



ISSN: 2319-5967

ISO 9001:2008 Certified

International Journal of Engineering Science and Innovative Technology (IJESIT)

Volume 2, Issue 3, May 2013

Study of Data Mining Techniques used for Financial Data Analysis

Abhijit A. Sawant and P. M. Chawan

Department of Computer Technology, VJTI, Mumbai, INDIA

Abstract— This paper describes about different data mining techniques used in financial data analysis. Financial data analysis is used in many financial institutes for accurate analysis of consumer data to find defaulter and valid customer. For this different data mining techniques can be used. The information thus obtained can be used for Decision making. In this paper we study about loan default risk analysis, Type of scoring and different data mining techniques like Bayes classification, Decision Tree, Boosting, Bagging, Random forest algorithm and other techniques.

Index Terms— Data mining, Bayes Classification, Decision Tree, Boosting, Bagging, Random forest algorithm

I. INTRODUCTION

In Banking Sectors and other such leading organization the accurate assessment of consumer is of uttermost importance. Credit loans and finances have risk of being defaulted. These loans involve large amounts of capital and their non-retrieval can lead to major loss for the financial institution. Therefore, the accurate assessment of the risk involved is a crucial matter for banks and other such organizations. Not only is it important to minimize risk in granting credit but also the errors in declining any valid customer. This is to save the banks from lawsuits. Increasing the demand for consumer credit has led to the competition in credit industry. In India, there are an increasing number of people opting for loans for houses and cars and there is also a large demand for credit cards. Such credits can be defaulted upon and have a great impact on the economy of the country. Thus, assessing loan default risk is important for our economy. Earlier Assessing of credit was usually done using statistical and mathematical methods by analysts. Nowadays Data mining techniques have gained popularity over the years because of their ability in discovering practical knowledge from the database and transforming them into useful information.

II. LOAN DEFAULT RISK ANALYSIS

Loan default risk assessment is one of the crucial issues which financial institutions, particularly banks are faced and determining the effective variables is one of the critical parts in this type of studies. In this Credit scoring is a widely used technique that helps financial institutions evaluates 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. Credit scoring models are known as statistical models which have been widely used to predict the default risk of individuals or companies. These are multivariate models which use the main economic and financial indicators of a company or individuals' characteristic such as age, income and marital status as input, assign them a weight which reflect its relative importance in predicting default. The result is an index of credit worthiness that is expressed as a numerical score, which measures the borrower's probability of default. The initial credit scoring models are devised in the 1930s by authors such as Fisher and Durand. 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. 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 customer.

Type of scoring

Based on Paleologo, G., Elisseeff, A., and Antonini, G., research there are different kinds of scoring:

1) Application Scoring:

This kind of scoring consists of estimation credit worthiness of a new applicant who applies for credit. It estimates financial risk with respect to social, demographic financial conditions of a new applicant to decide whether to grand credit to them or not.



ISSN: 2319-5967

ISO 9001:2008 Certified

International Journal of Engineering Science and Innovative Technology (IJESIT)

Volume 2, Issue 3, May 2013

2) Behavioral Scoring:

It is similar to application scoring with difference that it involves existing customers, so the lender has some evidences about borrower's behavior to dynamic management of portfolio process.

3) Collection Scoring:

Collection scoring classifies customers into different groups according to their levels of insolvency. In the other words, it separates the customers who need more decisive actions from those who do not require to be attended to immediately. These models are used in order to management of delinquent customers from the first signs of delinquency.

4) Fraud Detection:

It categorizes the applicants according to the probability that an applicant be guilty.

