt", respectively. The accuracy is poor (63.71%), and the error of false negative is high (30.53%). It is difficult to further improve the accuracy of decision tree on the whole population, however, some leaves of higher accuracy can be discovered by focusing on smaller groups. We found that the current version of

Table 3. Correlation between debt indicator and customer circumstances

Attributes	Correlation
CUSTOMER_SERVICE_CENTRE_CHANGE_IND	0.143
ADDRESS_CHANGE_TIMES	0.139
CUSTOMER_SERVICE_CENTRE _CHANGE_TIMES	0.135
ADDRESS_CHANGE_IND	0.129
RENT_AMOUNT	0.090
MARITAL_CHANGE_TIMES	0.085
MARITAL_CHANGE_IND	0.083
RA_ENTITLEMENT_AMOUNT	0.058
AGE	-0.092
LODGEMENT_FREQUENCY	-0.098

Table 4. Result of chi-square analysis for customer circumstances and debt indicator

Attributes	Chi-square
CUSTOMER_SERVICE_CENTRE _CHANGE_TIMES	13026
ADDRESS_CHANGE_TIMES	10889
LODGEMENT_FREQUENCY	6057
RENT_TYPE	4276
AGE	3940
RENT_AMOUNT	3903
MARITAL_CHANGE_TIMES	3745
RA_RATE_EXPLANATION	3424
RA_PRECLUSION_A13SON	3043
ACCOMMODATION	2924

Table 5. Confusion matrix of decision tree results

	Actual 0	Actual 1
Predicted 0	280,200 (56.20%)	152,229 (30.53%)
Predicted 1	28,734 (5.76%)	37,434 (7.51%)

Teradata warehouse miner cannot output leaves above a given accuracy threshold automatically and that it is difficult to navigate through the huge tree manually, we then turned to association rule mining to discover interesting patterns with higher accuracy on small groups of population.

3.4 Association Rule Mining on Demographic Data

Association mining (Agrawal, Imielinski, and Swami 1993) is used to find frequent customer circumstances patterns that are highly associated with debt or non-debt. It is implemented with “Association” module of TWM. In the module, personal ID is set as group column, while item_code is set as item column, where item_code is derived from customer circumstances and their values. In order to apply association rule analysis to our customer data, we regard each value of each feature as an item. Taking feature DEBT_IND as example, it has 2 values, which are DEBT_IND_0 and DEBT_IND_1. So DEBT_IND_0 is regarded as an item and DEBT_IND_1 is regarded as another item.

Table 6. Results of association rule mining

Association Rule	Support	Confidence	Lift
RA_RATE_EXPLANATION=P and age 21 to 28 =>debt	0.003	0.65	1.69
MARITAL_CHANGE_TIMES =2 and age 21 to 28 =>debt	0.004	0.60	1.57
age 21 to 28 and PARTNER_CASUAL_INCOME_SUM>0 and rent amount ranging from $200 to $400 => debt	0.003	0.65	1.70
MARITAL_CHANGE_TIMES =1 and PARTNER_CASUAL_INCOME_SUM>0 and HOME_OWNERSHIP=NHO => debt	0.004	0.65	1.69
age 21 to 28 and BAS_RATE_EXPLAN=PO and MARITAL_CHANGE_TIMES=1 and rent amount in $200 to $400 =>debt	0.003	0.65	1.71
CURRENT_OCCUPATION_STATUS=CDP => no debt	0.017	0.827	1.34
CURRENT_OCCUPATION_STATUS=CDP and SEX=male => no debt	0.013	0.851	1.38
HOME_OWNERSHIP=HOM and CUSTOMER_SERVICE_CENTRE_CHANGE_TIMES =0 and REGU_PAY_AMOUNT in $400 to $800 => no debt	0.011	0.810	1.31

Due to the limitation of spool space, we cannot run association rule analysis on the whole customer data. Therefore, we conduct association rule analysis on a 10% sample of the original data, and the discovered rules are then tested on the whole customer data. We select the top 15 features to run association rule analysis with minimum support as 0.003, and selected results are shown in **Table 6**. For example, the first rule in **Table 6** shows that 65% out of those customers with RA_RATE_EXPLANATION as "P" (Partnered) and aged from 21 to 28 have had debts in the previous financial year, and the lift of the rule is 1.69.

