International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-1 Issue-4, April 2012

A Study of Classification Based Credit Risk Analysis Algorithm

Ketaki Chopde, Pratik Gosar, Paras Kapadia, Niharika Maheshwari, Pramila M. Chawan

Abstract— Almost all business organizations these days generate large amounts of data regarding their work. Simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data. The information thus extracted can be used by organizations in decision making process. In this paper, we study the data mining techniques used for credit risk analysis, in particular the decision tree technique.

Index Terms— Data Mining, Credit risk analysis, Decision tree.

I. INTRODUCTION

Credit Risk Analysis is the process of identifying the risk involved in granting a loan to a customer. The classification is contingent on characteristics of the borrower (such as age, education level, occupation, marital status and income), the repayment performance on previous loans and the type of loan. These models are also applicable to small businesses since these may be regarded as extensions of an individual costumer. While financial institutions have faced difficulties over the years for a multitude of reasons, the major cause of serious banking problems continues to be directly related to lax credit standards for borrowers and counterparties, poor portfolio risk management, or a lack of attention to changes in economic or other circumstances that can lead to a deterioration in the credit standing of a bank's counterparties. Lending organizations such as banks are perpetually risking loss while granting credits. The large volume of loan portfolios also implies that modest improvements in scoring accuracy may result in significant savings for financial institutions. The goal of a credit scoring model is to classify credit applicants into two classes: the "good credit" class that is liable to reimburse the financial obligation and the "bad credit" class that should be denied credit due to the high probability of defaulting on the financial obligation. Credit loans and finances have risk of being defaulted. The accurate assessment of consumer credit risk is of uttermost importance for lending organizations. Increasing the demand for consumer credit has led to the competition in credit industry. So credit managers have to develop and apply machine learning methods to handle analyzing credit data in order to save time and reduction of errors. Credit scoring is a widely used technique that helps financial institutions evaluate the likelihood for a credit applicant to default on the financial obligation and decide whether to grant credit or not. The precise judgment of the credit worthiness of applicants allows financial institutions to increase the volume of granted credit while minimizing possible losses.

Manuscript published on 30 April 2012.
***** Correspondence Author (s)
Ketaki Chopde*, Computer Technology Department, Veermata Jijabai Technological Institute (V.J.T.I.), Mumbai,India, +91 8097418075, (e-mail: ketaki.2311@gmail.com).
Pratik Gosar, Computer Technology Department, Veermata Jijabai Technological Institute (V.J.T.I.), Mumbai,India, +91 9930300533, (e-mail: pratikgosar999@gmail.com).
Paras Kapadia, Computer Technology Department, Veermata Jijabai Technological Institute (V.J.T.I.), Mumbai,India, +91 9029261194, (e-mail: kapadiaparas@gmail.com).
Niharika Maheshwari, Computer Technology Department, Veermata Jijabai Technological Institute (V.J.T.I.), Mumbai,India, +91 9833664306, (e-mail: niharika.m90@gmail.com).
Pramila M Chawan, Associate Prof., Computer Technology Department, Veermata Jijabai Technological Institute (V.J.T.I.), Mumbai,India, +91 2224198154, (e-mail: pmchawan@vjti.org.in).

II. DATA MINING TECHNIQUES EMPLOYED FOR CREDIT RISK ANALYSIS

The decision-making of accepting or rejecting a client's credit by banks is commonly executed via judgmental techniques and/or credit scoring models. Most banks and financial institutions use the judgmental approach which is based on the 3C's, 4C's or 5C's which are character, capital, collateral, capacity and condition. However, to improve assessment of credit card applicants, banks can use credit scoring or predictive models to classify the applicants. Generally, Linear Discriminant Analysis and logistic regression are two popular statistical tools to construct credit scoring models. Some techniques are Decision trees, neural networks, Bayesian classifier, SVM, kernel networks etc.

Some of the popular techniques used are classified in [2] as a proposed classification of the credit risk assessment techniques. Some of the various techniques are also discussed in [3]. We describe the main decision tree and neural network techniques in detail in this paper.

