Journal of Intelligent Information Systems, Vol.9, No.2, November, 2003 (pp.135~154)

Mining Association Rules of Credit Card Delinquency of Bank Customers in Large Databases

Young-Chan Lee[a], Soo-Il Shin[b]

[a] Institute for Business Research, Sogang University
[b] MetLife Insurance Co. of Korea, Ltd. Marketing team
([a] *chanlee@sogang.ac.kr*, [b] *sooils@metlifekorea.co.kr*)

Credit scoring system (CSS) starts from an analysis of delinquency trend of each individual or industry. This paper conducts a research on credit card delinquency of bank customers as a preliminary step for building effective credit scoring system to prevent excess loan or bad credit status. To serve this purpose, we use association rules as a rule generating data mining technique. Specifically, we generate sets of rules of customers who are in bad credit status because of delinquency by association rule mining. We expect that the sets of rules generated by association rule mining could act as an estimator of good or bad credit status classifier and basic component of early warning system.

Key words: Association rules; credit scoring system

Received: October 2003 Accepted: November 2003 Corresponding Author: Yong-Chan Lee

1. Introduction

The total amount of credit in household economy exceeds about four times of budget of government in South Korea. Such an increasing of household credit is caused of activated personal consuming after overcoming economic crisis in 1998. In 2002, real estate and property worth rise steadily, and that affect to drive taking a loan from bank. With together, credit and bank industry had a full ability of lending money to household or industry. Since 1999, household consumption is more than Gross National Income (GNI), and also total amount of household credit exceeds Gross Domestic Product (GDP). These situations are one of the reasons of increasing personal bad credit status. In October 2002, the total amount of credit in household economy exceeds about four times of budget of government in South Korea, and the number of persons who registered in government as a bad credit status is over 2.5 million.

We must build effective credit scoring system to prevent excess loan or bad credit status.

To decide whether to grant credit to customers has gained more and more attention for credit industry due to the industry has been experiencing high growth rate during the past few decades. With the rapid growth in credit industry, credit scoring models have been extensively used for the credit admission evaluation. Today credit scoring is used by 97% of banks that approve credit card applications and by 82% of banks that determine whom to solicit for credit cards. With the growth in financial services there have been mountain losses from delinquent loans. In response, many organizations in the credit industry are developing new models to support the credit decision. The objective of credit scoring models is to assign credit applicants to either a 'good credit' group that is likely to repay financial obligation or a 'bad credit' group whose application will be denied because of its high possibility of defaulting on the financial obligation more accurately which means more creditworthy applicants are granted credit thereby increasing profits; non-creditworthy applicants are denied credit thus decreasing losses. Therefore credit scoring problems are basically in the scope of the more general and widely discussed discrimination and classification problems (Anderson, 1984; Chen and Huang, 2003; Dillion and Goldstein, 1984; Hand, 1981; Johnson and Wichern 1998; Lee et al., 2002; Morrison, 1990; West, 2000).

So far, the statistical methods, nonparametric statistical methods, and artificial intelligence approaches have been proposed to support the credit decision (Bryant, 1997; Buta, 1994; Coakley and Brown, 2000; Davis et al., 1992; Dimitras et al., 1996; Emel et al., 2003; Falbo, 1991; Frydman et al., 1985; Martin, 1997; Reichert et al., 1983; Tam and Kiang, 1992; Troutt et al., 1996).

Classification is a commonly encountered decision making tasks in business. Categorizing an object into a predefined group or class based on a number of observed attributes related to that object is a typical classification problem (Zhang, 2000). In addition to credit scoring and corporate distress prediction, neural networks (NNs) have been successfully applied to a variety of real world classification tasks in industry, business and science. A number of performance comparisons between neural and conventional classifiers have been made by many studies (Curram & Mingers, 1994; Lee et al., 1997; Lee et al., 1999; Desai et al., 1996; Desai et al., 1997; Jensen, 1992; Markham & Ragsdale, 1995; Piramuthu, 1999; West, 2000).

Conventional statistical classification procedures such as linear discriminant analysis (LDA) and logistic regression analysis (LRA) are constructed on the Bayesian decision theory. In these classification techniques, an underlying probability model must be assumed in order to calculate the posterior probability upon which the classification decision is made (Chen and Huang, 2003).

In credit industry, NN has recently been claimed to be an accurate tool for credit analysis (Desai et al., 1996; Malhotra and Malhotra, 2002; West, 2000). Desai et al. (1996) have explored the ability of NN and traditional statistical techniques such as LDA and LRA, in constructing credit scoring models. Their results indicated that NN shows promise if the performance measure is percentage of bad loans accurately classified. However, if the performance measure is percentage of good and bad loans accurately classified, LRA is as good as NN. The percentage of bad loans correctly classified is an important performance measure for credit scoring models since the cost of granting a loan to a defaulter is much larger than that of rejecting a good applicant (Desai et al., 1996). West (2000) has investigated the accuracy of quantitative models commonly used by the credit industry. The results indicated that NN can improve the credit scoring accuracy. West (2000) also suggested that LRA is a good alternative to NN. On the other hand, LDA, KNN, and classification and regression tree (CART) did not produce encouraging results.