4 Discovering Activity Sequence Patterns of Debtors

The previous section describes our work on "static" demographic data, which tries to find the demographic patterns of debtors. In addition to the demographic characteristics, a debtor may have interesting activity patterns which show the interactions between him/her and Centrelink. This section will present the work on mining "dynamic" activity data, which is composed of customer activities sequences from 1/1/2006 to 31/3/2006.

An activity represents a unit of work. It records: 1) which customer is the activity related to (e.g. personal id), 2) the type of actions (e.g. activity code), 3) the date and the time the actions were done, 4) the reason for the actions (e.g. receipt of source documents and manual notes), etc. An activity can be caused by new claims or circumstance changes. It can also be caused by transaction of somebody else's. For example, when a customer is granted Newstart, an activity will be created to check the partner's record to ensure that the Family Tax Benefit is paid correctly. An activity can also be generated automatically by the system or manually registered. For example, in cases where maintenance action is in progress, automatic activities are generated to ensure the progress is reviewed.

An activity sequence is a series of actions for a customer, and mining such sequences may help to discover which sequences are highly associated with debts and which are not. In the following an activity sequence data model, data mining models and the results will be presented.

4.1 Building Activity Sequence Data Model

This data model organises activity data into baskets or sequences for association and sequence analysis and two

different kinds of baskets/sequences are built respectively for debt and non-debt. According to domain expert's opinion and the frequency of activity, the following strategy is used to build the baskets/sequences. For debt baskets/sequences, we look back one month before each debt to build up the basket/sequence for analyzing the activities. That is, if there is a debt of a customer, those activities of the customer happened in 30 days immediately before the debt are put into one basket. For debt basket, those debts beginning in January 2006 and their associated activities are excluded because the available activities before them are less than one month. For non-debt basket, those activities in January and February 2006 can be taken as a basket for a customer having no debts in the first 3 months of 2006. Those activities from 16/1/2006 to 15/2/2006 are used to build baskets for non-debt, because there are too few activities in the beginning days of the year.

An activity sequence is built for each debt of a person, and personal ID and debt ID are concatenated as the identifier of the sequence. This is to make the data ready for both association mining and sequence mining with TWM. The following are two examples of debt activity sequence and non-debt activity sequence, where A_i stands for an activity, and D and N denote "debt" and "no debt", respectively.

Debt activity sequence: A_1, A_2, A_3, A_5, A_6, A_7, D.

Non-debt activity sequence: A_4, A_6, A_3, A_2, A_5, A_9, A_1, N.

Table 7 summarizes the results of constructing debt and non-debt activity baskets/sequences. Activities are also reorganised in terms of income and non-income, where debt baskets are only built for all income-related debts, and all other types of debts are ignored while non-debt baskets are constructed for all persons not having a debt. This is based on that income-related debt is one of the concerns of this project, and also because of the imbalance of debt and non-debt classes of activities.

4.2 Sequence Mining Models and Results

The following data mining approaches are used to discover activity sequence patterns of debtors. Some examples of the discovered results of sequence patterns will follow. Note that the activity codes are replaced with A_i for the consideration of business confidentiality.

4.2.1 Sequence Mining on Activity Sequences on Mixed Data

Because the two classes, debt and non-debt, are highly unbalanced, it is difficult to find meaningful patterns from such data, and the measures of confidence and lift are skewed. Therefore, we draw a sample from non-debt data, which is of the same size as that of debt data, and conducted sequence mining on the sampled data. Then sequence mining was employed to discover those activity sequences co-occurring frequently with debt or non-debt. The support threshold is set to 0.01 and some mined results are given in **Table 8**.

