Chapter 6
Data Mining Applications in Social Security

Yanchang Zhao, Huaifeng Zhang, Longbing Cao, Hans Bohlscheid, Yuming Ou, and Chengqi Zhang

Abstract This chapter presents four applications of data mining in social security. The first is an application of decision tree and association rules to find the demographic patterns of customers. Sequence mining is used in the second application to find activity sequence patterns related to debt occurrence. In the third application, combined association rules are mined from heterogeneous data sources to discover patterns of slow payers and quick payers. In the last application, clustering and analysis of variance are employed to check the effectiveness of a new policy.

Key words: Data mining, decision tree, association rules, sequential patterns, clustering, analysis of variance.

6.1 Introduction and Background

Data mining is becoming an increasingly hot research field, but a large gap remains between the research of data mining and its application in real-world business. In this chapter we present four applications of data mining which we conducted in Centrelink, a Commonwealth government agency delivering a range of welfare services to the Australian community. Data mining in Centrelink involved the application of techniques such as decision trees, association rules, sequential patterns and combined association rules. Statistical methods such as the chi-square test and analysis of variance were also employed. The data used included demographic data, transactional data and time series data and we were confronted with problems

Yanchang Zhao, Huaifeng Zhang, Longbing Cao, Yuming Ou, Chengqi Zhang
Faculty of Engineering and Information Technology, University of Technology, Sydney, Australia, e-mail: {yczhao,hfzhang,lbcao,yuming,chengqi}@it.uts.edu.au

Hans Bohlscheid
Data Mining Section, Business Integrity Programs Branch, Centrelink, Australia, e-mail: hans.bohlscheid@centrelink.gov.au

such as imbalanced data, business interestingness, rule pruning and multi-relational data. Some related work include association rule mining [1], sequential pattern mining [13], decision trees [16], clustering [10], interestingness measures [15], redundancy removing [17], mining imbalanced data [11,19], emerging patterns [8], multi-relational data mining [5–7, 9] and distributed data mining [4, 12, 14].

Centrelink is one of the largest data users in Australia, distributing approximately $63 billion annually in social security payments to 6.4 million customers. Centrelink administers in excess of 140 different products and services on behalf of 25 Commonwealth government agencies, making 9.98 million individual entitlement payments and recording 5.2 billion electronic customer transactions each year [3]. These statistics reveal not only a very large population, but also a significant volume of customer data. Centrelink's already significant transactional database is further added to by its average yearly mailout of 87.2 million letters and the 32.68 million telephone calls, 39.5 million website hits, 2.77 million new claims, 98,700 field officer reviews and 7.8 million booked office appointments it deals with annually.

Qualification for payment of an entitlement is assessed against a customer's personal circumstances and if all criteria are met, payment will continue until such time as a change of circumstances precludes the customer from obtaining further benefit. However, customer debt may occur when changes of customer circumstances are not properly advised or processed to Centrelink. For example, in a carer/caree relationship, the carer may receive a Carer Allowance from Centrelink. Should the caree pass away and the carer not advise Centrelink of the event, Centrelink may continue to pay the Carer Allowance until such time as the event is notified or discovered through a random review process. Once notified or discovered, a debt is raised for the amount equivalent to the time period for which 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. If the customer cannot repay the total amount in full, a repayment arrangement is negotiated between the parties. The above debt prevention and recovery are two of the most important issues in Centrelink and are the target problems in our applications.

In this chapter we present four applications of data mining in the field of social security, with a focus on the debt related issues in Centrelink, an Australia Commonwealth agency. Section 6.2 describes the application of decision tree and association rules to find the demographic patterns of customers. Section 6.3 demonstrates an application of sequence mining techniques to find activity sequences related to debt occurrence. Section 6.4 presents combined association rule mining from heterogeneous data sources to discover patterns of slow payers and quick payers. Section 6.5 uses clustering and analysis of variance to check the effectiveness of a new policy. Conclusions and some discussion will be presented in the last section.

6.2 Case Study I: Discovering Debtor Demographic Patterns with Decision Tree and Association Rules

This section presents an application of decision tree and association rules to discover the demographic patterns of the customers who were in debt to Centrelink [20].