From the extensive survey of NN applications in business (Vellido et al., 1999), it indicates that NN shows promise in various areas where nonlinear relationships are believed to exist within the datasets, and traditional statistical approaches are deficient. In credit prediction, the nonlinear features of NNs make them a potential alternative to traditional parametric (e.g. LDA and LRA) and nonparametric (e.g. case-based reasoning and decision tree) methods (Chen and Huang, 2003). However, NN is commonly considered as a black-box technique without logic or rule-based explanations for the input-output approximation. A main shortage of applying NN for credit scoring is the difficulty in explaining the underlying principle for the decision to rejected applications (West, 2000).

Unfortunately, most of companies in Korea have weak credit scoring system and they have not activated joint ownership of credit information. In fact, credit scoring system starts from an analysis of delinquency trend of each individual or industry. Therefore, an effective analysis of delinquency is the starting point of good estimation of credit status in anywhere. This study performs a research about credit card delinquency of bank customers as a preliminary step for building effective credit scoring system.

For this research, we use association rules as a rule generating data mining technique. Because association rule mining finds all the rules existing in the database that satisfy constraints user-specified in advance, it is well suited to find meaningful delinquency patterns and estimate customer's future credit status as possible. Specifically, we generate sets of rules of customers who

are in bad credit status because of delinquency by association rule mining. We expect that the result of this study can be a standard of estimating good or bad credit status of personal and basic component of early warning system of default or overdue in various financial products.

On the other hand, the aim of data mining is to discover useful or interesting rules (Fayyad et al., 1996). However, past applications have found that it is easy to generate a large number of rules from a database, and most of them are not useful to the user (Piatesky-Shapiro and Matheus, 1994; Piatesky-Shapiro et al., 1994, Silberschatz and Tuzhilin, 1996; Liu and Hsu, 1996). The presence of the huge number of rules makes it difficult for the user to analyze them and to identify those that are of interest to him or her. Therefore, post-analysis, including rule validation, becomes an important issue for building accurate behaviors of customers (Adomavicius and Tuzhilin, 2001).

As mentioned in the above, this paper captures the comprehensive patterns of credit card delinquency of bank customers using conjunctive rules that are learned from customer transactional databases using association rule mining. However, there are drawbacks to this approach due to the nature of personalization applications. In particular, as will be explained in the paper, the credit card delinquency rules learned about individual customers can be unreliable, irrelevant, or obvious. In this paper, during the post-analysis stage of the data mining process, this validation process is performed by the domain expert who can iteratively apply various rule validation operators.

2. Related Works

Association rules

Association rules are a class of important regularities that exist in database. Since, it was first introduced in Agrawal et al. (1993), the problem of mining associations has received a great deal of attention and has been studied extensively in the past (Aggarwal and Yu, 1998; Agrawal et al., 1993; Agrawal and Srikant, 1994; Brin et al., 1997; Chiang et al., forthcoming; Han and Fu, 1995; Liu et al., 1999; Ng et al., 1998; Park et al., 1995; Rastogi and Shim, 1998; Srikant et al., 1997). The basic model of association rules is as follows.

Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form $X \rightarrow Y$, where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y.

An example of an association rule is: "10% of transactions contain *cheese* and *beer* together, and transactions that contain *cheese* also contain *beer* 80% of the time." Here 10% is the *support* of the rule and 80% is called the *confidence* of the rule (Liu at al., 1999). The problem is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints. Applications include discovering affinities for market basket analysis and cross-marketing, catalog design, loss-leader analysis, store layout, customer segmentation based on buying patterns, etc. (Srikant and Agrawal, 1995).

The definition of association rule is as following: Make $I = \{i_1, i_2, \mathrm{K}, i_m\}$ as the itemset, in which each item represents a specific commodity. D stands for a trading database in which each transaction T represents a itemset. That is $T \subseteq I$. Each itemset is a non-empty sub-itemset of I and the only identify code is *TID* (Transaction ID): Each itemset, $X \subset I$ has a measure standard - *Support*, to evaluate the statistical importance in D. $Support(X, D)$ denotes the rate of merchandising in transaction D (Agrawal and Srikant 1994).

The format of the association rule is $X \rightarrow Y$ in which $X, Y \subset I$ and $X \cap Y = \varnothing$. The interpretation of this association rule is that if X is purchased, Y can be bought at the same time. Each rule has a measuring standard called *Confidence*; i.e. $Confidence(X \rightarrow Y) = Support(X \cup Y, D) / Support(X, D)$. In this case, Confidence denotes if the merchandise including X, the chance of buying Y is relatively high (Chiang et al., forthcoming).

An association mining algorithm works in two steps. First step is to generate all large itemsets that satisfy minimum support (*Minsup*). Second step is to generate all association rules that satisfy minimum confidence (*Minconf*) using the large itemset. Therefore, exploring the association rules means to find out all the association rules of formats and meet the following conditions:

$$Support(X \cup Y, D) \geq Minsup$$

$$Confidence(X \rightarrow Y) \geq Minconf$$

The *Minsup* and *Minconf* are both set by the users. In general, the numbers of the transactions that comprising X is called the support of X denoted by σx. Make *Minsup* the minimum value of support. If the support of X meets the condition, $\sigma x \geq Minsup$, X is the large itemset (Chinag et al., forthcoming).

