International Journal of Advances in Computer Science and its Applications – IJCSIA
Volume 4 : Issue 1 *[ISSN 2250 – 3765]*
Publication Date : 09 January 2014

Predicting personal credit ratings using ubiquitous data mining

Jae Kwon Bae

***Abstract*—Ubiquitous data mining (UDM) is a methodology for creating new knowledge by building an integrated financial database in a ubiquitous computing environment, extracting useful rules by using diverse rule-extraction-based data mining techniques, and combining these rules. In this study, we built six credit rating forecasting models using traditional statistical methods (i.e., logistic regression and Bayesian networks), multilayer perceptron (i.e., MLP), classification tree algorithms (i.e., C5.0), neural network rule extraction algorithms (i.e., NeuroRule), and UDM in order to predict personal credit ratings. To verify the feasibility and effectiveness of UDM, credit ratings and credit loan data provided by *A* Financial Group in Korea were used in this study. Empirical results indicated that UDM outperforms other single traditional classifiers such as logistic regression, neural networks, frequency matrix, C5.0, and NeuroRule. UDM always outperforms other single classifiers in credit rating forecasting; it can predict future personal credit ratings more accurately than any other single classifier.**

***Keywords*—Ubiquitous data mining (UDM), Ubiquitous computing environment, Credit rating forecasting models, Rule extraction algorithms, Integrated financial database.**

I. Introduction

The U.S. subprime mortgage crisis in 2007 resulted from unreliable credit ratings assigned to mortgage-related securities (i.e., mortgage-backed securities) by world-renowned credit rating agencies with vested interests. Consequently, credit rating evaluation (forecasting) has emerged as a critical issue in global finance. Numerous classification techniques have been adopted for forecasting credit ratings, including traditional statistical methods (i.e., multiple discriminant analysis, logistic regression, and Bayesian networks), decision tree methods (i.e., CHAID, CART, QUEST, and C5.0), and neural network analysis (i.e., BPN and MLP). Desai et al. (1996) investigated the accuracy of credit rating forecasting models by using the personal loan information of three U.S. credit unions. They empirically compared the performance of various data mining techniques such as logistic regression, multiple discriminant analysis, and neural network analysis. Malhotra and Malhotra (2002) used the loan information of U.S. credit unions to develop a personal credit rating forecasting model by using an adaptive neuro-fuzzy inference system and back-propagation network.

Jae Kwon Bae

Department of Management Information Systems, Keimyung University
1095 Dalgubeoldaero, Dalseo-gu, Daegu 704-701, Republic of Korea
jkbae99@kmu.ac.kr

Their model exhibited better performance than traditional statistical models. Chung and Suh (2009) adopted classification models to categorize customers into three types of credit card delinquents (i.e., good, bad, and potentially good) by using neural networks and decision trees. They constructed neural networks and decision tree models to estimate the utility of individual credit card delinquency and demonstrated that the classification model with the best hit rate does not necessarily result in the best utility value. Huang (2011) developed reliable credit rating forecasting models using Gaussian-process-based classifiers inspired by artificial intelligence. Empirical results indicated that Gaussian-process-based classifiers outperform other statistical and artificial intelligence techniques. Bae and Kim (2011) investigated the feasibility and effectiveness of knowledge consolidation models as multiple classifiers for credit rating forecasting. Such models combine rules extracted using neural network rule extraction algorithms, Naïve Bayesian classifiers, and C5.0. Moreover, they are superior to single classifiers based on multiple discriminant analysis, logistic regression, frequency matrices, neural networks, and decision trees.

Studies on credit rating forecasting have been conducted extensively. However, financial institutions do not actively adopt the diverse credit rating forecasting models presented in these studies for real-world applications because of their poor accuracy in such scenarios. Previous studies have identified and attempted to address the structural drawbacks of credit rating forecasting models, such as the problem of setting parameters inherent to the algorithms and the infeasibility of statistical assumptions. To the best of our knowledge, the structural problems of financial institutions, which can contribute to the poor classification accuracy of credit rating forecasting models, have not been considered thus far. Therefore, this study focuses on such problems.

Financial Group in Korea has various subsidiaries such as banking, credit cards, securities, life insurance, real estate trust, venture capital, credit information, and data systems. These subsidiaries have been building independent financial databases (DBs) and decision-making models for solutions to financial decision-making problems. Some financial firms integrate financial DBs that can be efficiently used by their subsidiaries, but most subsidiaries build their own independent financial DBs. Consequently, data are duplicated and inconsistent, resulting in poor classification accuracy of decision-making models developed to solve financial decision-making problems. In order to facilitate decision making, the financial DBs of all the subsidiaries should be integrated, and the financial data scattered among them should be stored and managed in the integrated DB. For this purpose, it is essential to build a ubiquitous computing environment among subsidiary companies. Accordingly, this study attempts to present a ubiquitous data mining (UDM)-based credit rating forecasting model that integrates the financial data of subsidiary companies, and the integrated DB is used to solve financial decision-making problems. Here, UDM denotes utilizing data mining techniques to derive useful rules and knowledge from the vast amount of data accumulated in a ubiquitous computing environment.

