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

Young-chan Lee[a] and Soo-il Shin[b]

[a]*Institute for Business research, Sogang University*
1-1 Shinsu-dong, Mapo-ku, Seoul 121-742, South Korea
Tel: +82-2-705-8224, Fax: +82-2-703-8224, E-mail: chanlee@sogang.ac.kr

[b]*MetLife Insurance Co. of Korea Ltd. Marketing Team*
141 Sungwon B/D 7FSamsung-dong, Kangnam-ku, Seoul 135-716, South Korea
Tel: +82-2-3469-9847, Fax: +82-2-3469-9703, E-mail: sooils@metlifekorea.co.kr

Abstract

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 that are rule generating method. Specifically, we generate sets of rules of customers who are in bad credit status because of delinquency by using association rules. We expect that the sets of rules generated by association rules could act as an estimator of good or bad credit status classifier.

Keywords:

Association rules; credit scoring system

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 number of persons who registered in government as a bad credit status are over 2.5 million.

To prevent excess loan or credit and bad credit status, we must build effective credit scoring system in every field of industry. However, most of companies in Korea have weak credit scoring system, and not activated joint ownership of credit information. In fact, the beginning of credit scoring system is analysis of delinquency trend of each individual or industry. Effective analysis of delinquency is starting point of good estimation of credit status in anywhere.

This study performed a research about credit card delinquency of bank customers as a preliminary step for building effective credit scoring system. For this research, we used rule generating method such as association rules. Specifically, we generated sets of rules of customers who are in bad credit status because of delinquency by using association rules. 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.

Literature review

Association rules

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: "30% of transactions that contain beer also contain diapers; 2% of all transactions contain both of these items." Here 30% is called the *confidence* of the rule, and 2% the *support* of the rule. 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. [12].

The definition of association rule is as following: Make $I = \{i_1, i_2, K, i_m\}$ 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 [3].

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 [5].

There are two steps to find out the association rules. First step is to detect the large itemset. Second step is to generate the association rules by utilizing 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 [5].

```
$L_1$={large 1-itemsets};
For ($k$=2; $L_{k-1} \neq \varnothing$ ; $k$++) do begin
    $C_k$ = Apriori-gen ($L_{k-1}$); //New candidates
    For all transactions $t \in D$ do begin
        $C_t$ = subset($C_k$, $t$); //Candidates contained in $t$
        For all candidates $c \in C_t$ do
            $c$.count + +;
    End
    $L_k = \{c \in C_k \mid c.count \geq \text{min sup}\}$
End
Answer = $\cup_k L_k$ ;
```

Figure 1. Apriori algorithm

As for the exploration of the association rules, many researchers take the Apriori algorithm [3] supported by Agrawal et al. [1] as the basic formulation. Figure 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.

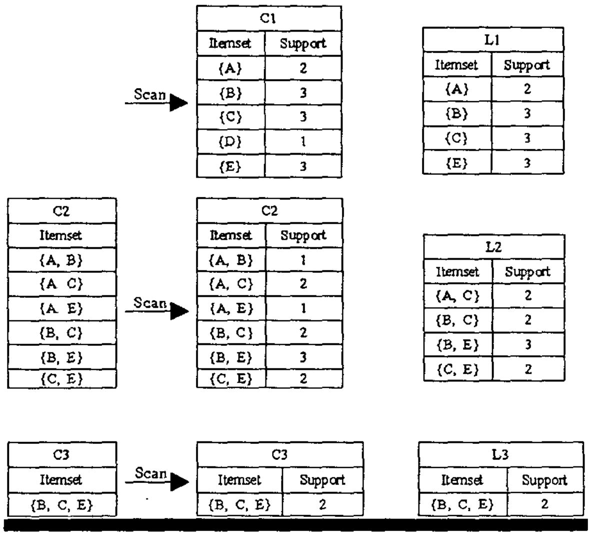
```
Apriori-gen( )
{
//Join step
Insert into $C_k$
Select $p$.item$_1$, $p$.item$_2$, …, $p$.item$_{k-1}$, $q$.item$_{k-1}$
From $L_{k-1}$ $p$, $L_{k-1}$ $q$
Where $p$.item$_1$ = $q$.item$_1$, …, $p$.item$_{k-2}$ = $q$.item$_{k-2}$,
      $p$.item$_{k-1}$ < $q$.item$_{k-1}$;
//Prune step
For itemsets $c \in C_k$ do
    For all ($k$-1)-subsets $s$ of $c$ do
        If ( $s \notin L_{k-1}$ ) then
            Delete $c$ from $C_k$;}
```