As for the exploration of the association rules, many researchers take the Apriori algorithm

(Agrawal and Srikant 1994) supported by Agrawal et al. (1993) as the basic formulation. Fig. 1 gives the Apriori algorithm. The first pass of the algorithm simply counts item occurrences to determine the large 1-itemsets. A subsequent pass, say pass k, consists of two phases. First, the large itemsets L_{k-1} found in the (k-1)th pass are used to generate the candidate itemsets C_k, using the apriori-gen function described in Figure 2. Next, the database is scanned and the support of candidates in C_k is counted. For fast counting, we need to efficiently determine the candidates in C_k that are contained in a given transaction t.

```
1) L1={large 1-itemsets};
2) For (k=2; Lk-1≠Ø; k++) do begin
3)     Ck = Apriori-gen (Lk-1); //New candidates
4)     For all transactions t ∈ D do begin
5)         Ct = subset(Ck, t); //Candidates contained in t
6)         For all candidates c ∈ Ct do
7)             c.count + +;
8)     End
9)     Lk = {c ∈ Ck| c.count ≥ min sup}
10) End
11) Answer = ∪kLk ;
```

Fig. 1. Apriori algorithm

The apriori-gen function takes as argument L_{k-1}, the set of all large (k-1)-itemsets. It returns a superset of the set of all large k-itemsets. The function works as follows. First, in the join step, we join L_{k-1} with L_{k-1}. Next, in the prune step, we delete all itemsets $c \in C_k$ such that some (k-1)-subset of c is not in L_{k-1}. This algorithm can be ceased when no further candidate itemset can be generated.

```
Apriori-gen( )
{
//Join step
Insert into Ck
Select p.item1, p.item2, ..., p.itemk-1, q.itemk-1
From Lk-1 p, Lk-1 q
Where p.item1 = q.item1, ..., p.itemk-2 = q.itemk-2,
      p.itemk-1 < q.itemk-1;
//Prune step
For itemsets c ∈ Ck do
    For all (k-1)-subsets s of c do
        If (s ∉ Lk-1) then
            Delete c from Ck;}
```

Fig. 2. Apriori-gen function

For example, in Fig. 3, database D is the original transaction database. We assume that minimum support is 2 transactions in Fig. 4. First, calculate the number of each item that appears in the transaction database, which is to calculate the support and to evaluate whether the number is bigger than or equal to the minimal support and determine the Large 1-itemsets, L_1. Next, generate candidate 2-itemsets, C_2 from $L_1 \times L_1$. Further, calculate the support of C_2 to create L_2. From $L_2 \times L_2$ brings about Candidate 3-itemsets, C_3. In the phase of join, we have {B, C, E}. And in the phase of pruning, the sub-itemsets {B, C}, {B, E}, {C, E} of {B, C, E} are all comprised in L_2. Thus, it does meet Candidate 3-itemset C_3. However, Candidate 4-itemset, C_4 is not generated from $L_3 \times L_3$. The algorithm, therefore, is terminated. In consequence, the large itemsets generated are L1 = {A}, {B}, {C}, {E}, L2 = {A, C}, {B, C}, {B, E}, {C, E}, and L3 = {B, C, E} as demonstrated in Fig. 4 (Chiang et al., forthcoming).

TID	Items			
100	A	C	D	
200	B	C	E	
300	A	B	C	E
400	B	E		

Fig. 3. Original transaction database

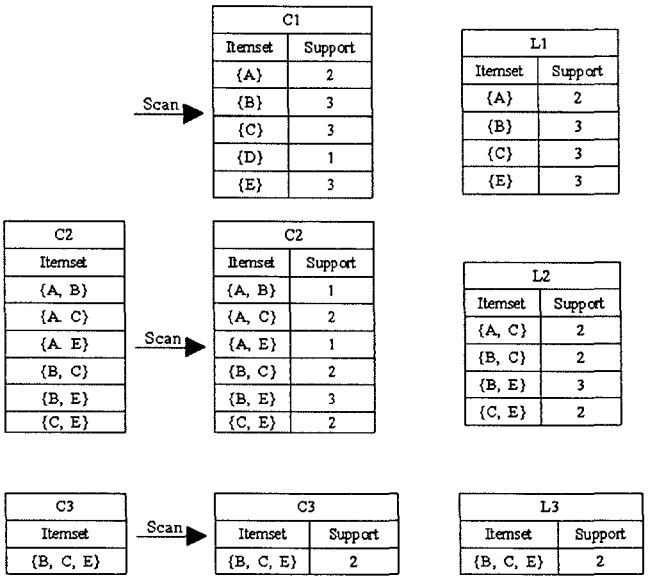

Fig. 4. Generation of candidate itemsets and large itemsets

Rule validation

Generally, rules of pattern can be discovered using various data mining algorithms. For example, to discover association rules, we can use Apriori (Agrawal et al., 1996) and its numerous variations. Similarly, to discover classification rules, we can use CART (Breiman et al., 1984), C4.5 (Quinlan, 1993), or other classification rule discovery methods.

The "quality" of rules can be defined in several ways. In particular, rules can be "good" because they are (1) statistically valid, (2) acceptable to a human expert in a given application, (3) "effective" in the sense that they result in certain benefits obtained in an application. In this paper, we focus on the first two aspects, i.e., statistical validity and acceptability to an expert (Adomavicius and Tuzhilin, 2001).

The rule validation problem in the post-analysis stage of the data mining process has been addressed before in the data mining community. In particular, there has been work done on specifying filtering constraints that select only certain types of rules from the set of all the discovered rules; examples of this research include (Klemettinen et al., 1994; Liu and Hsu, 1996; Liu et al., 1999). In these approaches the user specifies constraints but does not do it iteratively. In contrast to this, it has been observed by several researchers, for example, Brachman and Anand (1996), Fayyad et al. (1996), Silberschatz and Tuzhilin (1996a), Provost and Jensen (1998), Lee et al. (1998), Adomavicius and Tuzhilin (1999, 2001), Sahar (1999), that knowledge discovery should be an iterative process that involves an explicit participation of the domain expert, and we apply this point of view to the rule validation process.