6.2.1 Business Problem and Data

For various reasons, customers on benefit payments or allowances sometimes get overpaid and these overpayments collectively lead to a large amount of debt owed to Centrelink. For example, Centrelink statistical data for the period 1 July 2004 to 30 June 2005 [3] shows that:

- 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 totalling $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; and
- there were 3,446 convictions for welfare fraud involving $41.2 million in debts.

The above figures indicate that debt detection is a very important task for Centrelink staff and we can see from the statistics examined that approximately 14 per cent of all entitlement reviews resulted in a customer debt. However, 86 per cent of reviews resulted in a NIL and therefore it becomes obvious that much effort can be saved by identifying and reviewing only those customers who display a high probability of having or acquiring a debt. Based on the above observation, this application of decision tree and association rules aimed to discover demographic characteristics of debtors; expecting that the results may help to target customer groups associated with a high probability of having a debt. On the basis of the discovered patterns, more data mining work could be done in the near future on developing debt detection and debt prevention systems.

Two kinds of data relate to the above problem: customer demographic data and customer debt data. The data used to tackle this problem have been extracted from Centrelink's database for the period 1/7/2004 to 30/6/2005 (financial year 2004-05).

6.2.2 Discovering Demographic Patterns of Debtors

Customer circumstances data and debt information is organized into one table, based on which the characteristics of debtors and non-debtors are discovered (see

Table 6.1 Demographic data model

Fields	Notes
Customer current circumstances	These fields are from the current customer circumstances in customer data, which are 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 6.2 Confusion matrix of decision tree result

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

Table 6.1). In the data model, each customer has one record, which shows the aggregated information of that customer's 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, for example, the number of address changes. Debt indicator is defined as a binary attribute which indicates whether a customer had debts in the financial year. In the built data model, there are 498,597 customers, of which 189,663 are debtors.

There are over 80 features in the constructed demographic data model, which proved to be too much for available data mining software to deal with due to the huge search space. The following methods were used to select features: 1) the correlation between variables and debt indicator; 2) the contingency difference of variables to debt indicators with chi-square test ; and 3) data exploration based on the impact difference of a variable on debtors and non-debtors. 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, were selected as input for decision tree and association rule mining.

Decision tree was first used to build a classification model for debtors/non-debtors. It was implemented in Teradata Warehouse Miner (TWM) module of “Decision Tree". In the module, debt indicator was set to dependent column, while customer circumstances variables were set as independent columns. The best result obtained is a tree of 676 nodes, and its accuracy is shown in Table 6.2, where “0" and “1" stand for “no debt" and “debt", respectively. However, 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 were discovered by focusing on smaller groups.

Association mining [1] was then used to find frequent customer circumstances patterns that were highly associated with debt or non-debt. It was implemented with “Association" module of TWM. In the module, personal ID was set as group column, while item-code was set as item column, where item-code is derived from

Table 6.3 Selected association rules

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

customer circumstances and their values. In order to apply association rule analysis to our customer data, we took each pair of feature and value as a single item. Taking feature DEBT-IND as example, it had 2 values, DEBT-IND-0 and DEBT-IND-1. So DEBT-IND-0 was regarded as an item and DEBT-IND-1 was regarded as another. Due to the limitation of spool space, we conducted association rule analysis on a 10 per cent sample of the original data, and the discovered rules were then tested on the whole customer data. We selected the top 15 features to run association rule analysis with minimum support as 0.003, and some selected results are shown in Table 6.3. For example, the first rule shows that 65 per cent of customers with RA-RATE-EXPLANATION as “P" (Partnered) and aged from 21 to 28 had debts in the financial year, and the lift of the rule was 1.69.

6.3 Case Study II: Sequential Pattern Mining to Find Activity Sequences of Debt Occurrence

This section presents an application of impact-targeted sequential pattern mining to find activity sequences of debt occurrence [2]. Impact-targeted activities specifically refer to those activities associated with or leading to specific impact of interest to business. The impact can be an event, a disaster, a government-customer debt, or any other interesting entities. This application was to find out which activities or activity sequences directly triggered or were closely associated with debt occurrence.

6.3.1 Impact-Targeted Activity Sequences

We designed impact-targeted activity patterns in three forms, *impact-oriented activity patterns*, *impact-contrasted activity patterns* and *impact-reversed activity patterns*.