4.2.2 Sequence Mining on Activity Sequences on Separated Data

To avoid the undesirable effect of unbalanced class sizes, we mined separately on debt activity sequences and non-debt activity sequences. The difference from mining on the mixed data is that this task discovers frequent activity patterns without worrying about debt or non-debt and that only support is computed for these patterns. Note that the support here is local support (Chen et al 2005), that is, the support of the pattern on debt data or non-debt data, not on the whole dataset. The support threshold is also set to 0.01.

4.2.3 Discovering Dual-Target and Contrast-Target Activity Patterns

For those activity patterns discovered on separated data, we look for those patterns which are frequent in both datasets, and those patterns which are frequent in one dataset but infrequent in the other. The former is named as *dual-target pattern*, while the latter is as *contrast-target pattern*. The way to find dual-target patterns is simply selecting those common frequent patterns in both debtor dataset and non-debtor dataset. The method to discover contrast-target patterns is computing the ratio of the supports of a pattern on the two datasets and then selecting those patterns with ratios much different from one.

Dual-target activity patterns given in **Table 9** are those frequent activity patterns identified in both debt group and non-debt group. For these activities and patterns, the local support indicates how frequently the activity/pattern occurs in each group. Activity A13 occurs in 86.3% of the debt records and in 69.4% of the non-debt records. For activity A17, a debt is raised within 30 days after A17 for 25.5% activity sequences, and no debt is raised within 30 days of activity A17 in 17.8% of activity sequences with A17. In this case, probably there are some other activities that make the sequences with A17 lead to different results. Further investigation is needed to tell what makes the sequences result in debt or non-debt. Obviously more research needs to be conducted into these activities. The inference is that an activity or sequence occurring in both debt and non-debt groups and strongly associated with debt is associated with missing information in the non-debt group. The potential in this area is for activities/sequences strongly associated with debt to act as a trigger for some kind of information where they occur in the non-debt group.

Contrast-target activity patterns are those activity associations or sequences much more frequent in debt group (or non-debt group) than in non-debt group (or debt group), which are shown in **Table 10**. For example, the first pattern "A13, A19" is 1.6 times more associated with debt than with no debt. For the activities of a customer, if A13 happens first and then A19 occurs, the customer should be checked carefully in case he may have a debt in the near future, or the activity sequences with "A13, A19" should be checked to make sure what can be done to prevent the occurrence of debt. These sequences, if confirmed by further research, can act as markers of debt or potential debt. Where they occur they could be a clear

trigger for targeted intervention strategies to prevent or minimize debt.

4.2.4 Discovering Reverse-Target Activity Patterns

For *reverse-target activity patterns*, we look for those frequent pattern pairs like "P=>debt" and "PQ=>no debt", or "P=>no debt" and "PQ=>debt", where P and Q are activities or activity sequences. In both cases, the occurrence of Q has a significant impact on the result, by changing the result to its opposite. This kind of patterns help to find which activity or sequence Q has significant impact on the result. For frequent pattern pairs like "P=>debt" and "PQ=>no debt", when activity sequence P happened, then activities in Q are suggested to conduct to reduce or prevent debt. For frequent pattern pairs like "P=>debt" and "PQ=>no debt", activities in Q are suggested not to conduct to reduce or prevent debt.

To measure the interestingness of the above patterns, we designed a measure of "impact" for pattern pair "P=>result1" and "PQ=>result2"as follows (see (Cao, Zhao, and Zhang 2006) for details).

$$\text{Impact} = \frac{Sup_2 / Sup_4}{Sup_3 / Sup_1},$$

where Sup_1 is the local support of an underlying pattern, e.g., "P=>result1", Sup_2 is the local support of a derivative pattern, e.g., "PQ=>result2", Sup_3 is the local support of the rule ("P=>result2") contrary to the underlying pattern, and Sup_4 is the local support of the rule ("PQ=>result1") contrary to the derivative pattern. Impact should be greater than 1.0 for useful patterns. The larger impact is, the more interesting is the pattern pair.