A common way to perform the post-analysis of data mining results is to let the domain expert perform this task, and several data mining systems support this capability. In this paper, association rules discovered during the data mining stage are validated by the expert, and, depending on how well they represent the actual credit card usage behaviors of the customers, some rules are "accepted" and some "rejected" by the expert. The overall process of rule validation is represented in Fig. 5 (Adomavicius and Tuzhilin, 2001).

We provide a framework allowing the human expert validate large numbers of rules at a time with relatively little input from the expert. This is done by applying different *rule validation operators*. This examination process can be performed in the following two ways. First, the expert may already know some types of rules that he or she wants to examine and accept or reject based on the prior experience. Therefore, it is important to provide capabilities allowing him or her to specify such types of rules in advance. Second, the expert may not know all the relevant types

of rules in advance, and it is important to provide methods that group discovered rules into classes that he or she can subsequently examine and validate. The former is called *template-based filtering operators*, the latter is called *similarity-based rule grouping operator*. In this paper, we use *template-based rule filtering* to assess the validation of generated rules (Adomavicius and Tuzhilin, 2001; Lent et al., 1997; Toivonen et al., 1995; Wang et al., 1998).

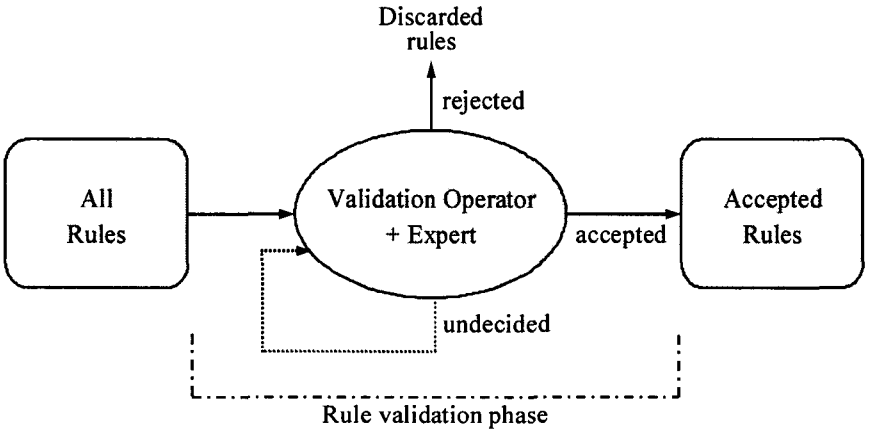

Fig. 5. Rule validation process

3. Research Methodology

Raw data

This paper used data extracting based on the customer profiles of a domestic K bank. This study focused on the analysis of customer data that settlement amount was withdrawn in a K bank account among customer's credit cards. If delinquency was occurred, we divided customer's credit status into good and bad at that time. The numbers of whole customers is total 2,016,223, and predict variables about each person are consisted of total 63. The amount and usage of credit cards included in each record is consisted of one year accumulation number and amount (see Table. 1).

Table 1. Data record

Age	Sex	Zip Code-Home Address	Zip Code-Office
Trust Balance	Deposit Balance	Loan Balance	Lump Sum Payment Balance
Installment Amount	Installment Usage	Foreign Cash Amount	Foreign Cash Usage
Revolving Amount	Revolving Usage	Cash Service Amount	Cash Service Usage
Other Credit Cash Amount	Other Credit Sales Amount	Account Opening Date	Net Profit
Cash Service Channels - CD/ATM, BC Card ARS, KCI, Tele-Banking, PC Banking			
Major Affiliation I, II, III Usage		Maximum Usage of Affiliation I, II, III	
6 Month Credit Card Overdue Amount	Overdue Days	Maximum Amount of Affiliation I, II, III	
Cash Service Overdue Amount	Installment Balance	Installment Overdue Amount	Card Loan Overdue Amount
Number of Other Firm's Credit Cards	Number of Own Bank Credit Cards	Card Loan Balance	Agreed Card Loan Amount
Cash Back Usage	Cash Back Amount	Mileage Usage	Mileage Amount
Railroad Usage	Railroad Amount	Event Usage	Event Amount
Gas Station Usage	Gas Station Amount	Free Movie Usage	Free Movie Amount
Cell-Phone Usage	Cell-Phone Amount	Bookstore Usage	Bookstore Amount

Variables

We defined good and bad credit status via 60 delinquency days. Therefore, if delinquency days are over 60 via withdrawal day, it is defined as bad credit status.

As previously mentioned, original data records are consisted of 63 variables. In this paper, we discarded several variables before constructing data set. Specifically, variables related commission and foreign usage, affecting directly on delinquency days, occurred from additional credit card function, and address etc. are eliminated. Also, we converted variables divided into usage and amount separately into amount/usage. Table 2 shows variables included in final data set.

Table 2. Final variables

Age	Sex	Trust Balance	Deposit Balance
Loan Balance	Installment Amount/Usage	Installment Usage	Lump Sum Payment Usage
Lump Sum Payment Amount/Usage	Cash Service Amount/Usage	Cash Service Usage	Revolving Usage
Revolving Amount/Usage	Card Loan Balance	Number of Other Firm's Cards	Overdue Days
Cash Service Channel (CD/ATM, BC Card ARS, KCI, Tele-Banking, PC Banking)			
Major Affiliation Ⅰ,Ⅱ,Ⅲ Usage			

Data set and partition

This paper constituted a data set with total 1,000 data records. 1,000 data records are consisted of 500 good and 500 bad credit status data in respectively. In order to solve the problem of generalization and underfitting (or overfitting), this paper divided data set into training and validation set.