Impact-Oriented Activity Patterns

Mining frequent debt-oriented activity patterns was used to find out which activity sequences were likely to lead to a debt or non-debt. An impact-oriented activity pattern is in the form of $P \rightarrow T$, where the left hand side P is a sequence of activities and the right side is always the target T, which can be a targeted activity, event or other types of business impact. Positive frequent impact-oriented activity patterns ($P \rightarrow T$, or $\bar{P} \rightarrow T$) refer to the patterns likely lead to the occurrence of the targeted impact, say leading to a debt, resulting from either an appeared pattern (P) or a disappeared pattern ($\bar{P}$). On the other hand, negative frequent impact-oriented activity patterns ($P \rightarrow \bar{T}$, or $\bar{P} \rightarrow \bar{T}$) indicate that the target unlikely occurs ($\bar{T}$), say leading to no debt.

Given an activity data set $D = D_T \bigcup D_{\bar{T}}$, where D_T consists of all activity sequences associated with targeted impact and $D_{\bar{T}}$ contains all activity sequences related to non-occurrence of the targeted impact. The count of debts (namely the count of sequences enclosing P) resulting from P in D is $Cnt_D(P)$. The *risk* of pattern $P \rightarrow T$ is defined as $Risk(P \rightarrow T) = \frac{Cost(P \rightarrow T)}{TotalCost(P)}$, where $Cost(P \rightarrow T)$ is the sum of the cost associated with $P \rightarrow T$ and $TotalCost(P)$ is the total cost associated with P. The *average cost* of pattern $P \rightarrow T$ is defined as $AvgCost(P \rightarrow T) = \frac{Cost(P \rightarrow T)}{Cnt(P \rightarrow T)}$.

Impact-Contrasted Activity Patterns

Impact-contrasted activity patterns are sequential patterns having contrasted impacts, and they can be in the following two forms.

- $Supp_{D_T}(P \rightarrow T)$ is high but $Supp_{D_{\bar{T}}}(P \rightarrow \bar{T})$ is low,
- $Supp_{D_T}(P \rightarrow T)$ is low but $Supp_{D_{\bar{T}}}(P \rightarrow \bar{T})$ is high.

We use FP_T to denote those frequent itemsets discovered in those impact-targeted sequences, while $FP_{\bar{T}}$ stands for those frequent itemsets discovered in non-target activity sequences. We define *impact-contrasted patterns* as $ICP_T = FP_T \backslash FP_{\bar{T}}$ and $ICP_{\bar{T}} = FP_{\bar{T}} \backslash FP_T$. The *class difference* of P in two datasets D_T and $D_{\bar{T}}$ is defined as $Cd_{T,\bar{T}}(P) = Supp_{D_T}(P \rightarrow T) - Supp_{D_{\bar{T}}}(P \rightarrow \bar{T})$. The *class difference ratio* of P in D_T and $D_{\bar{T}}$ is defined as $Cdr_{T,\bar{T}}(P) = \frac{Supp_{D_T}(P \rightarrow T)}{Supp_{D_{\bar{T}}}(P \rightarrow \bar{T})}$.

Impact-Reversed Activity Patterns

An impact-reversed activity pattern is composed of a pair of frequent patterns: an underlying frequent impact-targeted pattern 1: $P \rightarrow T$, and a derived activity pattern

2: $PQ \rightarrow \bar{T}$. Patterns 1 and 2 make a contrasted pattern pair, where the occurrence of Q directly results in the reversal of the impact of activity sequences. We call such activity patterns as *impact-reversed activity patterns*. Another scenario of impact-reversed activity pattern mining is the reversal from negative impact-targeted activity pattern $P \rightarrow \bar{T}$ to positive impact $PQ \rightarrow T$ after joining with a trigger activity or activity sequence Q.

To measure the significance of Q leading to impact reversal from positive to negative or vice versa, a metric *conditional impact ratio* (*Cir*) is defined as $Cir(Q\bar{T}|P) = \frac{Prob(Q\bar{T}|P)}{Prob(Q|P) \times Prob(\bar{T}|P)}$. *Cir* measures the statistical probability of activity sequence Q leading to non-debt given pattern P happens in activity set D. Another metric is *conditional Piatetsky-Shapiro's ratio* (*Cps*), which is defined as $Cps(Q\bar{T}|P) = Prob(Q\bar{T}|P) - Prob(Q|P) \times Prob(\bar{T}|P)$.