III. DECISION TREES

A classification tree is a tree-like graph of decisions and their possible consequences. Topmost node in this tree is the root node which a decision is supposed to take on it. In each inner node, a test is done on an attribute or input variable. Each branch which follows the node lead to the result of the test, and the classes are represented by leaf nodes. Classification trees are used when the response variable is quantitative discrete or qualitative. CT is based on maximizing purity measure

of the response variables of the observations. The advantage of this method is that it is a white box model and so it is simple to understand and explanation, but the limitation of this model is that, it cannot be generalized a designed structure for one context to the other contexts. Decision tree divides customer loan segments into smaller sub segments recursively. At each segment, splitting is made in a way that boosts proportions of either defaulted loans or fully-recovered loans, in each resulting sub segment. This process repeats until no further improvement can be made. Customer loan segments are partitioned recursively in a way that increases the proportion of either defaulted or fully-recovered loans. Once the classification tree is created, it can be tested using test data for checking the accuracy. It can then be used for the credit risk analysis.

IV. VARIOUS IMPLEMENTATION METHODS OF DECISION TREES FOR CREDIT RISK ANALYSIS

Decision trees are used to perform credit risk assessment by banks.

- The simplest way to utilize decision trees is directly. We can use the attributes of the customers as specified by the bank's database to form the tree. The classification tree thus formed will be used for analysing new customer data. An implementation of such a method is shown in [4].

- Another variation proposed is the Meta Decision Tree. Meta decision trees (MDTs) adopt the stacking framework of combining base-level classifiers. The difference between meta and ordinary decision trees (ODTs) is that MDT leaves specify which base-level classifier should be used, instead of predicting the class value directly. Stacking with MDTs performs better than bagging and boosting of decision trees, which are the state of the art methods for learning ensembles of classifiers. Another advantage of the MDTs is their understandability: they provide information about the relative areas of expertise of the base-level classifiers. This method is explained in [5].

- Random forest: The RF technique is based on ensemble learning of the decision tree. Ensemble learning is useful for improving the model accuracy with the integration of a set of learning machine.

  RF has the following features:

  1) It is capable of dealing with high dimensional data.

  2) It allows the users to estimate the variable importance that indicates the index what extent each variable is important.

  3) Variables *x*1-*xm* are randomly selected from all the variables at each node of the tree. It provides more accurate classification than CART.

  4) It stops growing the tree if the number of nodes reaches at a certain threshold value.

  5) It does not carry out pruning

This method allows the market participants to evaluate the credit risk appropriately. It is explained in [9].

- Another way to improve the decision tree is integrating domain knowledge constraint. A lot of information about the domain is useful while analysing the data. Including it will improve its performance. First the information gain is calculated for each attribute and using its value decision tree is constructed and the prediction is improved. The choice of attribute's measure is the key to algorithm of decision tree. Using information gain as measure to choose testing attribute, which is called attribute selection measure. The attribute with the highest information gain is selected as the testing attribute of current node, which makes sample grouping results reflect minimum randomness or impurity. The expected testing numbers can be reduced to minimum when classifying the objects by this approach and a simple tree are ensured to be found.

- Most banks have a large number of indices used in the predictive analysis. Another variation is decreasing the number of indices using cluster analysis of their attributes. A smaller decision tree is constructed using the reduced index system. This yields a better prediction model than the normal indexed system decision tree.

- Bagging is another technique employed to improve the performance of decision trees. Bagging (Bootstrap aggregating) (Breiman, 1994) is a meta algorithm to improve classification or regression models in terms of stability and classification accuracy in machine learning. Also it reduces the variance and helps to avoid overfitting. It is seen that the prediction accuracy is improved by aggregating many models rather than building one high capacity model.

- A comparison of CART decision trees with other popular techniques such as decision trees is presented in [12]. Their implementations are compared using the same dataset. Prediction accuracy of each method for different misjudgements is also included.

- Boosting (Freund and Schapire, 1991; Schapire, 2002; Friedman, 2003) is a procedure that aggregates many "weak" classifiers in order to achieve a high classification performance. Additionally, boosting helps stabilizing the response of classifiers with respect to changes in the training sample. The boosting algorithm initiates by giving all credit applicants the same weight w(0). After a classifier is built, the weight of each applicant is changed according to the classification given by that classifier. Then, a second classifier is built using the reweighted training sample. This procedure is typically repeated several hundreds of times. The final classification of a credit applicant is a weighted average of the individual classifications over all classifiers. There are several methods to update the weights and combine the individual classifiers.

