International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.5 (2015), pp.17-24
http://dx.doi.org/10.14257/ijmue.2015.10.5.03

Feature Selection in Credit Scoring Model for Credit Card Applicants in XYZ Bank: A Comparative Study

Mediana Aryuni and Evaristus Didik Madyatmadja

School of Information Systems, Bina Nusantara University, Jakarta, Indonesia
mediana.aryuni@binus.ac.id

Abstract

The performance of credit scoring models is determined by the used features. The relevant features for credit scoring usually are determined unsystematic and dominate by arbitrary trial. This paper presents a comparative study of four feature selection methods, which use data mining approach in reducing the feature space. The final results show that among the four feature selection methods, the Gini Index and Information Gain algorithms perform better than others with the classification accuracy of 75.46% and 75.44% respectively.

Keywords: *credit card; credit scoring; bank; feature selection*

1. Introduction

The US mortgage crisis in 2007 [1] and the world financial crisis from 2008 [2] made credit scoring model become major and fascinating area for research.

According to Bellotti and Crook [3] a credit scoring model is the set of decision models and techniques to help lenders in evaluating credit applicants.

Credit scoring models are built using classification algorithms. The problem is to determine which features will be used in classification algorithm. Some of the features may be irrelevant and redundant. The irrelevant and redundant features increase the learning process time of classification algorithm and makes the model become more complex. In addition, the accuracy of model is possibly decreased [4-7].

Some previous researches about feature selection were conducted [3-5, 7-12] to identify the most relevant features. By reducing the irrelevant and redundant features, the accuracy of credit scoring model was improved [3-4, 6-7].

Additionally, some researches about credit scoring model for credit card dataset [3] [13] were done to deal with the growth of credit card usage includes risk increment of bad debts. A credit card applicant does not use security collateral as warranty. So, it is very crucial in determining which credit card applicants to be approved or rejected.

By analyzing the results of previous researches, the objective of this study is to present a comparative study of four feature selection methods for credit card applicants in XYZ Bank. These methods consist of Information Gain, Gain Ratio, GINI Index, and CHI-Squared Statistics in order to determine the proper feature selection methods, which decrease the learning process time and improve the accuracy of the model.

2. Data Mining and Feature Selection

Data mining is the process of knowledge extraction from very large size data [14]. It is also called Knowledge Discovery from Data, or KDD.

Classification is one of data mining functionalities. It finds a model or function that separates classes or data concepts in order to predict the class of an unknown object [14]. For example, a loan officer requires data analysis to determine which loan applicants are "safe" or "risky". The data analysis task is classification, where a model or classifier is

Copyright ⓒ 2015 SERSC

constructed to predict class (categorical) labels, such as "safe" or "risky" for the loan application data. These categories can be represented by discrete values, where the ordering among values has no meaning. Because the class labels of training data is already known, it is also called supervised learning [14].

Classification consist two processes: (1) training and (2) testing. The first process, training, builds a classification model by analyzing training data containing class labels. While the second process, testing, examines a classifier (using testing data) for accuracy (in which case the test data contains the class labels) or its ability to classify unknown objects (records) for prediction [15].

According to Han and Kamber [14], feature selection is dimensionality reduction to reduce the number of random variables or attributes under consideration. Automatic data mining technique is used in feature selection to find a best subset of features, from the original set of features in a given data set [4].

Some feature selection methods use a measure to evaluate the goodness of individual features. Features are ranked according to their values on this measure. The first X features are chosen as the selected feature subset. X is decided according to some domain knowledge or a user-specified threshold value [4].

2.1. Information Gain

Information gain measure is based on pioneering work by Claude Shannon on information theory, which studied the value or "information content" of messages. The expected information is given by [14]:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i), \tag{1}$$

where p_i is the nonzero probability that an arbitrary tuple in D belongs to class C_i and is estimated by $|C_{i,D}|/|D|$. The information is encoded in bits, so it uses a log function to the base 2. Info (D) is also known as the entropy of D [14].

To know the impurity of each attribute, this amount is measured by [14]:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j). \tag{2}$$

The term |Dj|/|D| acts as the weight of the *j*th partition. Info_A(D) is the expected information based on the attribute A. The smaller the expected information (still) required, the greater the purity of the attribute.