Support and confidence

This paper set the minimal support to 2% to include at least 20 cases (2% of total 1,000 data records) in one rule. And the minimal confidence is set to be 80%. This means that at least 80% records among records satisfying conditions of rules can satisfy the results of rules. Both of support and confidence are set via bad credit status. The number of conditions to be included in conditions of rules is set to be 3.

Rule validation test

This paper used *template-based rule filtering* to assess the validation of generated rules. Template-based rule filtering is divided into *template using statistical parameter* and *rule syntax template* (Adomavicius and Tuzhilin, 2001). In case of template using statistical parameter, we can prevent over-generating rules by setting support and confidence parameters in advance. Also, in case of rule syntax template, we can eliminate corresponding rules when specific variable or syntax is used in rules.

4. Empirical Analysis

This paper used association rules to generate rules of credit card delinquency of bank customers. To serve this purpose, we use *Wiz Why 4.02*. This software is designed for association rules, and based on Apriori algorithm.

Classification result of association rule

Association rules generate rules to predict good or bad credit status. Therefore, we can judge the predicting power with comparing predicted credit status from rules and actual credit status. Table 3 shows classification result of association rules.

Table 3. Classification result of association rules

		Actual Group			Predicting Power	
		Good	Bad	Total	Sensitivity	89.2%
Classified Group	Good	388 (77.6%)	54 (10.8%)	425	Specificity	77.6%
	Bad	112 (22.4%)	446 (89.2%)	575	Accuracy	83.4%
	Total	500	500	1000	Misclassification	16.6%
					Positive Predicted	91.2%
					Negative Predicted	77.6%

Association rule

Total number of rules is 130 through association rule generating program, these rules represent that the rule of each person's credit status is bad. Also, 130 rules satisfied both support (20 cases in one rule) and confidence (80%) previously set by us. The order of generated rules is numbered by confidence level. So, the confidence of 1st rule is most high, and 130th rule has most low confidence value, 80%. Through rule validation process, 27 rules are eliminated, final 103 rules are selected. Table 4 shows total number of rules generated, eliminated rules, and final rules selected.

Table 4. Number of rules

Total # of Rules	# of Eliminated Rules	# of Final Selected Rules
130	27	103

According to confidence level, this paper found top 10 rules, and arranged frequency of predict variables used in every rules.

As shown in Table 5, rule 3) has most high confidence level, 95.2%, includes rules of age, lump sum payment usage, and cash service amount/usage. Specifically, the person whose age is ranged from 19 to 24, lump sum payment usage ranged from 4 to 11, and cash service amount/usage ranged from 20,000 to 600,000 can be classified into bad credit status. Also, the number of bank customers satisfying conditions of rules is 20, and 19 customers of them (95.2%) satisfy results of rules.

Table 5. Confidence Top 10 Rules

Rules	Reference
3) If Age = 19.00 ~ 24.00 (Average = 22.29) and Lump Sum Payment Usage = 4.00 ~ 11.00 (Average = 6.71) and Cash Service Amount/Usage = ₩ 20,000 ~ ₩ 600,000 (Average = ₩ 269,036) Then Bad	Confidence: 0.952 # of Records: 20. Significance: p < 0.00001
4) If Lump Sum Payment Usage = 4.00 ~ 11.00 (Average = 6.68) and Installment Usage = 4.00 ~ 8.00 (Average = 5.36) and Installment Amount/Usage = ₩ 278,108 ~ ₩ 664,500 (Average = ₩ 423,445) Then Bad	Confidence: 0.929 # of Records: 26. Significance: p < 0.00001
5) If Sex = Male and Lump Sum Payment Usage = 4.00 ~ 11.00 (Average = 7.36) and Major Affiliation III Usage = 2.00 Then Bad	Confidence: 0.929 # of Records: 26. Significance: p < 0.00001
6) If # of Other Cards = 5.00 and CD/ATM = 2.00 ~ 12.00 (Average = 4.85) and Co-Network = 2.00 ~ 17.00 (Average = 9.00) Then Bad	Confidence: 0.926 # of Records: 25. Significance: p < 0.00001
7) If Lump Sum Payment Usage = 4.00 ~ 11.00 (Average = 6.54) and Installment Amount/Usage = ₩ 278,108 ~ ₩ 611,000 (Average = ₩ 415,022) and Cash Service Amount/Usage = ₩ 60,625 ~ ₩ 630,000 (Average = ₩ 377,537) Then Bad	Confidence: 0.923 # of Records: 36. Significance: p < 0.0000001
11) If Age = 19.00 ~ 24.00 (Average = 22.75) and Major Affiliation III Usage = 0.00 ~ 1.00 (Average = 0.58) and Cash Service Amount/Usage = ₩ 20,000 ~ ₩ 642,857 (Average = ₩ 248,595) Then Bad	Confidence: 0.917 # of Records: 22. Significance: p < 0.0001
16) If Co-Network = 2.00 ~ 21.00 (Average = 7.91) and Lump Sum Payment Amount/Usage = ₩ 93,951 ~ ₩ 116,262 (Average = ₩ 101,350) Then Bad	Confidence: 0.909 # of Records: 20. Significance: p < 0.0001
17) If Installment Usage = 4.00 ~ 8.00 (Average = 5.38) and Co-Network = 2.00 ~ 27.00 (Average = 8.00) and Installment Amount/Usage = ₩ 278,108 ~ ₩ 650,000 (Average = ₩ 426,845) Then Bad	Confidence: 0.905 # of Records: 38. Significance: p < 0.0000001
18) If Major Affiliation III Usage = 2.00 and Installment Amount/Usage = ₩ 276,292 ~ ₩ 660,810 (Average = ₩ 412,704) and Cash Service Amount/Usage = ₩ 60,625 ~ ₩ 623,076 (Average = ₩ 385,568) Then Bad	Confidence: 0.903 # of Records: 28. Significance: p < 0.00001
20) If Deposit Balance = ₩ 0 ~ ₩ 900,000 (Average = ₩ 68,755) and Loan Balance = ₩ 314,045 ~ ₩ 2,539,000 (Average = ₩ 1,343,201) Then Bad	Confidence: 0.897 # of Records: 26. Significance: p < 0.00001

