5th SASTech Iran | Mashhad May 12 - 17 | 2011
5th Symposium on Advances in Science & Technology

Evaluation of Data Mining Classification Techniques and Performances to Banking Customers Credit Scoring

Narges khalesi, Master Student of Islamic Azad University, Zanjan
narges.khalesi@gmail.com
Amir hoeiyn shokuhi, MS Computer Engineering – Iran university of Science and Technology, shokuhi@comp.iust.ac.ir
Paper Reference Number: 31
Name of the Presenter: Narges khalesi

Abstract

The clarity of the information about Iran's banking system is increased at the result of the globalization of trade and the growing economy of the country's economy and this is playing a significant role in the process of the banks in primary customer. Such a system which assist the banks in reaching their goals will require diversity ground. This system have some problems in Iran's banking system due to the lack of comprehensive data bank of bank's customer that these problems will be reveled in the payment facility to customers or presentation long-term programs about the amount of the facilities that will paid in the future and also this can't considered as a guarantee for receipt of claims. Thus the lack of the full validation and scoring customers in Iranian banks with the aim of transparency and competition in the banking system is one of the main reasons for increased bank demand.

There are many data mining techniques for predicting and validation of bank's customers, that the most famous techniques is considered in this article which included: support vector machines, genetic programming, denotative analysis, logical regression, C4.5, Bagging, Boosting.

KEYWORDS
Banking facilities, Data mining, Validation

1. Introduction

The most important goal of all banks is the collect savings entities and allocate them as a facilities to the industrial, manufacturing and services companies. In other words, directed the financial resources of the country towards economic activity and the increasing employment in the community is the goal of all banks. From the bank view point the customer who is good for the allocation of bank credit that in addition to spending the intake facilities in different economic sectors also it could return the intake facilities to the banking system on time, and therefore not repayment on time of the facilities indicate that facilities receiver have not enjoyed much success in operation of intake facilities. In the other words, they were encountered problems in the time of repayment because the

efficiency of the facilities is the lower than banking interest. Thus, these facilities comes to banking postpone demands. The postpone demands is the issue that is causing bank affected and severe shortage of banking resources in recent years.
The competition market among banks in providing financial services to customer on one hand and the balance between supply and demand on resources and banking facilities on the other hand is led to that management and reduce outstanding claims are reveals as a major issue in the country's banking system, therefore the necessity of implement the ranking system and validation in the country's banking system has been raised as a major issue in previous years. So it is very important in the banking system to create a validation system which can evaluated the customers in reimbursement the facilities before granting facilities to them. This is meaning that the banks grant the advantages to their customers based on the credit indices and ultimately they specify the rating scores of their customer for granting facilities based on these advantages. Thus the importance of granting the facilities in the country's banking and its critical role in economic growth and increased the employment is leading to the development of various models for evaluating the customer's credit whom requesting this facilities, and this issue has been considered by researches in recent years.

Classification plays an important role in financial predictions, discovered the fraud, marketing strategies and such issues [3] [6] [22]. Categorized can be done with the help of different statistical techniques and artificial intelligence techniques. In [27] statistical methods, nonparametric statistical methods and artificial intelligence techniques have been proposed for credit assessment.

Statistical methods including regression, analysis of separate linear [33] and nonlinear models [1] [29], have been used in the validation models. Logistic regression models [9] and is denotative analysis models [2] [17] [27] [29] are two models which have been used in previous years and the major problems of these methods is that they are not suitable for inputs with large and small sample sizes, and in most of these methods assumes that there is a linear relationship between variables, while usually there is a non-linear relationship. Hence, automatic modeling process is difficult. In static models usually when the environment is changing it was doomed to failure and therefore the models may be re-created. Moreover, in recent year the classical models of artificial intelligence have been used in the credit scoring that is included: nearest neighbor [12], neural networks [20] programming genetics [23], decision tree [7] and Support Vector Machine models [26]. Of course encouraging researches has been done about hybrid Data mining in previous years [17] [18] [19].

2. Classification algorithms

In recent years, researchers have consistently followed the model which have accuracy and well performance. This assessment of methods the issue which always has been considered by researchers is that, which subset of variables is selected for predict and the random selection of variables that can increased the accuracy and efficiency of model. In this paper, the most common methods have been evaluated for credit scoring the customers on two-sample ranking bank in terms of accuracy and efficiency of the actual use of classification techniques.