III. DATA MINING TECHNIQUES

Data mining a field at the intersection of computer science and statistics is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Mostly use Data mining techniques are as follows:

A. Bayes Classification

A Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independent assumptions and is particularly suited when the dimensionality of the inputs is high. A naive Bayes classifier assumes that the existence (or nonexistence) of a specific feature of a class is unrelated to the existence (or nonexistence) of any other feature. Classification is a form of data analysis which can be used to extract models describing important data classes. Classification predicts categorical labels (or discrete values). Data classification is a two step process. In the first step, a model is built describing a predetermined set of data classes or concepts. The model is constructed by analyzing database tuples described by attributes. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute. In the second step, the model is used for classification. First, the predictive accuracy of the model is estimated. The accuracy of a model on a given test set is the percentage of test set samples that are correctly classified by the model. . If the accuracy of the model is considered acceptable, the model can be used to classify future data tuples or objects for which the class label is not known.

Bayesian Algorithm:

1. Order the nodes according to their topological order.
2. Initiate importance function $Pr^0(X|E)$, the desired number of samples m , the updating interval l , and the score arrays for every node.
3. $k \leftarrow 0, T \leftarrow \emptyset$
4. for $i \leftarrow 1$ to m do
5. if $(i \bmod l == 0)$ then
6. $k \leftarrow k+1$
7. update importance function $Pr^k(X|E)$ based on T
 end if
8. $s_i \leftarrow$ generate a sample according to $Pr^k(X|E)$
9. $T \leftarrow T \cup \{s_i\}$
10. Calculate $Score(s_i, Pr(X|E, e), Pr^k(X|E))$ and add it to the corresponding entry of every array according to the instantiated states.
11. Normalize the score arrays for every node.

The major disadvantage of this model is that the predictive accuracy is highly correlated with this assumption.

An advantage of this method is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.

B. Decision Tree

A classification (decision) tree is a tree-like graph of decisions and their possible consequences. Topmost node in this tree is the root node on which a decision is supposed to be taken. In each inner node, it is done as a test on an attribute or input variable. Specifically each branch of the tree is a classification question and the leaves of the tree



ISSN: 2319-5967

ISO 9001:2008 Certified

International Journal of Engineering Science and Innovative Technology (IJESIT)

Volume 2, Issue 3, May 2013

are partitions of the dataset with their classification. The processes in decision tree algorithms are very similar when they build trees.

These algorithms look at all possible distinguishing questions that could possibly break up the original training dataset into segments that are nearly homogeneous with respect to the different classes being predicted. Some decision tree algorithms may use heuristics in order to pick the questions or even pick them at random. The advantage of this method is that it is a white box model and so it is simple to understand and explain, but the limitation of this model is that, it cannot be generalized into a designed structure for all contexts.

A decision tree is a mapping from observations about an item to conclusion about its target value as a predictive model in data mining and machine learning. Generally, for such tree models, other descriptive names are classification tree (discrete target) or regression tree (continuous target). In these tree structures, the leaf nodes represent classifications, the inner nodes represent the current predictive attributes and branches represent conjunctions of attributions that lead to the final classifications. The popular decision trees algorithms include ID3, C4.5 which is an extension of ID3 algorithm and CART.

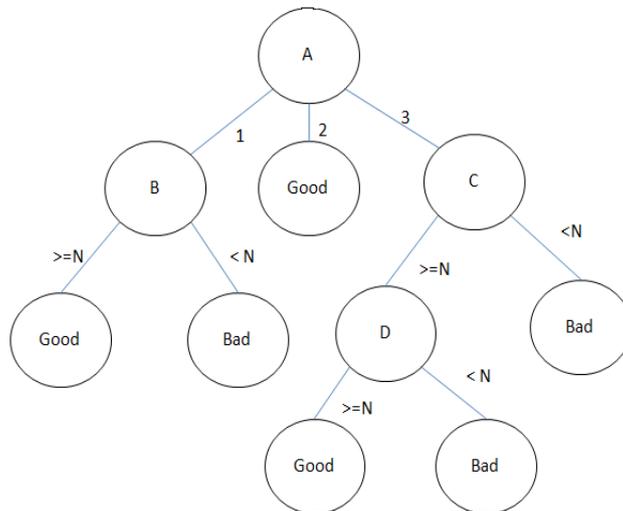


Fig 1- Decision Tree

C. Boosting

The concept of boosting applies to the area of predictive data mining, to generate multiple models or classifiers (for prediction or classification), and to derive weights to combine the predictions from those models into a single prediction or predicted classification.