In confidence top 10 rules, installment usage, lump sum payment usage, and installment amount/usage are used most frequently. Also, age ranged only from 19 to 24, installment usage from 4 to 8, and lump sum payment usage from 4 to 11. The frequency of variable usage in confidence top 10 rules is summarized in Table 6.

Table 6. Usage frequency of confidence top 10 rules

Variable	Frequency	Variable	Frequency
Age	2	Sex	1
Installment Usage	2	Installment Amount/Usage	4
Lump Sum Payment Usage	4	Lump Sum Amount/Usage	1
Cash Service Channel (Co-Network)	3	Cash Service Channel(CD/ATM)	1
# of Other Firm's Cards	1	Major Affiliation III Usage	3
Cash Service Amount/Usage	4	Deposit Balance / Loan Balance	1

Association rule validation test

As previously described, this paper used *template using statistical parameter* and *rule syntax template* as a rule validation. First, we used support 2% (20 cases in one rule) and confidence 80% as statistical parameters. Second, trust balance, revolving usage, revolving amount, PC banking, ARS, etc. are used as rule syntax template to eliminate rules including variables having no value in advance.

5. Conclusion

This study performed the research on credit card delinquency of bank customer as a preliminary step for building effective credit scoring system. The result of this study can be a standard of estimating customer's good or bad credit status and basic component of early warning system of default or overdue in various financial products.

For this research, association rules method is used to generate rules from large databases. In association rules, sets of rules classified good or bad credit status. The sets of rules might act as an estimator of good or bad credit status classifier of new customers in the future.

To summarize the result of this paper is as follows. First, total number of rules is 130

through association rule generating program, 27 rules are eliminated through rule validation process, and final 103 rules are selected. Second, from the confidence top 10 rules indicates that variables included in rules are related to cash service and installment service. There are no rules which disobeyed completeness, but several rules are discarded because of inconsistency. Association rules generate 127 complete rules and 27 rules are eliminated or discarded.

In this paper, rule validation process was performed for optimal sets of rules. Among generated rules, a lot of rules are object to completeness and inconsistency. For minimize these rules, post-analysis of rule validation process was conducted by template-based rule filtering method.

Unlike a transactional database normally used in association rule mining that does not have many associations, classification data tends to contain a huge number of associations. In addition, classification datasets often contain many continuous (or numeric) attributes. Mining of association rules with continuous attributes is still a major research issue (Srikant and Agrawal, 1995; Liu et al., 1997, 1998; Yoda et al., 1997; Wang et al., 1998).

This paper also contains various continuous variables in customer profile databases. Therefore, it must be needed to involve discretizing continuous attributes based on the classification pre-determined class target, generate all the class association rules, and build a classifier based on the generated class association rules. In the future, we will propose the method of integrating classification and association rule mining.

References

[1] Adomavicius, G. and A. Tuzhilin, "Expert-Driven Validation of Rule-Based User Models in Personalization Application," *Data Mining and Knowledge Discovery*, Vol.5(2001), 33-58.

[2] Adomavicius, G. and A. Tuzhilin, A., "User Profiling in Personalization Applications through Rule Discovery and Validation," In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999.

[3] Aggarwal, C. and P. Yu, "Online Generation of Association Rules," *ICDE-98*, 1998, 402-411.

[4] Agrawal, R. and R. Srikant, "Fast Algorithms for Mining Association Rules," *Proceedings of the 20th International Conference on VLDB*, IBM Almaden Research Center, 1994, 478-499.

[5] Agrawal, R., H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Fast Discovery of Association Rules," *Advances in Knowledge Discovery and Data Mining*, Chapter 12, 1996, AAAI/MIT Press.

[6] Agrawal, R., T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *Proceedings of the 1993 ACM SIGMOD conferences*, IBM Almaden Research Center, 1993, 207-216.

[7] Anderson, T. W., *An Introduction to Multivariate Statistical Analysis*, New York, NY: Wiley, 1984.

[8] Brachman, R. J. and T. Anand, "The Process of Knowledge Discovery in Databases: A Human-centered Approach," In *Advances in Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park, CA, Ch. 2, 1996.

[9] Breiman, L., J. H. Freidman, R. Olson, and C. Stone, *Classification and Regression Trees*, Wadsworth Publishers, 1997.

[10] Brill, J., "The Importance of Credit Scoring Models in Improving Cash Flow and Collections," *Business Credit*, Vol.1(1998), 16-17.