2.1. C4.5 tree

Decision trees were introduced in 1980 and they were used for develop accredit scoring models [8]. However decision trees have restrictions, but they are counted as a power full and flexible tools in category. There are different methods for decision trees that the decision tree method C4.5 is one of them. C4.5 trees is a type of generalized and revised of ID3 tree which was raised by Cheng and Chen in 2008 for credit scoring of bank's customers and it's a standard measure in machine learning. Based on the decision trees, C4.5 occurs theoretical aspects of data [21]. Attribute selection in C4.5 is based on minimizing the scale of information in a node. Each path from the root toward the node represents a category of low.

2.2. Boosting

Boosting [7] [8] [25] a process in which the poor classifiers order are combined with each other to achieve a high performance classification. Boosting algorithm begins with the same weight assigned $w^{(0)}$ to all credit requests. After that a classifier is made where the weight of each request is varied based on a specific classifier. A second classifier is made by using the re-weighting training sample. This process is repeated several times. Final classification of a credit request is in the average weight of all unique categories from all classifiers. Always there are several ways to update weights and combines the unique classifications.

One of the most common algorithms boosting is a Adaboost that was the brief of the adapting boosting and it is an adaptive learning algorithm of the machine. This algorithm was able to solve many problems boosted basic boosted algorithm [8][9] [25].

2.3. Bagging

This method was introduced first by Breiman in 1994. This method is based on machine learning algorithm that combine the samples of training data sets to learning and combination for making ultimate forecasts. This method is a way to reduce variance in made models. Bagging is technique where several examples of the training data set are chosen by replacing for build predicting models, that each of the models varies and bagging were done this with a simple vote of the obtained models from selected samples, and the final section categories, is the categories which obtained by the different samples obtained. Indeed bagging is a automatic condensation method.

2.4. Genetic Programming

In recent years most research done in the field of Credit Rating is based on artificial intelligence techniques like neural networks and genetic programming to credit analyze and these methods usually have highly accurate [23] [29]. Genetic Programming [16] is one of the youngest evolutionary algorithms that introduced in America in 1990 by Koza. The goal of the genetic programming is to resolve the issue without an explicit programming.

General features of genetic programming are like neural networks. One of the advantage of genetic programming is to require to large data collection. More crowded usually cause to goes up the efficient and the speed [4][14].

2.5. *Support Vector Machines*

Support vector machine is the other method in the filed of customers validation which it's highly regarded by researchers in recent years. Support Vector Machines was considered for the first time in 1995 by Vapnik [28].

There is robust mathematical method in classification, a method similar to the neural networks that instead the line separator is following the best line separator that has the maximum margin. It means the best line separator that has a minimal gap with the nearest point. choose the optimal features has a considerable impact on the accuracy of the model in support vector machines, so in previous years many research on support vector machines, are focused on the optimal selection of parameters. To improve performance of classifier is better to select important features and dealing the attributes with less important should be avoided. For this purpose we can used strategies such as grid search and genetic algorithms.

2.6. Logical Regression

Logic regression is a statistical regression model for binary dependent variables. Logic regression is a more general case of linear regression. In the past, this method was used to predict the values of binary or multi-value discrete variables. Since the values are intended to predict discrete values, so it can't be modeled using linear regression, the purpose of this discrete variables are a way to convert numeric and continuous variable and thus is considered likely logarithm value corresponding variable, then can be identified with the inverse ratio of the desired logarithm, and it identified the desired classes [10].

2.7. k-nearest neighbors method

A technique class credit scheme in the Americas since ranking method is used for k close to my neighbors, that version of this method is used. In this method measures decided that a new sample in each category should be examining a number of the k most similar cases or neighbor is done. Cases are considered for each class, and new cases to handle the increasing number of neighbors belong to it will be attributed [15].

In k-nearest neighbors algorithm for classification uses all the record features as the same. On the other hand may not have the same role in the entire record in character classification, and irrelevant attributes cause the near distance records, are diagnosed far apart, and do not correctly classified. This problem is called the scourge of dimensions [31]. The first requires the use of k-nearest neighbors, find a criterion to determine the distance between characters in the data and calculate it.

2.8. Discriminant Analysis

This method is one of the oldest methods in mathematics are grouped data; the first time in 1936 was used by Fisher. This method is the following that review the data as multidimensional data, and data are created between the boundaries (line separator for the two-dimensional data, separator level for three-dimensional data and ...) that the boundaries identified Visitors are of various classes, and then we just have to specify the location to determine the class to new data [30].