6.3.2 Experimental Results

The data used in this case study was Centrelink activity data from 1 January 2006 to 31 March 2006. Extracted activity data included 15,932,832 activity records recording government-customer contacts with 495,891 customers, which lead to 30,546 debts in the first three months of 2006. For customers who incurred a debt between 1 February 2006 and 31 March 2006, the activity sequences were built by putting all activities in one month immediately before the debt occurrence. The activities used for building non-debt baskets and sequences were activities from 16 January 2006 to 15 February 2006 for customers having no debts in the first three months of 2006. The date of the virtual non-debt event in a non-debt activity sequence was set to the latest date in the sequence. After the above activity sequence construction, 454,934 sequences were built, out of which 16,540 (3.6 per cent) activity sequences were associated with debts and 438,394 (96.4 per cent) sequences with non-debt. T and $\bar{T}$ denote debt and non-debt respectively, and a_i represents an activity.

Table 6.4 shows some selected impact-oriented activity patterns discovered. The first three rules, $a_1, a_2 \rightarrow T$, $a_3, a_1 \rightarrow T$ and $a_1, a_4 \rightarrow T$ have high *confidences* and *lifts* but low *supports* (caused by class imbalance). They are interesting to business because their *confidences* and *lifts* are high and their *supports* and *AvgAmts* are not too low. The third rule $a_1, a_4 \rightarrow T$ is the most interesting because it has $risk_{amt}$ as high as 0.424, which means that it accounts for 42.4% of total amount of debts.

Table 6.5 presents some examples of impact-contrasted sequential patterns discovered. Pattern "a_{14}, a_{14}, a_4" has $Cdr_{T,\bar{T}}(P)$ as 4.04, which means that it is 3 times more likely to lead to debt than non-debt. Its $risk_{amt}$ shows that it appears before 41.5% of all debts. According to *AvgAmt* and *AvgDur*, the debts related to the second pattern a_8 have both large average amount (26789 cents) and long duration (9.9 days). Its $Cdr_{T,\bar{T}}(P)$ shows that it is triple likely associated with debt than non-debt.

Table 6.6 shows an excerpt of impact-reversed sequential activity patterns. One is *underlying pattern* $P \rightarrow$ *Impact 1*, the other is *derived pattern* $PQ \rightarrow$ *Impact 2*,

Table 6.4 Selected impact-oriented activity patterns

Patterns	$Supp_D(P)$	$Supp_D(T)$	$Supp_D(P \rightarrow T)$	Confidence	Lift	*AvgAmt*	*AvgDur*	$risk_{amt}$	$risk_{dur}$
$P \rightarrow T$						(cents)	(days)		
$a_1, a_2 \rightarrow T$	0.0015	0.0364	0.0011	0.7040	19.4	22074	1.7	0.034	0.007
$a_3, a_1 \rightarrow T$	0.0018	0.0364	0.0011	0.6222	17.1	22872	1.8	0.037	0.008
$a_1, a_4 \rightarrow T$	0.0200	0.0364	0.0125	0.6229	17.1	23784	1.2	0.424	0.058
$a_1 \rightarrow T$	0.0626	0.0364	0.0147	0.2347	6.5	23281	2.0	0.490	0.111
$a_6 \rightarrow T$	0.2613	0.0364	0.0133	0.0511	1.4	18947	7.2	0.362	0.370

Table 6.5 Selected impact-contrasted activity patterns

Patterns (P)	$Supp_{D_T}(P)$	$Supp_{D_{\bar{T}}}(P)$	$Cd_{T,\bar{T}}(P)$	$Cdr_{T,\bar{T}}(P)$	$Cd_{\bar{T},T}(P)$	$Cdr_{\bar{T},T}(P)$	*AvgAmt*	*AvgDur*	$risk_{amt}$	$risk_{dur}$
							(cents)	(days)		
a_4	0.446	0.138	0.309	3.24	-0.309	0.31	21749	3.2	0.505	0.203
a_8	0.176	0.060	0.117	2.97	-0.117	0.34	26789	9.9	0.246	0.245
a_4, a_{15}	0.255	0.092	0.163	2.78	-0.163	0.36	21127	3.9	0.280	0.141
a_{14}, a_{14}, a_4	0.367	0.091	0.276	4.04	-0.276	0.25	21761	2.9	0.415	0.151