A simple algorithm for boosting works like this: Start by applying some method to the learning data, where each observation is assigned an equal weight. Compute the predicted classifications, and apply weights to the observations in the learning sample that are inversely proportional to the accuracy of the classification. In other words, assign greater weight to those observations that were difficult to classify (where the misclassification rate was high), and lower weights to those that were easy to classify (where the misclassification rate was low).

Boosting will generate a sequence of classifiers, where each consecutive classifier in the sequence is an "expert" in classifying observations that were not well classified by those preceding it. During deployment (for prediction or classification of new cases), the predictions from the different classifiers can then be combined (e.g., via voting, or some weighted voting procedure) to derive a single best prediction or classification.

The most popular boosting algorithm is AdaBoost. AdaBoost, short for Adaptive Boosting, is a machine learning algorithm, formulated by Yoav Freund and Robert Schapire. It is a meta-algorithm, and can be used in conjunction with many other learning algorithms to improve their performance.

The Boosting Algorithm AdaBoost

Given: $(x_1, y_1), \dots, (x_m, y_m); x_i \in X, y_i \in \{-1, 1\}$

Initialize weights $D_1(i) = 1/m$

For $t = 1, \dots, T$:

1. (Call WeakLearn), which returns the weak classifier $h_t : X \rightarrow \{-1, 1\}$ with minimum error w.r.t distribution D_t ;
2. Choose $\alpha_t \in \mathbb{R}$,
3. Update



ISSN: 2319-5967

ISO 9001:2008 Certified

International Journal of Engineering Science and Innovative Technology (IJESIT)

Volume 2, Issue 3, May 2013

$$D_{t+1}(i) = (D_t(i) \exp(-\alpha_t y_i h_t(x_i))) / Z_t$$

Where Z_t is a normalization factor Chosen so that D_{t+1} is a distribution

Output the strong Classifier

$$H(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(x) \right)$$

D. Bagging

Bagging (Bootstrap aggregating) was proposed by Leo Breiman in 1994 to improve the classification by combining classifications of randomly generated training sets.

Bootstrap aggregating (bagging) is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid over fitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging is a special case of the model averaging approach.

Given a standard training set D of size n , bagging generates m new training sets D_i , each of size $n' < n$, by sampling from D uniformly and with replacement. By sampling with replacement, some observations may be repeated in each D_i . If $n'=n$, then for large n the set D_i is expected to have the fraction $(1 - 1/e)$ of the unique examples of D , the rest being duplicates. This kind of sample is known as a bootstrap sample. The m models are fitted using the above m bootstrap samples and combined by averaging the output voting.

Bagging leads to "improvements for unstable procedures", which include, for example, neural nets, classification and regression trees, and subset selection in linear regression? On the other hand, it can mildly degrade the performance of stable methods such as K -nearest neighbours.

The Bagging Algorithm

Training phase

1. Initialize the parameter
 - $D = \emptyset$, the ensemble.
 - L_k the number of classifiers to train.
2. For $k=1, \dots, L$
 - Take a bootstrap sample S_k from Z .
 - Build a classifier D_k using S_k as the training set
 - Add the classifier to the current ensemble,
 $D = D \cup D_k$.
3. Return D .

Classification phase

4. Run D_1, \dots, D_k on the input x .
5. The class with the maximum number of votes is chosen as the label for x .

E. Random forest Algorithm

Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. The algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler, and "Random Forests" is their trademark. The term came from random decision forest which was first proposed by Tin Kam Ho of Bell Labs in 1995. The method combines Breiman's "bagging" idea and the random selection of features, introduced independently by Ho and Amit and Geman in order to construct a collection of decision trees with controlled variation. The selection of a random subset of features is an example of the random subspace method, which, in Ho's formulation, is a way to implement stochastic discrimination proposed by Eugene Kleinberg. The introduction of random forests proper was first made in a paper by Leo Breiman. This paper describes a method of building a forest of uncorrelated trees using a CART like procedure, combined with randomized node optimization and bagging. It is better to think of random forests as a framework rather than as a particular model. The framework consists of several interchangeable parts which can be mixed and matched to create a large number of particular models, all built around the same central theme. Constructing a model in this framework requires making several choices:

1. The shape of the decision to use in each node.
2. The type of predictor to use in each leaf.
3. The splitting objective to optimize in each node.
4. The method for injecting randomness into the trees.