Information gain is defined as the difference between the original information requirement (Info(D)) and the new requirement (Info_A(D)) [14]. That is,

$$Gain(A) = Info(D) - Info_A(D). \tag{3}$$

The attribute A with the highest information gain, Gain(A) has the highest weight being the relevant features.

2.2. Gain Ratio

The information gain measure is biased toward tests with many outcomes. That is, it prefers to select attributes having a large number of values [14].

C4.5, a successor of ID3, uses an extension to information gain known as gain ratio, which attempts to overcome this bias [14]. It applies a kind of normalization to information gain using a "split information" value defined analogously with Info (D) as [14]:

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right). \tag{4}$$

The gain ratio is defined as [14]:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)}. \tag{5}$$

The attribute with the maximum gain ratio is selected as the most relevant attribute.

2.3. Gini Index

Gini index [16] observes the decrease of impurity yielded by the use of a particular feature. It is frequently used for estimating feature quality in high dimensional domains, where the number of features incurs less manageable computational complexity of the more powerful feature selection methods.

Gini index is derived from the decrease of impurity where a prior and posterior impurity estimation is approximated using the Gini coefficient (this coefficient is defined using a sum of squared class probabilities). Gini index for a feature A is defined as [16]:

$$\mathrm{Gini}(A) = \sum_{j} p\,.\,j \sum_{k} p^2_{k|j} - \sum_{k} p^2_{k.}, \tag{6}$$

where $p._j$ denotes the probability that feature A takes value j, $p_{k|j}$ probability that a random example from the dataset belongs to class k, its feature A having value j. Symbol $p_{k.}$ denotes the probability that a random example from the dataset belongs to class k.

2.4. Chi-Squared Statistics

The X^2 (CHI-Squared) statistics is defined by the following expression [17]:

$$X^2(t,c) = \frac{N \text{ x } (AD - CB)^2}{(A+C) \text{ x } (B+D) \text{ x } (A+B) \text{ x } (C+D)} \tag{7}$$

where N is the number of documents, A is the number of data of class c containing the feature t, B is the number of data of other class (not c) containing t, C is the number of data of class c not containing the feature t and D is the number of data of other class not containing t.

3. Credit Scoring Applications

Credit scoring is formally defined as a statistical method (or quantitative) which is used to predict the probability of the applicant's credit worthiness [18]. Credit scoring goal is to measure the financial risk of the loan so that the loan provider can make credit lending decisions quickly and objectively.

Credit scoring is not only useful for credit providers, but also for the credit borrowers. For example, credit scoring help reduce discrimination because the credit scoring model provides an objective analysis of the feasibility of the applicant. In addition, the credit providers focus only on information related to credit risk and avoid subjectivity of credit analysts [18].

In the United States, the variables related to discrimination such as race, sex, and age are not included in the credit scoring model. Only information that is not related to discrimination and have proven predictive for the performance of credit payments can be included in the model [18].

Credit scoring also supports to increase the speed and consistency of the credit application process and enables the automation of the credit application process. So, human intervention in the credit evaluation and cost can be reduced. The usage of credit scoring will support the financial institutions to measure the risk associated to lending to the applicant in a short time. In addition, the financial institutions can make better decisions [18].

Peng [18] discussed the advantages and usage of credit scoring as well as the development its model using data mining. Data mining techniques which used for credit scoring models such as logistic regression, neural networks, and decision tree.

Hsieh [19] demonstrated that identifying customers by a behavioral scoring model is helpful to know the characteristics of customer and facilitate marketing strategy development.

Kotsiantis [20] mentioned that credit risk analysis became the main focus on the financial and banking industries. To improve accuracy, the research developed a hybrid method that combined several representative algorithms and then used selective voting methodology.

Kocenda and Vojtek [1] built the two credit risk models based on logistic regression and Classification and Regression Trees (CART) using a retail credit data of banks in Czech Republic.

4. The Proposed Model

The method used in this research was Knowledge Discovery from Data, or KDD [14], which consisted of business understanding, data understanding, data preparation, modeling, and evaluation.