The objectives of this study are as follows:

(1) to develop UDM for credit rating forecasting that combines rules extracted using Bayesian networks (frequency matrix), decision trees (C5.0), and neural networks (NeuroRule). UDM simplifies the process of knowledge acquisition and refines the initial domain knowledge. In addition, it provides reasoning and expression capability, and can support mutual cross-referencing and verification;

(2) in order to verify the feasibility and effectiveness of the proposed UDM, we use the personal credit ratings provided by *A* Financial Group in Seoul, Republic of Korea. After building a credit rating forecasting model with a dataset of personal credit ratings using logistic regression, neural networks, frequency matrix, C5.0, and NeuroRule, the forecasting capabilities of the traditional single classifiers and those of UDM were compared.

II. Ubiquitous Data Mining

UDM is a methodology for creating new knowledge by building an integrated financial DB in a ubiquitous computing environment, extracting useful rules by using diverse rule-extraction-based data mining techniques (i.e., frequency matrices, decision trees, neural network rule extraction algorithms), and combining these rules. UDM can effectively integrate multiple rule sets into a single centralized knowledge base. Figure 1 shows the conceptual model of UDM. The cumulative rules from rule extraction algorithms can improve the overall performance, as they can reduce the error terms and increase R-square for financial decision-making. The key idea of UDM is to combine a number of classifiers such that the resulting integrated system achieves higher classification accuracy and efficiency than the original individual classifiers. The objective of UDM is to design a composite system that outperforms any individual classifier by pooling together the decisions of all classifiers. UDM exhibits better performance than single classifiers because it consolidates knowledge from single classifiers. Another advantage of UDM is that it does not require memory space for storing datasets, as only extracted knowledge is necessary to build it. Furthermore, it can reduce storage allocation, memory, and time schedule costs.

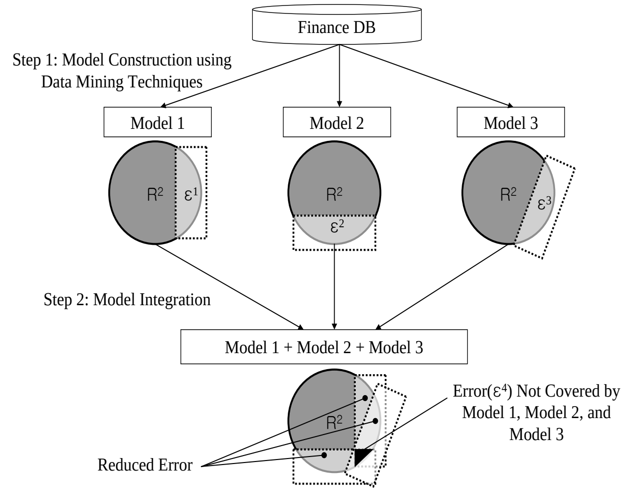

Figure 1. The conceptual model of UDM.

As shown in Figure 2, the proposed UDM framework consists of three stages. Details of the UDM development process are described below.

Step 1: Model construction using logistic regression, Bayesian networks, neural networks, and decision trees

In the first stage, credit rating forecasting models are built using four data mining techniques (i.e., logistic regression, Bayesian networks, neural networks, and decision trees) based on an integrated financial DB in a ubiquitous computing environment. In this study, we used a stepwise logistic regression method to build a logistic-regression-based credit rating forecasting model, and we used a case-based learning (CBL) algorithm (i.e., frequency matrix), which is a conditional independency-based algorithm, to build a Bayesian-network-based credit rating forecasting model. In the case of neural networks, we established a standard three-layered back-propagation network. In MLP, we varied the number of nodes in the hidden layer and the stopping criteria for training. In particular, *n/2*, *n*, *2n*, *3n*, and *4n* hidden nodes are used for each stopping criterion because MLP does not have a general rule for determining the optimal number of hidden nodes (Hornik, 1991). For the stopping criteria of MLP, we allow 100, 200, and 300 learning epochs per training example because there is little general knowledge for selecting the number of epochs. The learning rate is set to 0.1 and the momentum term is set to 0.7. The hidden nodes and the output node use the *hyperbolic tangent* transfer function. In the case of decision trees, we used the measure of *entropy index* (i.e., C5.0), which is used for categorical target variables.