The top-ranking reverse-target activity patterns are given in **Table 11**. It is found that A1, A22, A23, A20 are not related to debts in our data, while A13, A14 or A15 is more likely to be associated with debt than non-debt. However, when A1, A22, A23 or A20 follow A13, A14 or A15, the composite patterns have higher likelihood of resulting in non-debt rather than debt.

4.2.5 Pruning Redundant Patterns

There are many redundant patterns in the discovered rules. Assume that "A=>C" and "AB=>C" are two rules of the same confidence, then "AB=>C" is redundant because it provides no more information given "A=>C" is known. The method is to remove those patterns if there are any shorter sub-patterns having the same or roughly the same confidence. For contrast-target patterns, those patterns whose support ratios are the same or less than the support ratios of their sub-patterns are removed. For reverse-target patterns, those pattern pairs whose impacts are the same or less than the impacts of their sub-patterns are removed. For more details on pruning redundant patterns, please refer to (Zaki 2004).

5 Concluding Remarks

Data mining techniques have been used to discover the debtor patterns of Centrelink customers. Two data models, customer demographic data model and activity sequence data model, were first built and then techniques of decision tree and sequence mining were employed to mine the demographic data and activity data. The discovered patterns maybe used to find those customer groups with high probability of debts, so that reviews on those customers can be conducted or letters can be mailed to them to help reduce debts. Moreover, by discovering activity sequence patterns associated with debt/non-debt, appropriate actions can be suggested for next step under a given situation to reduce the probability of leading to debt. This is one of our efforts to solve real-world business problems with advanced data mining techniques, and it shows promising applications of data mining to solve real-life problems in near future.

However, there are still many open problems to be solved. First, there are still hundreds or even thousands of discovered rules after redundant patterns have been pruned. How can interesting patterns be efficiently selected from them? Second, most rules obtained with existing statistical measures of interestingness are not interesting at all from business perspective, and many business interesting rules may be pruned during data mining procedure, so post-mining rules pruning helps little. The use of domain knowledge when mining can not only help to find "business interesting" patterns, but also help to reduce the search space and running time of data mining algorithms. How can domain knowledge be effectively incorporated in data mining procedure? Third, the business data is complicated and the customers and customer debts/activities are linked to many other customers, such as spouse, dependents and tenants. How can existing data mining approaches be improved to discover more useful patterns by utilizing those additional information? For example, an activity A1 of a customer C1 may lead to an activity A2 of his/her spouse C2, and A2 may activate activity A3 of a third customer C3. How can existing approaches for sequence mining be improved to take into consideration the linkage and interaction between activity sequences of different customers? Last and the most important, how to use these discovered rules to help predict and prevent debt? How to build an efficient debt prevention system to effectively detect debt in advance and give appropriate suggestions to help reduce or prevent debt? How to evaluate the risk of debt when an action is taken? All the above problems remain open and will be part of our future work.

6 Acknowledgments

The authors would like to thank Dr. Jie Chen, Mr. Peter Newbigin, Mr. Fernando Figueiredo, Mr. Rick Schurmann and Mr. Mark Norrie for their work and support. This work was supported in part by the Australian Research Council (ARC) Discovery Projects (DP0449535 and DP0667060), National Science Foundation of China (NSFC) (60496327) and Overseas Outstanding Talent Research Program of Chinese Academy of Sciences (06S3011S01).

Table 7. Summary of activity sequences

	# of activities	# of sequences	# of debt-related sequences
All debts	6,063,703	454,934	16,540 (3.6%)
Income debts	5,770,523	439,953	1,559 (0.35%)

Table 8. Results of activity sequence mining

Sequence Rule	Support	Confidence	Lift
A1 => no debt	0.0529	0.753	1.51
A2 => debt	0.0173	0.831	1.66
A3 => debt	0.0712	0.716	1.43
A4, A3 => debt	0.0157	0.845	1.69
A5, A3 => debt	0.0144	0.833	1.67
A6, A3 => debt	0.0141	0.815	1.63
A7, A4 => debt	0.0141	0.759	1.52
A8, A7 => debt	0.0388	0.738	1.48
A8, A8 => debt	0.0260	0.704	1.41
A8, A6 => debt	0.0173	0.692	1.38
A9, A10 => debt	0.0154	0.686	1.37
A11, A9 => debt	0.0138	0.682	1.36
A12, A9 => debt	0.0157	0.681	1.36
A8, A4 => debt	0.0209	0.677	1.35
A8, A8, A7 => debt	0.0138	0.860	1.72