KDD was used to build the proposed model. Figure 1 shows the block diagram of the proposed model. First, four feature selection algorithms were used; Information Gain, Gain Ratio, GINI Index and CHI-Squared Statistics in credit card applicant dataset. The feature selection algorithms were applied by using Rapid Miner Software to choose the relevant features or attributes of each algorithm. Second, it was dimensionality reduction to produce new dataset using only the relevant attributes after feature selection applied. Third, classification algorithm using Naïve Bayes classifier was applied to build credit-scoring model for credit card applicants.

In this research, some parameters; the learning process time and the accuracy of the model before and after feature selection applied were going to be measured. The accuracy was measured by using cross-validation method on the dataset to evaluate the classifier model.

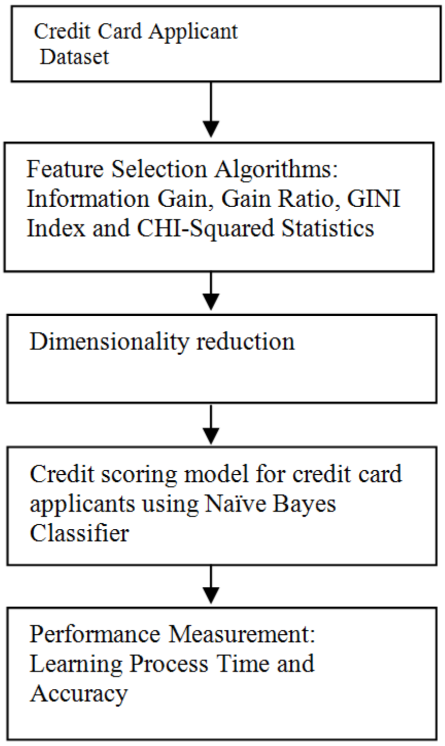
Credit Card Applicant Dataset
Feature Selection Algorithms: Information Gain, Gain Ratio, GINI Index and CHI-Squared Statistics
Dimensionality reduction
Credit scoring model for credit card applicants using Naïve Bayes Classifier
Performance Measurement: Learning Process Time and Accuracy

Figure 1. Block Diagram of the Proposed Model

5. Results

The experimental work was conducted using Rapid Miner Software. The data were obtained from credit card applicant dataset of Bank XYZ in Indonesia. Available records in the dataset were classified into two class labels, 'approved' and 'rejected'. The class label was determined by credit experts' knowledge. There were 4634 records in total, among them 1054 were 'approved' and 3580 were 'rejected'. Each record is described by 18 attributes.

We proposed the block diagram of model as shown in figure 1 to continue our previous research [13] in order to improve the accuracy of credit scoring model after feature selection applied.

The comparison of accuracy is shown in table 1. We can see that the accuracy of model using feature selection is better than the model without feature selection applied. Among the four methods, GINI Index and Information Gain are better both in the dimensionality reduction and in the improvement of model accuracy.

Table 1. The Comparison of Accuracy

Without Feature Selection	Using Feature Selection			
	Information Gain	Gain Ratio	Gini Index	Chi-Squared Statistics
66.83%	75.44%	68.10%	75.46%	74.77%

Based on the Table 1, the accuracy was improved after feature selection applied because we reduced irrelevant and redundant features and only used the relevant ones for the credit scoring model. The highest accuracy was achieved by GINI Index.

6. Conclusions

The comparative study of the feature selection methods to build the credit scoring model for credit card applicants in this paper illustrates how different feature selection methods perform on one real dataset.

Among the four feature selection methods, the GINI Index and Information Gain feature selection methods performed relatively better.

After feature selection applied, the model accuracy was increased. Furthermore, the training time was decreased and the final model became more simple because the reduction in the number of features.

Acknowledgements

Supported by Hibah Penelitan Bina Nusantara University, year 2014.

References

[1] E. Kocenda and M. Vojtek, “Default predictors in retail credit scoring: Evidence from Czech banking data”, William Davidson Institute Working Paper Number, vol. 1015, **(2011**).

[2] A. Dzelihodzic and D. Donko, “Data Mining Techniques for Credit Risk Assessment Task”, Recent Advances in Computer Science and Applications, **(2013)** August 6-8; Valencia, Spain.

[3] T. Bellotti and J. Crook, “Support vector machines for credit scoring and discovery of significant features”, Expert Systems with Applications, vol. 36, no. 2, **(2009)**.