F. Other Techniques

1] The Back Propagation Algorithm



ISSN: 2319-5967

ISO 9001:2008 Certified

International Journal of Engineering Science and Innovative Technology (IJESIT)

Volume 2, Issue 3, May 2013

The back propagation algorithm (Rumelhart and McClelland, 1986) is used in layered feed-forward ANNs. This means that the artificial neurons are organized in layers, and send their signals “forward”, and then the errors are propagated backwards. The network receives inputs by neurons in the input layer, and the output of the network is given by the neurons on an output layer. There may be one or more intermediate hidden layers. The back propagation algorithm uses supervised learning, which means that we provide the algorithm with examples of the inputs and outputs we want the network to compute, and then the error (difference between actual and expected results) is calculated. The idea of the back propagation algorithm is to reduce this error, until the ANN learns the training data. The training begins with random weights, and the goal is to adjust them so that the error will be minimal.

II] Genetic Algorithm:

Genetic Algorithm (GA) was developed by Holland in 1970. This incorporates Darwinian evolutionary theory with sexual reproduction. GA is stochastic search algorithm modeled on the process of natural selection, which underlines biological evolution. A has been successfully applied in many search, optimization, and machine learning problems. GA process in an iteration manner by generating new populations of strings from old ones. Every string is the encoded binary, real etc., version of a candidate solution. An evaluation function associates a fitness measure to every string indicating its fitness for the problem. Standard GA apply genetic operators such selection, crossover and mutation on an initially random population in order to compute a whole generation of new strings.

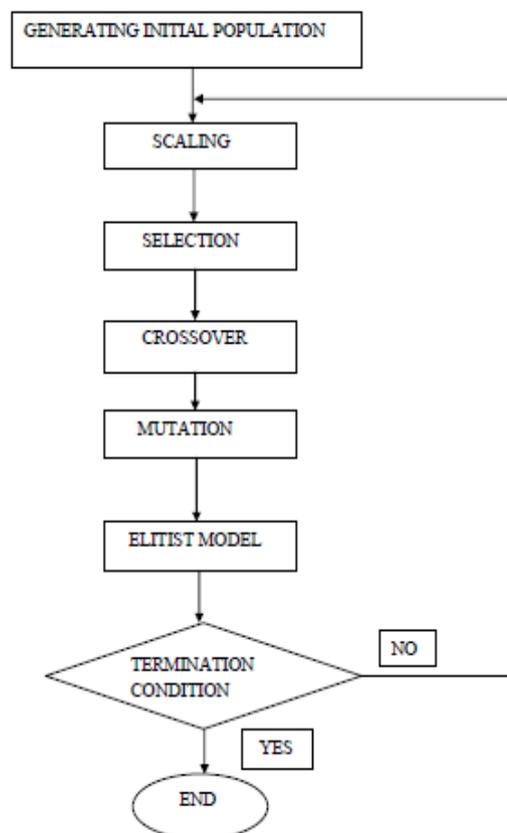


Fig 2- Genetic Algorithm Flowchart

Selection deals with the probabilistic survival of the fittest, in those more fit chromosomes are chosen to survive. Where fitness is a comparable measure of how well a chromosome solves the problem at hand.

Crossover takes individual chromosomes from P combines them to form new ones.

Mutation alters the new solutions so as to add stochasticity in the search for better solutions.

In general the main motivation for using GAs in the discovery of high-level prediction rules is that they perform a global search and cope better with attribute interaction than the greedy rule induction algorithms often used in data mining.