Step 2: Rule extraction using frequency matrix, C5.0, and NeuroRule

All knowledge sources are represented by rules, because almost all knowledge derived by knowledge acquisition tools, or induced by machine learning algorithms, may be easily translated into or represented by rules. Multiple rules are extracted from a frequency matrix and the C5.0 algorithm, and they are sequentially accumulated in the knowledge base. In Figure 2, rules extracted from the frequency matrix are denoted by "R_F," whereas those extracted from C5.0 are denoted by "R_C." Moreover, multiple rules are extracted from neural networks using the NeuroRule algorithm. First, trained neural network models for credit rating forecasting were built, and a simple model was developed by pruning the link weights, which were unnecessary and would hence not affect the results. After discretizing the activation values of the hidden units of the trained neural network by pruning and clustering them in a meaningful section, rules regarding groups and the forecast result of each hidden unit were sought. Finally, we fixed the rules for credit rating forecasting that could be explained in terms of input variables by combining the rules regarding hidden units and the forecasting result obtained by determining the rules that explain the relationship between the input and the hidden units. In Figure 2, rules extracted using the NeuroRule algorithm are denoted by "R_N".

Step 3: Rule (knowledge) integration of frequency matrix, C5.0, and NeuroRule for UDM

UDM is a methodology developed in accordance with rules derived from frequency matrix, C5.0, and NeuroRule with training datasets. By combining separate knowledge in the form of *If-then* rules induced from the datasets, it builds a meta-model. The rule sets R_F, R_C, and R_N in step 3 are merged into a cumulative rule set (R_F + R_C + R_N) as an integrated model shown in Figure 2. Each rule in this model serves as an agent. UDM can effectively integrate multiple rule sets into a single centralized knowledge base.

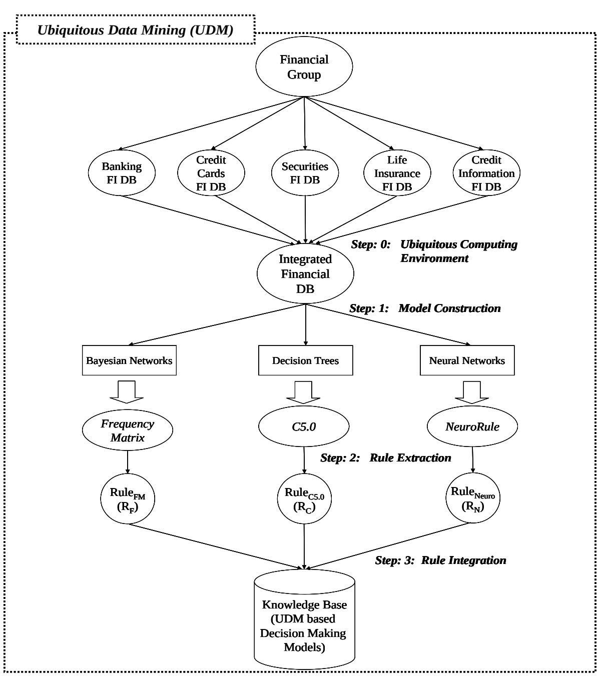

Figure 2. The conceptual model of UDM.

III. Experimental Design

The raw data used in our experiment comprises credit ratings and credit loan data acquired from *A* Financial Group, Korea's leading financial institution, with the largest customer base and the most extensive network in the country. With total assets exceeding US$ 330 billion, it wields considerable capital influence and enjoys strong brand loyalty. As of June 2013, the Group has eight domestic subsidiaries (banking, credit cards, securities, life insurance, real estate trust, venture capital, credit information, and data system) and four overseas subsidiaries in Hong Kong, England, Cambodia, and Kazakhstan.

The data used herein comprises loan transactions of 8234 customers, from January 2004 to December 2008, and includes various types of information used by managers to make decisions regarding loan approval for customers, including customer profile information, loan information, and credit information. Typically, the number of bad customers (customers with bad credit histories) is much smaller than the number of good customers. This inevitably results in zero learning from data mining techniques. To compensate for this asymmetric distribution, we draw two samples of equal size−one each from the two subgroups (bad and good customers). The data consist of 4117 bad customers and 4117 good customers. The variables are adjusted to follow standard normal distribution, as this helps reduce measurement errors. Of 28 variables in total, 12 variables are selected via preliminary screening in the form of two-sample *t*-tests, and the remaining 4 variables are finally selected by stepwise logistic regression. The variables finally selected include the average balance for last six months, days in arrears, cash dispenser amount, and other cash dispenser total amounts.