[4] Y. Liu and M. Schumann, “Data mining feature selection for credit scoring models”, Journal of the Operational Research Society, vol. 56, no. 9, **(2005**).

[5] B. Waad, B. M. Ghazi, and L. Mohamed, “Rank aggregation for filter feature selection in credit scoring”, Proceedings of International Conference on Control, Engineering, and Information Technology (CEIT’13), **(2013**) Jun 4-7; Sousse, Tunisia.

[6] J. Wang, A. -R. Hedar, S. Wang, and J. Ma, “Rough set andscatter search metaheuristic based feature selection for credit scoring”, Expert Systems with Applications, vol. 39, no. 6, **(2012)**.

[7] F. L. Chen and F. C. Li, “Combination of feature selection approaches with SVM in credit scoring”, Expert Systems with Applications, vol. 37, no. 7 **(2010)**.

[8] P. E. N. Lutu and A. P. Engelbrecht, “A decision rule-based method for feature selection in predictive data mining”, Expert Systems with Applications, vol. 37, no. 1, **(2010)**.

[9] C. M. Wang and Y. F. Huang, “Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data”, Expert Systems with Applications, vol. 36, no. 3, **(2009)**.

[10] P. Somol, B. Baesens, P. Pudil, and J. Vanthienen, “Filter- versus wrapper-based feature selection for credit scoring”, International Journal of Intelligent Systems, vol. 20, no. 10, **(2005)**.

[11] L. K. Sheng and T. Y. Wah, “A comparative study of data mining techniques in predicting consumers’ credit card risk in banks”, African Journal of Business Management, vol. 5, no. 20, **(2011)**.

[12] J. Li, H. Wei, and W. Hao, “Weight-selected attribute bagging for credit scoring”, Mathematical Problems in Engineering, vol. 2013 **(2013)**.

[13] E. D. Madyatmadja and M. Aryuni, “Comparative Study of Data Mining Model for Credit Card Application Scoring in Bank”, Journal of Theoretical and Applied Information Technology, vol. 59, no. 2, **(2014)**.

[14] J. Han and M. Kamber, “Data mining: concepts and techniques”, Morgan Kaufmann Publishers, San Francisco **(2012)**.

[15] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J. Chang, and L. Hua, “Data Mining in Healthcare and Biomedicine: A Survey of the Literature”, J. Med Syst., vol. 36, no. 4, **(2012)**.

[16] L. Cehovin and Z. Bosnic, “Empirical evaluation of feature selection methods in classification”, Intelligent Data Analysis, vol. 14, no. 3, **(2010)**.

[17] S. R. Singh, H. A. Murthy, and T. A. Gonsalves, “Feature selection for text classification based on gini coefficient of inequality”, Proceedings of the Fourth International Workshop on Feature Selection in Data Mining, **(2010)** June 21; Hyderabad, India.

[18] G. C. Peng, “Credit scoring using data mining techniques”, Journal of Singapore Management Review **(2004)**.

[19] N. C. Hsieh, “An integrated data mining and behavioral scoring model for analyzing bank customers”, Expert Systems with Applications, vol. 27, no. 4. **(2004)**.

Copyright ⓒ 2015 SERSC

[20] S. Kotsiantis, “Credit Risk Analysis using Hybrid Data Mining Model”, International Journal of Intelligent Systems Technologies and Applications, vol. 2, no. 4, **(2007)**.

Authors

Mediana Aryuni, She received the bachelor degree in Informatics from the Sepuluh Nopember Institute of Technology (ITS), Surabaya, Indonesia, in 2004. She received the master degree Informatics from the Sepuluh Nopember Institute of Technology (ITS), Surabaya, Indonesia, in 2006. Currently, she is a lecturer at Bina Nusantara University, Jakarta, Indonesia. Her research interests include data mining and Business Intelligence.

Evaristus Didik Madyatmadja, He received the master degree Computer Science from Gadjah Mada University (UGM), Yogyakarta, Indonesia, in 2005. Currently, he is a lecturer at Bina Nusantara University, Jakarta, Indonesia. His interests are in decision support system, data warehouse, data mining and business intelligence.

Copyright ⓒ 2015 SERSC

Copyright ⓒ 2015 SERSC