[11] Brin, S., R. Motwani, J. Ullman, and S. Tsur, "Dynamics Internet Counting and Implication Rules for Market Basket Data," *SIGMOD-97*, 1997, 255-264.

[12] Bryant, S. M., "A Case-based Reasoning Approach to Bankruptcy Prediction Modeling," *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol.6, No.3(1997), 195-214.

[13] Buta, P., "Mining for Financial Knowledge with CBR," *AI Expert*, Vol.9, No.2(1994), 34-41.

[14] Chen, M-C. and S.-H. Huang, "Credit Scoring and Rejected Instances Reassigning through Evolutionary Computation Techniques," *Expert Systems with Applications*, Vol.24(2003), 433-441.

[15] Chiang, D.-A., Wang, Y.-F., Lee, S.-L., and Lin, C.-J., "Goal-Oriented Sequential Pattern for Network Banking Churn Analysis," *Expert Systems with Applications*, forthcoming.

[16] Coakley, J. R. and C. E. Brown, C. E., "Artificial Neural Networks in Accounting and Finance: Modeling Issues," *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol.9, No.2(2000), 119-144.

[17] Curram, S. P. and J. Mingers, "Neural Networks, Decision Tree Induction and Discriminant Analysis: An Empirical Comparison," *Journal of Operational Research Society*, Vol.45, No.4(1994), 440-450.

[18] Davis R. H., D. B. Edelman, and A. J. Gammerman, "Machine Learning Algorithms for Credit-Card Applications," *IMA Journal of Mathematics Applied in Business and Industry*, Vol.4(1992), 43-51.

[19] Desai, V. S., J. N. Conway, and G. A. Overstreet Jr., "Credit Scoring Models in the Credit Union Environment Using Neural Networks and Genetic Algorithms," *IMA Journal of Mathematics Applied in Business and Industry*, Vol.8(1997), 324-346.

[20] Desai, V. S., J. N. Crook, and G. A. Overstreet Jr., "A Comparison of Neural Networks and

Linear Scoring Models in the Credit Union Environment," *European Journal of Operational Research*, Vol.95(1996), 24-37.

[21] Dillon, W. R. and M. Goldstein, *Multivariate Analysis Methods and Applications*, New York: Wiley, 1984.

[22] Dimitras, A. I., S. H. Zanakis, and C. Zopounidis, C., "A Survey of Business Failure with an Emphasis on Prediction Methods and Industrial Applications," *European Journal of Operational Research*, Vol.90, No.3(1996), 487-513.

[23] Emel, A. B., M. Oral, A. Reisman, and R. Yolalan, "A Credit Scoring Approach for the Commercial Banking Sector," *Socio-Economic Planning Sciences*, Vol.27(2003), 103-123.

[24] Falbo, P., "Credit Scoring by Enlarged Discriminant Analysis," *OMEGA*, Vol.19, No.4(1991), 275-289.

[25] Fayyad, U., G. Piatesky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, 1996, 37-54.

[26] Frydman H. E., E. I. Altman, and D. Kao, "Introducing Recursive Partitioning for Financial Classification: the case of Financial Distress," *Journal of Finance*, Vol.40, No.1(1985), 269-291.

[27] Han, J. and Y. Fu, "Discovery of Multiple-level Association Rules from Large Databases," *VLDB-95*.

[28] Hand, D. J. *Discrimination and Classification*, New York, NY: Wiley, 1981.

[29] Jensen, H. L., "Using Neural Networks for Credit Scoring," *Managerial Finance*, Vol.18(1992), 15-26.

[30] Johnson, R. A. and D. W. Wichern, *Applied Multivariate Statistical Analysis (Fourth Edition)*, Upper Saddle River, NJ: Prentice-Hall, 1998.

[31] Klemettinen, M., H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo, "Finding Interesting Rules from Large Sets of Discovered Association Rules," *In Proceedings of the Third International Conference on Information and Knowledge Management*, 1994.

[32] Lee, Y., B. G. Buchanan, and J. M. Aronis, J.M., "Knowledge-based Learning in Exploratory Science: Learning Rules to Predict Rodent Carcinogenicity," *Machine Learning*, Vol.30(1998), 217-240.

[33] Lee, G., T. K. Sung, and N. Chang, N., "Dynamics of Modeling in Data Mining: Interpretive Approach to Bankruptcy Prediction," *Journal of Management Information Systems*, 16(1999), 63-85.

[34] Lee, H., H. Jo, and I. Han, I., "Bankruptcy Prediction Using Case-based Reasoning, Neural Networks, and Discriminant Analysis," *Expert Systems With Applications*, Vol.13(1997), 97-108.

[35] Lee, T.-S., C.-C. Chiu, C.-J. Lu, and I-F. Chen, "Credit Scoring Using Hybrid Neural Discriminant Technique," *Expert Systems with Applications*, Vol.23(2002), 245-254.

[36] Lent, B., A. N. Swami, and J. Widom, "Clustering Association Rule," *Proceedings of the 13th International Conference on Data Engineering*, 1997, 3-5.

[37] Liu, B. and W. Hsu, "Post-analysis of Learned Rules," AAAI-96, 1996, 828-834.

[38] Liu, B., W. Hsu, and Y. Ma, "Using General Impressions to Analyze Discovered Classification Rules," In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, 1997.

[39] Liu, B., W. Hsu, and Y. Ma, "Integrating Classification and Association Rule Mining," In Proceedings of the *Fourth International Conference on Knowledge Discovery and Data Mining*, 1998.