Each dataset is split into two subsets: a training set and a validation (holdout) set. The training subset is used to train the prediction models, whereas the validation subset is used to test the model's prediction performance for data that have not been used to develop the classification models. Both the training subset (70% of the larger dataset, with 5764 customers) and the validation subset (30% of the larger dataset, with 2470 customers) are randomly selected. We replicate the entire process of data selection, estimation, and testing five times (we refer to the five replications as datasets 1-5) in order to reduce the impact of random variation in the dataset composition (Weiss & Kulikowski, 1991). Cross-validation, a well known method, is applied to enhance the generalizability of the test results. For this purpose, five independent datasets are created wherein 5764 out of 8234 customers (approximately 70%) are randomly selected as the training dataset, and the rest are used for the validation dataset.

IV. Experimental Results

Table 1 shows the results of a *comprehensibility* analysis of the rules extracted using a frequency matrix, C5.0, NeuroRule, and UDM. In the analysis, the rules were compared in terms of the number of extracted rules and average number of rules. *Comprehensibility* is measured both by characterizing whole sets of rules and by examining individual rules. As a measure, *comprehensibility* is at least two dimensional. The first dimension along which we characterize the *comprehensibility* of sets of rules is rule-set size. Size is a concern because sets with a large number of rules can be difficult, if not impossible, to understand. From the five datasets (datasets 1-5), 35 rules were extracted using frequency matrix, 82 rules were extracted using C5.0, 41 rules were extracted using NeuroRule, and 158 rules were extracted using UDM. The average number of rules for the frequency matrix was found to be 7, representing the least number of rules extracted in the single models, and the average number of rules for C5.0 was found to be 16.4, representing the highest number of extracted rules. UDM is an accumulation of the rules extracted from the three above-mentioned rule extraction algorithms, and the average number of rules was found to be 31.6. The greater the number of accumulated rules, the more comprehensive is the credit rating forecasting model.

Table 2 compares the prediction performance of logistic regression, neural networks, frequency matrix, C5.0, NeuroRule, and UDM using fivefold cross-validation. We can evaluate the prediction performance using the accuracy rate (also referred to as hit ratio), which is calculated by dividing the number of correct predictions by the total number of predictions. The main purpose of this cross-validation procedure is to obtain the average accuracy rates for all iterations in the five sets (five iterations per set). Fivefold cross-validation is employed to enhance the generalizability of the test results (Zhang, Hu, Patuwo, & Indro, 1999). Among these models, UDM shows the highest average accuracy of 78.67% with the given validation sets, followed by NN with 72.75% and NeuroRule with 70.04%. The results from the tests show that the performance of UDM is superior to that of the other single classifiers such as logistic regression, neural networks, frequency matrix, C5.0, and NeuroRule. UDM always outperforms other single classifiers in credit rating forecasting; it can predict future personal credit ratings more accurately than any other single classifier.

TABLE 1. THE NUMBER OF RULES EXTRACTED FROM TRAINING DATASETS

Set no.	FM	C5.0	NeuroRule	UDM[a]
Data Set 1	5	16	7	28
Data Set 2	7	17	7	31
Data Set 3	8	16	8	32
Data Set 4	8	18	10	36
Data Set 5	7	15	9	31
Sum	35	82	41	158
Avg.	7.0	16.4	8.2	31.6

Note: FM (Frequency Matrix), NeuroRule (Neural network Rule extraction)

[a] UDM (Ubiquitous Data Mining): FM + C5.0 + NeuroRule.

V. Conclusions

Personal credit ratings provide important information on credit risk for banks or investors in financial markets. Moreover, credit rating forecasting influences important managerial decisions affecting financing and firm value. In this study, we built six credit rating forecasting models using traditional statistical methods, MLP, classification tree algorithms, neural network rule extraction algorithms, and UDM in order to predict personal credit ratings.

TABLE 2. COMPARISON OF PREDICTION MODELS

Data Set no.	Result	LR	NN	FM	C5.0	NeuroRule	UDM
Data Set 1	Training	70.69	72.22	69.00	71.79	70.06	81.02
	Validation	66.52	71.38	69.35	69.55	68.99	80.65
Data Set 2	Training	72.41	72.62	70.85	72.95	69.78	76.04
	Validation	70.53	72.19	68.87	68.42	66.56	74.53
Data Set 3	Training	69.83	73.11	72.22	72.05	71.34	77.20
	Validation	70.20	72.71	69.27	69.19	72.23	77.65
Data Set 4	Training	72.41	75.62	71.95	72.41	73.47	81.59
	Validation	69.88	75.06	69.15	67.37	70.61	81.21
Data Set 5	Training	69.83	73.06	71.48	70.35	72.31	78.45
	Validation	68.14	72.39	69.31	68.91	71.82	79.31
Avg.	Training	71.03	73.33	71.10	71.91	71.39	78.86
	Validation	69.05	72.75	69.19	68.69	70.04	78.67

Note: LR (Logistic Regression), NN (Neural Networks), FM (Frequency Matrix), NeuroRule (Neural network Rule extraction), UDM (Ubiquitous Data Mining): FM + C5.0 + NeuroRule.