The problem is that this method predicted variables are normally distributed, and can't be used for non-numeric data.

3. Practical Results

In classification problems, typically the data samples are divided into two classes training data set and test data set. Used to predict of training data set, and used to evaluate the accuracy predicted of test data set.

Here, in order to better evaluate the accuracy of the classification have been used 10-fold Cross Validation. This technique consists of random data sets in 10 divisions under a unique set of mutually equal sizes and the sequence evaluation that takes place each subset using the classifiers provisions in remaining subset.

4. Data Collection

Data set used in this article, two sets of very conventional set of credit data (UCI) [22], which are shown in Table 1. Dataset of Germany includes 1000 samples of which have used the version number it, and this collection includes 24 sample number, the characteristics of the sample data includes: History of Credit customer, account balance customer, the purpose of obtaining a loan, the loan, job, Personal data, age, job title and home ownership. In the Australian data set names and values have become all the symbols, to maintain data confidentiality. This data set includes 690 samples data. Each sample has 14 features, which are six numerical attributes, and 8 features are qualitative.

Dataset	Class	Number of samples	Features nominal	Numerical features	Total Features
German	2	1000	0	24	24
Australia	2	690	6	8	14

Table 1. General features of two data sets

In Tables 2 and 3 are shown results of evaluate methods in the Australian dataset and German dataset.

Model	Average classification accuracy	Average classification error
Bagging	%85/20	%14/80
Boosting	%84/20	%15/80
Genetic Programming	%87	%13
C4.5	%85/90	%14/10
Support Vector Machines	%85/92	%14/08
Discriminant analysis	%85/94	%14/06

K- nearest neighbor	%69/57	%30/43
logic regression	%73/88	%26/12

Table 2. Comparison of classification accuracy and model error in Australian dataset

Model	Average classification accuracy	Average classification error
Bagging	%74	%26
Boosting	%72/07	%27/92
Genetic Programming	%78/10	%21/90
C4.5	%73/60	%26/40
Support Vector Machines	%75/10	%24/90
Discriminant analysis	%73/80	%26/20
K- nearest neighbor	%71/50	%28/50
logic regression	%75/20	%24/80

Table 3. Comparison of classification accuracy and model error in the German dataset

Considering the two tables are observed, in the other hand with increasing average classification accuracy, reduced the total average error in classification. Method has increased efficiency, if the error reduced. Hence with increasing accuracy and reducing customer validation error to ensure they can give loans to applicants in financial institutions, and in the future will be caused to substantially reduce financial risks.

5. Conclusion

Advanced data mining techniques have significant contribution in the field of information science, and we matched them with credit scoring models. Professionals and researchers are searching a model that it can slightly increase the prediction accuracy. Few small changes are significant role in reducing risk, to give facilities to the borrowers.

In this paper, the most common methods of validation customers were evaluated in terms of accuracy in both data sets the actual bank. It helps to professionals and researchers whose decisions, as well as experts and researchers are able to use these techniques to

change the banking system, the country. In classification models, no formal theory to select the best model and choose the best parameters. Select the best set of parameters can be done by initiative rules or by the search grade. In this way, the overview of different values of parameters are studied and a set with the best prediction accuracy is selected. Hence the algorithm accurately predicted, may be a function of parameters. Thus, large amplitude changes in the parameters should be considered to reduce the face of possible local optimum.

Reference

[1] Baesens, B., Setiono, R., Mues, C.; and Vanthienen, J. (2003). *Using neural network rule extraction and decision tables for credit-risk evaluation*. Management Science, 49(3), 312–329.

[2] Baesens, B., Van Gestel, T. Viaene, S. Stepanova, M., Suykens, J.,Vanthienen, J. (2003). *Benchmarking state-of-art classification algorithms for credit scoring*, Journal of the Operational Research Society, 54, 627–635.

[3] Chen, M. S., Han, J., Yu, P. S. (1996). *Data mining: an overview from a database perspective*, IEEE Trans. Knowledge Data Engineering 8(6): 866–883. doi:10.1109/69.553155.

[4] Cheng-Lung Huang et al.(2006).*Credit scoring with a data mining based on support vector mashines*, Expert System with Application, doi: 0.1016/j.eswa,

[5] Chen, S. Y., & Liu, X. (2004).*The contribution of data mining to information science, Journal of Information Science*, 30(6), 550–558.