Table 6.6 Selected impact-reversed activity patterns

Underlying sequence(P)	Impact 1	Derivative activity$Q$	Impact 2	*Cir*	*Cps*	Local support of $P \rightarrow$Impact 1	Local support of $PQ \rightarrow$Impact 2
a_{14}	$\bar{T}$	a_4	T	2.5	0.013	0.684	0.428
a_{16}	$\bar{T}$	a_4	T	2.2	0.005	0.597	0.147
a_{14}	$\bar{T}$	a_5	T	2.0	0.007	0.684	0.292
a_{16}	$\bar{T}$	a_7	T	1.8	0.004	0.597	0.156
a_{14}, a_{14}	$\bar{T}$	a_4	T	2.3	0.016	0.474	0.367
a_{16}, a_{14}	$\bar{T}$	a_5	T	2.0	0.006	0.402	0.133
a_{16}, a_{15}	$\bar{T}$	a_5	T	1.8	0.006	0.339	0.128
a_{14}, a_{16}, a_{14}	$\bar{T}$	a_{15}	T	1.2	0.005	0.248	0.188

where *Impact 1* is opposite to *Impact 2*, and Q is a derived activity or sequence. Cir stands for *conditional impact ratio*, which shows the impact of the derived activity on *Impact 2* when the underlying pattern happens. Cps denotes *conditional P-S ratio*. Both Cir and Cps show how much the impact is reversed by the derived activity Q. For example, the first row shows that the appearance of a_4 tends to change the impact from $\bar{T}$ to T when a_{14} happens first. It indicates that, when a_{14} occurs first, the appearance of a_4 makes it more likely to become debtable. This pattern pairs indicate what effect an additional activity will have on the impact of the patterns.

6.4 Case Study III: Combining Association Rules from Heterogeneous Data Sources to Discover Repayment Patterns

This section presents an application of combined association rules to discover patterns of quick/slow payers [18, 21]. Heterogeneous data sources, such as demographic and transactional data, are part of everyday business applications and used for data mining research. From a business perspective, patterns extracted from a single normalized table or subject file are less interesting or useful than a full set of multiple patterns extracted from different datasets. A new technique has been designed to discover combined rules on multiple databases and applied to debt recovery in the social security domain. Association rules and sequential patterns from different datasets are combined into new rules, and then organized into groups. The rules produced are useful, understandable and interesting from business perspective.

6.4.1 Business Problem and Data

The purpose of this application 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.

Three datasets containing 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 that the debt amount is related to, and so on. The repayments dataset contains arrangement types, repayment types, date and time of repayment, repayment amount, repayment method (e.g., post office, direct debit, withholding payment), etc. Quick/moderate/slow payers are defined by domain experts based on the time taken to repay the debt, the forecasted time to repay and the frequency/amount of repayment.

6.4.2 Mining Combined Association Rules

The idea was to firstly 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 arrangement and repayment, customers were partitioned into groups according to their arrangement and repayment type. Secondly, pay-off timeframe distribution and statistics for each group were presented to domain knowledge experts, who then decided who were quick/slow payers by group. The criterion was applied to the data to tag every customer as quick/slow payer. Thirdly, association rules were generated for quick/slow payers in each single group. And lastly, the association rules from all groups were organized together to build potentially business-interesting rules.

To address the business problem, there are two types of rules to discover. The first type are rules with the same arrangement and repayment pattern but different demographic patterns leading to different customer classes (see Formula 6.1). The second type are rules with the same demographic pattern but different arrangement and repayment pattern leading to different customer classes (see Formula 6.2).

$$\text{Type A:}\ \begin{cases} A_1 + D_1 \rightarrow \text{quick payer} \\ A_1 + D_2 \rightarrow \text{moderate payer} \\ A_1 + D_3 \rightarrow \text{slow payer} \end{cases} \tag{6.1}$$

$$\text{Type B:}\ \begin{cases} A_1 + D_1 \rightarrow \text{quick payer} \\ A_2 + D_1 \rightarrow \text{moderate payer} \\ A_3 + D_1 \rightarrow \text{slow payer} \end{cases} \tag{6.2}$$

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

6.4.3 Experimental Results

The data used was debts raised in calendar year 2006 and the corresponding customers and repayments in the same year. Debts raised in calendar year 2006 were first selected, and then the customer data and repayment data in the same year related to the above debt data were extracted. The extracted data was then cleaned by removing noise and invalid values. The cleansed data contained 479,288 customers with demographic attributes and 2,627,348 repayments.