To verify the feasibility and effectiveness of UDM, credit ratings and credit loan data provided by *A* Financial Group in Korea were used in this study. The performance of UDM was examined using a dataset comprising a large amount of personal financial information. The results of the tests showed that the performance of UDM is superior to that of other single traditional classifiers such as logistic regression, neural networks, frequency matrix, C5.0, and NeuroRule. UDM always outperforms other single classifiers in credit rating forecasting; it can predict future personal credit ratings more accurately than any other single classifier. This enhancement in the predictability of personal credit ratings can significantly contribute to the correct credit admission evaluation of loan customers, and hence, domestic and international financial institutions can employ UDM for better lending decision making, which can eventually lead to higher profits and firm values. Moreover, by using UDM, financial institutions can reduce the financial loss caused by potential delinquent borrowers. This study can help financial institutions to assess credit risks that they encounter, thereby reducing their losses substantially.

Our study has the following limitations, which require further investigation. First, the results from the study should be generalized. Our study uses only a single selected dataset for system validation. However, only one dataset may not be reliable for making a conclusion. It is necessary to consider a certain number of different datasets for system validation. We believe that other problem domains (bankruptcy prediction, stock market prediction, dividend policy forecasting, and fraud detection) should be investigated in order to generalize the results of this study. Second, binary dependent variables (good or bad credit histories) were used to address binary classification problems. For lending or investment decision making, however, the construction of a credit rating prediction model for addressing multiclass classification problems, or that for forecasting continuous (numeric) dependent variables, will be more useful in actual practice. Therefore, a future study must establish UDM for a credit rating prediction model that enables forecast of multiclass or continuous dependent variables. Lastly, future research should consider non-financial and macroeconomic variables for UDM inputs.

References

[1] Bae, J. K. and Kim, J., “Combining models from neural networks and inductive learning algorithms,” Expert Systems with Applications, vol. 38, no. 1, pp. 4839-4850, 2011.

[2] Chung, S-H., and Suh, Y., “Estimating the utility value of individual credit card delinquents,” Expert Systems with Applications, vol. 36, no. 2, pp. 3975-3981, 2009.

[3] Desai, V. S., Crook, J. N., and Overstreet, G. A. J., “A comparison of neural networks and linear scoring models in the credit union environment,” European Journal of Operational Research, vol. 95, no. 1, pp. 24-37, 1996.

[4] Hornik, K., “Approximation capabilities of multilayer feedforward networks,” Neural Networks, vol. 4, no. 2, pp. 251-257, 1991.

[5] Huang, S-C., “Using Gaussian process based kernel classifiers for credit rating forecasting,” Expert Systems with Applications, vol. 38, no. 7, pp. 8607-8611, 2011.

[6] Malhotra, R., and Malhotra, D. K., “Differentiating between good credits and bad credits using neuro-fuzzy systems,” European Journal of Operational Research, vol. 136, pp. 190-211, 2002.

[7] Weiss, S. M., and Kulikowski, C. A., Computer systems that learn: Classification and prediction methods from statistics, neural nets, machine learning, and expert systems (1st ed.). CA: Morgan Kaufman, 1991.

[8] Wilcoxon, F., “Individual comparisons by ranking methods,” Biometrics Bulletin, vol. 1, no. 6, pp. 80-83, 1945.

[9] Zhang, G., Hu, M. Y., Patuwo, B. E., and Indro, D. C., “Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis,” European Journal of Operational Research, vol. 116, no. 1, pp. 16-32, 1999.

Corresponding Author: Jae Kwon Bae

Jae Kwon Bae is an assistant professor in the Department of Management Information System, at the Keimyung University in Korea. He received a BA in Management Information Systems from Hannam University, Korea, in 2004 and a MS in Finance and a PhD in Management Information Systems from the Sogang University, Korea in 2006 and 2009. His research interests include knowledge-based systems design and development, neural-net computing, decision support systems, intelligent systems, data mining, and artificial intelligence applications for business and electronic commerce.