ISSN: 2319-5967

ISO 9001:2008 Certified

International Journal of Engineering Science and Innovative Technology (IJESIT)

Volume 2, Issue 3, May 2013

III] Particle Swarm Optimization:

The investigation and analysis on the biologic colony demonstrated that intelligence generated from complex activities can provide efficient solutions for specific optimization problems. Inspired by the social behavior of animals such as fish schooling and bird flocking, Kennedy and Eberhart designed the Particle Swarm Optimization (PSO) in 1995. The basic PSO model consists of a swarm of particles moving in a d-dimensional search space. The direction and distance of each particle in the hyper-dimensional space is determined by its fitness and velocity. In general, the fitness is primarily related with the optimization objective and the velocity is updated according to a sophisticated rule.

Artificial neural networks (ANNs) are thus, non-linear statistical modeling based on the function of the human brain. They are powerful tools for unknown data relationship modeling. Artificial Neural Networks are able to recognize the complex pattern between input and output variables then predict the outcome of new independent input data.

IV] Support vector machine:

Support vector machine is a classifier technique. This method involves three elements. A score formula which is a linear combination of features selected for the classification problem, an objective function which considers both training and test samples to optimize the classification of new data, an optimizing algorithm for determining the optimal parameters of training sample objective function.

The advantages of the method are that, in the nonparametric case, SVM requires no data structure assumptions such as normal distribution and continuity. SVM can perform a nonlinear mapping from an original input space into a high dimensional feature space and this method is capable of handling both continuous and categorical predictions. The weaknesses of this method are that, it is difficult to interpret unless the features interpretable and standard formulations do not contain specification of business constraints.

IV. CONCLUSION

The system using data mining for loan Default risk analysis enables the bank to reduce the manual errors involved in the same. Decision trees are preferred by banks because they are a white box model. The discrimination made by decision trees is obvious and the people can understand its working easily. This enables the banks and other financial institutions to provide an account for accepting or rejecting an applicant. Boosting has already increased the efficiency of decision trees. The assessment of risk will enable banks to increase profit and can result in reduction of interest rate.

REFERENCES

- [1] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Technique", 2nd edition
- [2] Margaret H. Dunham, "Data Mining-Introductory and Advanced Topocs" Pearson Education, Sixtyh Impression, 2009.
- [3] Abbas Keramati, Niloofar Yousefi, "A Proposed Classification of Data Mining Techniques in Credit Scoring", International Conference on Industrial Engineering and Operations Management, 2011
- [4] Boris Kovalerchuk, Evgenii Vityaev, "DATA MINING FOR FINANCIAL APPLICATIONS", 2002
- [5] 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.
- [6] Raymond Anderson, "Credit risk assessment: Enterprise-credit frameworks".
- [7] Girisha Kumar Jha, "Artificial Neural Networks".
- [8] Hossein Falaki, "AdaBoost Algorithm".
- [9] Venkatesh Ganti, Johannes Gehrke, Raghu Ramakrishnan, "Mining Very Large Database" IEEE,1999.
- [10] <http://en.wikipedia.org>.

AUTHOR BIOGRAPHY



Abhijit A. Sawant is currently pursuing M. Tech in Computer engineering second year from "Veermata Jijabai Technological Institute (V.J.T.I), Matunga, Mumbai (INDIA)". He has received his Bachelors' Degree in Information technology from "Padmabhushan Vasantdada Patil Pratishan's College of Engineering (P.V.P.P.C.O.E), Sion, Mumbai (INDIA)" in 2010. He has published 2 papers in International Journal till date. His areas of interest are Software Engineering, Database, Data ware house, Data mining.



ISSN: 2319-5967

ISO 9001:2008 Certified

International Journal of Engineering Science and Innovative Technology (IJESIT)

Volume 2, Issue 3, May 2013



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, and 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 40 International Journal publications to her credit. She has guides 35 M. Tech. projects & 85 B. Tech. projects.