The most popular boosting algorithm is AdaBoost (Freund and Schapire, 1996). In this method n number of trees are created. By using a weighted method to classify the dataset the trees are normalized and finally used in combination. After the kth decision tree

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

is built, the total misclassification error of the tree, defined as the sum of the weights of misclassified credits over the sum of the weights of all credits, is calculated. Then, the weights of misclassified credit applicants are increased (boosted). Finally, the new weights are renormalized, w(k+1) and the tree k+1 is constructed. The final classification of the credit applicant is a weighted sum of the classifications over individual trees. Good credits tend to have large positive score and band credit tends to have large negative score. The performance of boosted decision trees (BDT) is optimized by adjusting two parameters: the number of decision trees that are aggregated to form the final classifier and the minimum number of credit applicants that a tree node must contain in order to be split.

V. CONCLUSION

Credit risk analysis (finance risk analysis, loan default risk analysis) and credit risk management is important to financial institutions which provide loans to businesses and individuals. The system using data mining for credit risk analysis enables the bank to reduce the manual errors involved in the same. Customers need to provide specific data and the vagueness of normal banking is removed. Once enough data is available, the analysis model will use this data to reach a terminal node that contains the decision regarding the customer's credit worthiness. This decision making process is quick and saves time and resources for the bank. Decision trees and Artificial Neural Network are some of the popular techniques. The challenge of reducing misjudgment is immense to overcome. The goal of each of these techniques is, in fact, to reduce the misjudgments. If we can completely eradicate it we can classify the customers directly and reduce loss for the financial institutions.

REFERENCES

[1] Jiawei Han, MichelineKamber, "Data Mining Concepts and Technique", 2nd edition.
[2] Abbas Keramati, NiloofarYousefi, "A Proposed Classification of Data Mining Techniques in Credit Scoring", International Conference on Industrial Engineering and Operations Management, 2011.
[3] Yap Bee Wah, Irma Rohaiza Ibrahim, "Using Data Mining Predictive Models to Classify Credit Card Applicants".
[4] QiweiGan, BinjieLuo, Zhangxi Lin, "Risk Management of Residential Mortgage in China Using Date Mining A Case Study", 2009 International Conference on New Trends in Information and Service Science.
[5] Bernard ienko, LjupCoTodorovski, and Sago Dieroski, "A comparison of stacking with meta decision trees to bagging, boosting, and stacking with other methods"
[6] Jingping Chen, Haiwei Pan, Qilong Han, Linghu Chen, Jun Ni, "Credit Risk Assessment Model Based On Domain Knowledge Constraint", 2008 International Multi-symposiums on Computer and Computational Sciences.
[7] Fei Li, JunXu, Zhi-Tong Mu, Ya-Lou Huang, "Data Mining-Based Credit Evaluation For Users Of CreditCard"
[8] Qiuju Yin, Ke Lu, "Data Mining Based Reduction on Credit Evaluation Index of Bank Personal Customer", 2010 International Conference on Future Information Technology and Management Engineering.
[9] Hiroyuki Mori, Yasushi Umezawa, "Credit Risk Evaluation in Power Market with Random Forest".
[10] Bastos, J. A., 2008, "Credit scoring with boosted decision trees". City: Munich Personal RePEc Archive, pp. 262-273.
[11] Defu Zhang, Xiyue Zhou, Stephen C.H. Leung, JieminZheng, "Vertical bagging decision trees model for credit scoring", Expert Systems with Applications 37 (2010) 7838–7843.
[12] Hong Yu, Xiaolei Huang, Xiaorong Hu, HengwenCai, "A Comparative Study on Data Mining Algorithms for Individual Credit Risk Evaluation", 2010 International Conference on Management of e-Commerce and e-Government.

AUTHOR PROFILE

Pramila M. Chawan is currently working as an Associate Professor in the Computer Technology Department of "Veermata Jijabai Technological Institute (V.J.T.I.), Matunga, Mumbai (INDIA)". She received her Bachelors' Degree in Computer Engineering from V.J.T.I., Mumbai University (INDIA) in 1991 & Masters' Degree in Computer Engineering from V.J.T.I., Mumbai University (INDIA) in 1997. She has an academic experience of 20 years. She has taught Computer related subjects at both Undergraduate & Post Graduate levels. Her areas of interest are Software Engineering, Software Project Management, Management Information Systems, Advanced Computer Architecture & Operating Systems. She has published 12 papers in National Conferences and 7 papers in International Conferences & Symposiums. She also has 16 International Journal publications to her credit. She has guides 35 M. Tech. projects & 85 B. Tech. projects.

Ketaki Chopde, **Pratik Gosar**, **Paras Kapadia** and **Niharika Maheshwari** are currently studying as Final Year B.Tech students in the Computer Technology Department of "Veermata Jijabai Technological Institute (V.J.T.I.), Matunga, Mumbai (INDIA)".

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)