Table 9. Dual-target activity sequence patterns

Activity	Local support of "activity=>debt"	Local support of "activity=>no debt"
A13	0.863	0.694
A14	0.845	0.601
A15	0.711	0.569
A16	0.322	0.257
A4	0.286	0.155
A17	0.255	0.178
A6	0.248	0.203
A8	0.226	0.155
A5	0.164	0.120
A12	0.162	0.138
A18	0.133	0.128

Table 10. Contrast-target activity patterns

Sequence patterns	DSUP: support of "pattern=>debt"	NSUP: support of "pattern =>no debt"	DSUP/NSUP
A3	0.142	0.053	2.7
A15, A3	0.094	0.029	3.2
A13, A19	0.033	0.013	2.6
A14, A19	0.028	0.011	2.6
A8, A4	0.042	0.017	2.6
A19, A14	0.026	0.010	2.4
A14, A4, A4	0.042	0.015	2.9
A4, A4, A14	0.038	0.014	2.7
A8, A7, A13	0.051	0.018	2.7
A8, A7, A14	0.055	0.020	2.7
A8, A8, A7	0.028	0.010	2.7
A14, A4, A14	0.126	0.047	2.7
A13, A4, A5	0.028	0.011	2.6
A20	0.026	0.035	0.8
A1	0.035	0.093	0.4

Table 11. Reverse-target activity patterns

Sequence patterns	Impact	SUP1, e.g., support of A1=> no debt	SUP2, e.g., support of A1,A14 =>debt	SUP3, e.g., support of A1=>debt	SUP4, e.g. Support of A1,A14=>no debt
A1 => no debt A1, A14=> debt	3.73	0.093	0.017	0.035	0.013
A1 => no debt A1, A15 => debt	2.93	0.093	0.019	0.035	0.017
A13 => debt A13, A1 => no debt	2.44	0.863	0.024	0.694	0.012
A15, A14 => debt A15, A14, A22 => no debt	2.00	0.527	0.015	0.340	0.012
A14, A15 => debt A14, A15, A22 => no debt	1.87	0.499	0.017	0.344	0.014
A13, A17 => debt A13, A17, A18 => no debt	1.64	0.166	0.012	0.108	0.012
A13, A17 => debt A13, A17, A21 => no debt	1.58	0.166	0.011	0.108	0.010

7 References

Agrawal, R., Imielinski, T. and Swami, A. (1993): Mining Association Rules between Sets of Items in Large Databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Washington DC, USA, 22:207-216, ACM Press.

Cao, L., Zhao, Y. and Zhang, C. (2006): Mining Impact-Targeted Activity Patterns in Unbalanced Data, submitted to IEEE TKDE special issue on Intelligence and Security Informatics, 30 April 2006.

Centrelink (2005): Centrelink Annual Report 2004-2005.

Centrelink (2006): Centrelink Fraud Statistics and Centrelink Facts and Figures, http://www.centrelink.gov.au/internet/internet.nsf/about_us/fraud_stats.htm, http://www.centrelink.gov.au/internet/internet.nsf/about_us/facts.htm.

Chen, J., He, H., Li, J., Jin, H., McAullay, D., Williams, G., Sparks, R. and Kelman C. (2005): Representing Association Classification Rules Mined from Health Data. *Proc. of 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2005)*, Melbourne, Australia, September 14-16, 2005, pp. 1225-1231.

Teradata (2005): Teradata Warehouse Miner User's Guide - Release 04.01.00, 2005.

Zaki, M. (2004): Mining non-redundant association rules. *Data Mining and Knowledge Discovery*, 9:223-248, 2004.