[6] Drucker, H ., Schapire, R., Simard P. (1993).*Boosting performance in neural networks, International Journal of Pattern Recognition and Artificial Intelligence*, 7(4):704-719.

[7] Freund, Y. , Robert, E., Schapire. (1997).*A decision-theoretic generalization of on-line learning and an application to boosting,* Journal of Computer and System Sciences, 55(1), 119-139, August.

[8] Friedman J. (2003). *Recent advances in predictive machine learning*, Proceedings of Phystat,Stanford University.

[9] Frydman H.E., Altman, E.I., and Kao D-L. (1985). *Introducing recursive partitioning for financial classification: the case of financial distress*. Journal of Finance.

[10] Han. J, .Kamber M. (2001). *Data Mining: Concepts and Techniques*, San Diego Academic Press.

[11] Henley, W. E.(1995). *Statistical aspects of credit scoring, Dissertation*, The Open University, Milton Keynes, UK.

[12] Henley W. E., and Hand, D. J. (1996). *A k-nearest neighbor classifier for assessing consumer credit risk*. Statistician, 44(1), 77–95.

[13] Hsieh. N.-C. (2005). *Hybrid mining approach in the design of credit scoring models*, Expert Systems with Applications, 28(4):655.

[14] Hussein A.Abdou. (2010). *Genetic Programming for credit scoring: The case of Egyptian Public sector banks*, Expert System with Application, 36: 11402-11417, doi:10.1016/j.esw.

[15] Ince. .H, Aktan .B. (2009). *A comparison of data mining technique for credit scoring in banking: A managerial perceptive*, Journal of Business Economics and Management, , doi: 10.3846/1611-1699. 10. 233-240.

[16] Koza, . John R. Koza. (1990). *Genetic programming: A paradigm for genetically breeding populations of computer programs to solve problems*, Technical Report STAN-CS-90-1314, Stanford University.

[17] Lee, T.-S., Chen, I.-F. (2005). *A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression*, splines. Expert Systems with Applications, 28(4), 743–752.

[18] Lee, T.-S., Chiu, C.-C., Lu, C.-J., & Chen, I.-F. (2005). *Credit scoring using the hybrid neural discriminate technique*. Expert Systems with Applications, 23(3), 245–254.

[19] Lee, T. S., Chiu, C. C., Chou, Y. C., Lu, C. J. (2006). *Mining the customer credit using classification and regression tree and multivariate adaptive regression splines*, Computational Statistics and Data Analysis 50: 1113–1130.

[20] Malhotra, R., & Malhotra, D. K. (2005).*Differentiating between good credits and bad credits using neuro fuzzy systems*, European Journal of Operational Research,136(1), 190–211.

[21]D. martens et al. (2007). *Comprehensible credit scoring models using extraction from support vector machine*, European Journal of Research.

[22] Murphy, P. M., Aha, D. W. (2005). *UCI repository of machine learning databases*, Department of Information and Computer Science, University of California, Irvine, CA, http://www.ics.uci.edu/mlearn/LRepository.

[23] Ong, C.-S., Huang, J.-J., Tzeng, G.-H. (2005). *Building credit scoring models using genetic programming*, Expert Systems with Applications, 29(1), 41–47.

[24] Reichert, A. K., Cho, C. C., Wagner, G. M. (1983). *An examination of the conceptual issues involved in developing credit-scoring models*, Journal of Business and Economic Statistics, 1(2), 101–114.

[25] Schapire, Robert E., Singer, Yoram, (1997). *Improved boosting algorithms using confidence-rated predictions*, Machine Learning, 37(3), pp.297-336, December.

[26] Schebesch KB, Stecking R. (2005). *Support vector machines for classifying and describing credit applicants: detecting typical and critical regions*, J Oper Res Soc 56(8):1082–1088.

[27] Thomas, L. C. (2000). *A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers*, International Journal of Forecasting, 16(2), 149–172.

[28] Vapnik V. N. (1995). *The nature of statistical learning Theory*, New York: Springer-Verlag.

[29] West, D.(2005). Neural network credit scoring models, Computers and Operations Research, 27(11–12), 1131– 1152.

[30]Xiujuan Xu et al. (2009). *Credit scoring algorithm based on link analysis ranking with support vector machine*, Expert System with Application, 36, 2625-2632.

[31]Yan Zhan,Hao Chen and Guo-Chun Zhang. (2006). *An optimization algorithm of k-NN classification, Proceedings of the Fifth International conference on Machine Learning and Cybernetics*, Dalian,13-16 August.