Selected combined association rules are given in Tables 6.7 and 6.8. Table 6.7 shows examples of rules with the same demographic characteristics. For those customers, different arrangements lead to different results. It 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*", and not to pay under “*Arrangement=Withholding and Voluntary Deduction, Repayment=Withholding and Direct Debit*" or “*Arrangement =Cash and Irregular, Repayment=Cash or Post Office*", so that the debt will likely be repaid quickly.

Table 6.7 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 6.8 Selected Results with the Same Arrangement-Repayment Patterns

Arrangement	Repayment	Demographic Pattern	Result	Expected Conf(%)	Conf (%)	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

Table 6.8 shows 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*".

6.5 Case Study IV: Using Clustering and Analysis of Variance to Verify the Effectiveness of a New Policy

This section presents an application of clustering and analysis of variance to study whether a new policy works or not. The aim of this application was to examine earnings related transactions and earnings declarations in order to ascertain, whether significant changes occurred after the implementation of the “welfare to work" initiative on 1st July 2006. The principal objective was to verify whether customers declare more earned income after the changes, the rules of which allowed them to keep more earned income and still keep part or all of their income benefit.

The population studied in this project were customers who had one or more non-zero declarations and were on the 10 benefit types affected by the “Welfare to Work" initiative across two financial years from 1/7/2005 to 30/6/2007. Three datasets were available and each of them contained 261,473 customer records. Altogether there were 13,596,596 declarations (including “zero declarations"), of which 4,488,751 were non-zero declarations. There are 54 columns in transformed earnings declaration data. Columns 1 and 2 are respectively customer ID and benefit type. The other 52 columns are declaration amounts over 52 fortnights.

6.5.1 Clustering Declarations with Contour and Clustering

At first we employed histograms, candlestick charts and heatmaps to study whether there were any changes between the two years. The result from histograms, candlestick charts and heatmaps all show that there was an increase of the earnings declaration amount for the whole population.

For the whole population, scatter plot indicates no well-separated clusters, while contour shows that some combinations of fortnights and declaration amounts had more customers than others (see Figure 6.1). It’s clear that the densely populated areas shifted from low amounts to large amounts from financial year 2005-2006 to financial year 2006-2007. Moreover, the sub-cluster of declarations ranging from \$50 to \$150 reduced over time, while the sub-cluster ranging from \$150 to \$250 expanded and shifted towards higher amounts.

The clustering with k-means algorithm did not generate any meaningful clusters. The declarations were divided into clusters by fortnights when the amount is small, while the dominant factor is not time, but amount, when the amount is high. A density-based clustering algorithm, DBSCAN [10], was then used to cluster the declarations below \$1000, and due to limited time and space, a random sample of 15,000 non-zero declarations was used as input into the algorithm. The clusters found for all benefit types are shown in Figure 6.2. There are four clusters, separated by the beginning of new year and financial year. From left to right, the four clusters shift towards larger amounts as time goes on, which shows that the earnings declarations increase after the new policy.