Figure 2. Apriori-gen function

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.

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

Figure 3. Original transaction database



C1

Itemset	Support
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L1

Itemset	Support
{A}	2
{B}	3
{C}	3
{E}	3

C2

Itemset
{A, B}
{A C}
{A E}
{B, C}
{B, E}
{C, E}

C2

Itemset	Support
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

L2

Itemset	Support
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C3

Itemset
{B, C, E}

C3

Itemset	Support
{B, C, E}	2

L3

Itemset	Support
{B, C, E}	2

Figure 4. Generation of candidate itemsets and large itemsets

For example, in Figure 3, database D is the original transaction database. We assume that minimum support is 2 transactions in Figure 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 x L_1. Further, calculate the support of C_2 to create L_2. From L_2 x 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 x 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 Figure 4 [5].

Rule validation

Generally, rules are generated from various data mining algorithms. For example, association rules are generated from Apriori algorithm [3] and classification rules are generated from CART algorithm [4]. When generating rules, the critical problem is generating too much rules including insiginificant, trivial, or unreachable rules [10]. Therefore, rule validation assessment has been recognized to be an important issue.

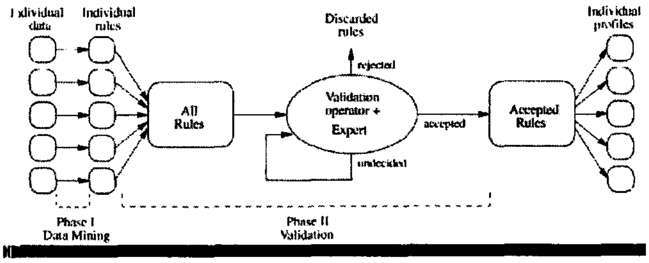


Figure 5. Rule validation assessment process

The process of rule validation assessment is performed on all rules generated by specific validation assessment method. The overall process of validation assessment is represented in Figure 5 [7]. Rule validation assessment is broadly divided into two methods such as *similarity based rule grouping* and *template-based rule filtering*. If we can know the patterns of generated rules by past experience, template-based rule filtering is used. This paper used *template-based rule filtering* to assess the validation of generated rules [7, 8, 14, 15].

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 overdue days. Therefore, if overdue 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 20 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*. 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.

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

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

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 (평균 = ₩ 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 = ₩ 50,625 ~ ₩ 623,076 (Average = ₩ 385,568) Then Bad	Confidence: 0.903 # of Records: 28. Significance: p < 0.00001
2 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

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.

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 20 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.

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 classifier.

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.

In this study, 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, pre-rule validation procedure was conducted by various methods. In this study, template-based rule filtering method is used.

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.

References

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

[2] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A.I. (1995), "Fast Discovery of Association Rules," *Advances in Knowledge Discovery and Data Mining*, Chapter 12, AAAI/MIT Press.

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

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

[5] 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.

[6] Mays, E. (2000), *Handbook of Credit Scoring*, American Management Association.

[7] Gediminas A. and Tuzhilin, A. (2001), "Expert-Driven Validation of Rule-Based User Models in Personalization Application," *Data Mining and Knowledge Discovery*, pp. 33-58.

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

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

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

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

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

[13] Stedman, C., Data mining for fool's gold, Computer world, 1997, pp. 109-111.

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

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

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