[40] Liu, B., W. Hsu, and Y. Ma, *Mining Association Rules with Multiple Minimum Supports*, In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1999.

[41] Liu, B., W. Hsu, and Y. Ma, "Pruning and Summarizing the Discovered Associations," *KDD-99*, 1999.

[42] Lyn, C. T., B. D. Edelman, and J. N. Crook, *Credit Scoring and Its Applications, Society for Industrial and Applied Mathematics*, Siam, 2002.

[43] Malhotra, R. and D. K. Malhotra, D. K., "Differentiating Between Good Credits and Bad Credits Using Neuro-fuzzy Systems," *European Journal of Operational Research*, Vol.136, No.2(2002), 190-211.

[44] Markham, I. S. and C. T. Ragsdale, C. T., "Combining Neural Networks and Statistical Predictions to Solve the Classification Problem in Discriminant Analysis," *Decision Sciences*, Vol.26, No.2(1995), 229-242.

[45] Martin, D., "Early Warning of Bank Failure: A Logit Regression Approach," *Journal of Banking and Finance*, Vol.1(1997), 249-276.

[46] Mays, E., *Handbook of Credit Scoring*, American Management Association, 2000.

[47] Morrison, D. F., *Multivariate Statistical Methods*, New York, NY: McGraw-Hill, 1990.

[48] Ng, R. T., L. Lakshmanan, and J. Han, "Exploratory Mining and Pruning Optimizations of Constrained Association Rules," *SIGMOD-98*, 1998.

[49] Park, J. S., M. S. Chen, and P. S. Yu, "An Effective Hash Based Algorithm for Mining Association Rules," *SIGMOD-95*, 1995, 175-186.

[50] Piatestsky-Shapiro, G. and C. Matheus, "The Interestingness of Deviations," *Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Database*, 1994, 3-10.

[51] Piatestsky-Shapiro, G., C. Matheus, P. Smyth, and R. Uthurusamy, "Progress and Challenge...," *AI Magazine*, Fall 1994, 77-87.

[52] Piramuthu, S., "Financial Credit-risk Evaluation with Neural and Neurofuzzy Systems," *European Journal of Operational Research*, Vol.112(1999), 310-321.

[53] Provost, F. and D. Jensen, D., "Evaluating Knowledge Discovery and Data Mining," In *Tutorial for the Fourth International Conference on Knowledge Discovery and Data Mining*, 1998

[54] Quinlan, J., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.

[55] Rastogi, R. and K. Shim, "Mining Optimized Association Rules with Categorical and Numeric Attributes," *ICDE-98.*

[56] Reichert, A. K., C. C. Cho, and G. M. Wagner, "An Examination of the Conceptual Issues Involved in Developing Credit-Scoring Models," *Journal of Business and Economic Statistics*, Vol.1(1983), 101-114.

[57] Sahar, S., "Interestingness via What is not Interesting," In *Proceedings of the Fifth ACMSIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999.

[58] Silberschatz, A. and A. Tuzhilin, "What Makes Patterns Interesting in Knowledge Discovery Systems," *IEEE Transactions on Knowledge and Data Engineering*, Vol.8, No.6(1996), 970-974.

[59] Srikant, R., and R. Agrawal, "Mining Genaralized Association Rules," *Proceedings of the 21st VLDB Conference*, Zurich, Switzerland, IBM Research Report RJ 9963, 1995.

[60] Srikant, R., Q. Vu, and R. Agrawal, "Mining Association Rules with Item Constraints," *Proceedings of the Third International Conference on Knowledge Discovery in Databases and Data Mining*, 1997, 67-73.

[61] Stedman, C., "Data Mining for Fool's Gold," *Computer World*, 1997, 109-111.

[62] Tam, K. Y and M. Y. Kiang, "Managerial Applications of Neural Networks: the Case of Bank Failure Predictions," *Management Science*, Vol.38, No.7(1992), 926-947.

[63] Toivonen, H., M. Klemettien, P. Ronkainen, K. Hatonen, and H. Mannila, "Pruning and Grouping Discovered Association Rules," *ECML-95 Workshop on Statistics*, Machine Learning and Knowledge Discovery in Databases, 1995, 25-52.

[64] Troutt, M. D., A. Rai, and A. Zhang, "The Potential Use of DEA for Credit Applicant Acceptance Systems," *Computers and Operations Research*, Vol.23, No.4(1996), 405-408.

[65] Vellido, A., P. J. G. Lisboa, and J. Vaughan, "Neural Network in Business: A Survey of Applications (1992-1998)," *Expert Systems with Application*, Vol.17, No.1(1999), 51-70.

[66] Wang, K., S. H. W. Tay, and B. Liu, "Interestingness-based Interval Merger for Numeric Association Rules," *Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining*, 1998, 1-14.

[67] West, D., "Neural Network Credit Scoring Models," *Computers & Operations Research*, Vol.27(2000), 1131-1152.

[68] West, D., "Neural Network Credit Scoring Models," *Computers and Operations Research*, Vol.27(2000), 1131-1152.

[69] Wizsoft Inc. (2002), *Wizwhy verision 4 User's Guide.*

[70] Yoda, K., T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama, "Computing Optimized Rectilinear Regions for Association Rules," *KDD-97*, 1997.

[71] Zhang, G. P., "Neural Networks for Classification: A Survey," *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, Vol.30, No.4(2000), 451-462.