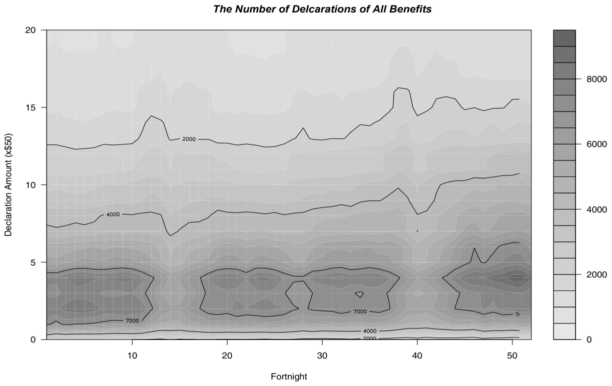

Fig. 6.1 Contour of Earnings Declaration

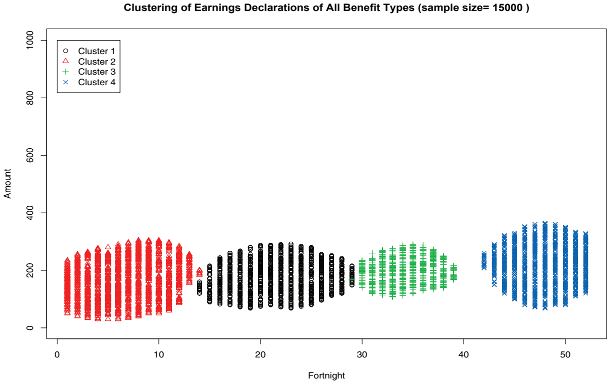

Fig. 6.2 Clustering of Earnings Declaration

The clustering with k-means algorithm does not generate any meaningful clusters. The declarations are divided into clusters by fortnights when the amount is small, while the dominant factor is not time but amount when the amount is high. A density-based clustering algorithm, DBSCAN [10], is then used to cluster the declarations below $1000, and due to limited time and space, a random sample of 15,000 non-zero declarations is used as input into the algorithm. The clusters found for all benefit types are shown in Figure 6.2. There are four clusters, separated by the beginning of new year and financial year. From left to right, the four clusters shift towards larger amounts as time goes on, which shows that the earnings declarations increase after the new policy.

Table 6.9 Hypothesis Test Results Using Mixed Model

Benefit Type	DenDF	FValue	ProbF
APT	1172	4844.50	<.0001
AUS	4872	2.94	0.0863
JSK	9351	2413.06	<.0001
NMA	317	5.67	0.0178
NSA	77801	1448.89	<.0001
PPP	11579	1782.00	<.0001
PTA	3102	421.04	<.0001
SKA	425	2.55	0.1112
STU	41623	16475.2	<.0001
WDA	3398	126.28	<.0001

6.5.2 Analysis of Variance

Hypothesis test was also used to study whether there were significant changes of the earnings declaration before/after the initiative. We employed mixed models to test the changes of earnings declaration. The Mixed Procedure in SAS was used and tests were conducted for every benefit type. The results are shown in Table 6.9, where “DenDF" shows the sample size, “Fvalue" gives an understanding on the difference before/after initiative. The greater this value is, the bigger the difference. “ProbF<0.0001" means there is significant change before/after initiative. So the customers with payment/benefit types APT, JSK, NSA, PPP, PTA, STU and WDA are all with significant changes. “ProbF>0.0001" has two meanings: 1) the difference is not significant, or 2) the sample size is not large enough. Because the sample size of NMA and SKA are very small, the hypothesis test cannot give reliable result on whether there are significant changes or not. The changes on the AUS customers are not significant. The results show that there are significant change for most benefit types, which suggests that the new policy is effective.

6.6 Conclusions and Discussion

Data mining techniques have been used in a social security environment to check the effectiveness of a new policy and discover demographic patterns, activity sequence patterns and debt recovery patterns. The demographic patterns discovered may be used to identify customer groups with a high probability of debt, so that reviews can be conducted to assist in the reduction of debts. Moreover, by discovering activity sequence patterns associated with debt/non-debt, appropriate actions can be suggested to reduce the probability of customers’ acquiring a debt. We have also presented effective and practical techniques for discovering rare but significant *impact-targeted activity patterns* in imbalanced activity data and a framework for mining combined association rules from multiple datasets. The above is part of our initial effort to tackle business problems using 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. Firstly, given the likelihood that hundreds or possibly thousands of rules are identified after pruning redundant patterns, how can we efficiently select interesting patterns from them? Secondly, how can domain knowledge be effectively incorporated in data mining procedure to reduce the search space and running time of data mining algorithms? Thirdly, given that the business data is complicated and a single debt activity may be linked to several customers, how can existing approaches for sequence mining be improved to take into consideration the linkage and interaction between activity sequences? And lastly and perhaps most importantly, how can these discovered rules be used to build an efficient debt prevention system to effectively detect debt in advance and give appropriate suggestions to reduce or prevent debt? The above will be part of our future work.

Acknowledgments

We would like to thank Mr Fernando Figueiredo, Mr Peter Newbigin and Mr Brett Clark from Centrelink, Australia, for their support of domain knowledge and helpful suggestions.

This work was supported by the Australian Research Council (ARC) Linkage Project LP0775041 and Discovery Projects DP0667060 & DP0773412, and by the Early Career Researcher Grant from University of Technology, Sydney, Australia.

References

1. R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB*, pages 487–499, Santiago, Chile, September 1994.
2. L. Cao, Y. Zhao, and C. Zhang. Mining impact-targeted activity patterns in imbalanced data. *Accepted by IEEE Transactions on Knowledge and Data Engineering in July 2007. IEEE Computer Society Digital Library. IEEE Computer Society, http://doi.ieeecomputersociety.org/10.1109/TKDE.2007.190635*.
3. Centrelink. Centrelink fraud statistics and centrelink facts and figures, url: 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. Accessed in May 2006.
4. J. Chattratichat, J. Darlington, Y. Guo, S. Hedvall, M. Köler, and J. Syed. An architecture for distributed enterprise data mining. In *HPCN Europe '99: Proceedings of the 7th International Conference on High-Performance Computing and Networking*, pages 573–582, London, UK, 1999. Springer-Verlag.
5. V. Crestana-Jensen and N. Soparkar. Frequent itemset counting across multiple tables. In *PAKDD'00: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, pages 49–61, London, UK, 2000. Springer-Verlag.

6. L. Cristofor and D. Simovici. Mining association rules in entity-relationship modeled databases. Technical report, University of Massachusetts Boston, 2001.
7. P. Domingos. Prospects and challenges for multi-relational data mining. *SIGKDD Explor. Newsl.*, 5(1):80–83, 2003.
8. G. Dong and J. Li. Efficient mining of emerging patterns: discovering trends and differences. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 43–52, New York, NY, USA, 1999. ACM.
9. S. Dzeroski. Multi-relational data mining: an introduction. *SIGKDD Explor. Newsl.*, 5(1):1–16, 2003.
10. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996.
11. H. Guo and H. L. Viktor. Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. *SIGKDD Explor. Newsl.*, 6(1):30–39, 2004.
12. B. Park and H. Kargupta. Distributed data mining: Algorithms, systems, and applications. In N. Ye, editor, *Data Mining Handbook*. 2002.
13. J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. Mining sequential patterns by pattern-growth: The prefixspan approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1424–1440, 2004.
14. F. Provost. Distributed data mining: Scaling up and beyond. In *Advances in Distributed and Parallel Knowledge Discovery*. MIT Press, 2000.
15. A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970–974, 1996.
16. Q. Yang, J. Yin, C. X. Ling, and T. Chen. Postprocessing decision trees to extract actionable knowledge. In *ICDM '03: Proceedings of the Third IEEE International Conference on Data Mining*, page 685, Washington, DC, USA, 2003. IEEE Computer Society.
17. M. Zaki. Mining non-redundant association rules. *Data Mining and Knowledge Discovery*, 9:223–248, 2004.
18. H. Zhang, Y. Zhao, L. Cao, and C. Zhang. Combined association rule mining. In T. Washio, E. Suzuki, K. M. Ting, and A. Inokuchi, editors, *PAKDD*, volume 5012 of *Lecture Notes in Computer Science*, pages 1069–1074. Springer, 2008.
19. J. Zhang, E. Bloedorn, L. Rosen, and D. Venese. Learning rules from highly unbalanced data sets. In *ICDM '04: Proceedings of the Fourth IEEE International Conference on Data Mining*, pages 571–574, Washington, DC, USA, 2004. IEEE Computer Society.
20. Y. Zhao, L. Cao, Y. Morrow, Y. Ou, J. Ni, and C. Zhang. Discovering debtor patterns of centrelink customers. In *Proc. of The Australasian Data Mining Conference: AusDM 2006*, Sydney, Australia, November 2006.
21. Y. Zhao, H. Zhang, F. Figueiredo, L. Cao, and C. Zhang. Mining for combined association rules on multiple datasets. In *Proc. of 2007 ACM SIGKDD Workshop on Domain Driven Data Mining (DDDM 07)*, pages 18–23, 2007.