Credit Risk Analysis using Machine Learning Classifiers

Trilok Nath Pandey
Department of Computer Science and Engineering,
I.T.E.R
S 'O'A UNIVERSITY
Bhubaneswar, Odisha, India
triloknath75@gmail.com

Suman Kumar Mohapatra
Department of Computer Science and Engineering
I.T.E.R
S 'O'A UNIVERSITY
Bhubaneswar, Odisha, India
sumanbutu@gmail.com

Alok Kumar Jagadev
Department of Computer Science and Engineering,
KIIT University
Bhubaneswar, Odisha, India
alok.jagadev@gmail.com

Satchidananda Dehuri
Department of Information and Communication Technology
Fakir Mohan University
Balasore, Odisha, India
satchi.lapa@gmail.com

***Abstract*— Banking industry has the major activity of lending money to those who are in need of money. In order to payback the principle borrowed from the depositor bank collects the interest made by the principle borrowers. Credit risk analysis is becoming an important field in financial risk management. Many credit risk analysis techniques are used for the evaluation of credit risk of the customer dataset. The evaluation of the credit risk datasets leads to the decision to issue the loan of the customer or reject the application of the customer is the difficult task which involves the deep analysis of the customer credit dataset or the data provided by the customer. In this paper we are surveying different techniques for the credit risk analysis which are used for the evaluation for the credit risk datasets.**

Keywords*—*Credit risk; Machine learning; Bayesian classifier; Naive bayes classifier; Decision tree; KNN; K-means clustering; MLP; ELM; SVM; ANN

I. INTRODUCTION

The banking system evaluates the accuracy of the datasets in order to classify the loan applicants into good and bad classes. The applicants which are in the good classes have the high probability of returning the money to the bank. The applicants which are in the bad classes have the low probability of returning of the money to the bank so, they are the defaulters of the loans. To minimize the defaulter's rate in the credit dataset different types of credit risk evaluation techniques are used [1], [2], [3]. Sometime huge losses can be reduced even with a small improvement in the accuracy of credit evaluation. The benefits of the reliable credit risk dataset is it reduces the cost of credit scoring, good decision making in very less time and avoid less risk associates with loan collection. As credit risk evaluation plays a important role in the banking field and it is a very critical and biggest challenge faced by banks, accuracy plays a very important role in classification of credit data to avoid the financial loss. The increase in defaulter's rate in the credit risk data set which is not reliable gives motivation towards this field [25].

II. METHODOLOGY

Different types of techniques are used for the evaluation of credit datasets for the better and reliable credit risk analysis. In this literature survey we have discussed different approaches to the credit risk analysis.

A. BAYSIAN CLASSIFIER

Bayesian network is referred as belief network. Bayesian is a statistical model which is represented by direct acyclic graph or DAG [25]. Each node in the graph represents a random variable, where edges represent the functional dependencies among the corresponding random variable. The Bayesian network is an annotated acyclic graph that represents a joint probability distribution over a set of random variable.

Bayesian theorem states that,

$$P(h\backslash d) = \frac{p\left(\frac{D}{h}\right)P(h)}{P(D)} \quad (1)$$

Where,
P(h) : Prior probability of hypothesis h-prior
P(D) : prior probability of training data D-Evidence
P(D\h) : probability of D given h-likelihood
P(h\D) : probability of h given D-posterior probability

B. NAÏVE-BAYES CLASSIFIER

It is a simple probabilistic classifier based on bayes rule. This classifier is called naïve because it assumes attributes of a class are independent of one another [25], [14]. The naïve

978-1-5386-1887-5/17/$31.00 ©2017 IEEE

bayes classifier need small amount of data to calculate the parameters like mean and variance other variable essential for classification.

C. DECISION TREE

Decision tree is a predictive model which maps the observation about an item represented in branches to conclusion about a target value represent in leafs. It is one of the most successful techniques in supervised learning. In this learning technique each internal node or non leaf node is labeled with an input feature. Each leaf node in the tree is labeled with a class or probabilistic distribution over the classes [1], [3], [22]. The branches between the nodes tells the possible values that these attributes can have in the observed samples, while the terminal nodes tells us the final value of the dependant variables[30].

D. K-NEAREST NEIGHBOR

KNN is the non-parametric method used for classification and regression. It involves a training set of both positive and negative cases. It is also called lazy algorithm. It does not use any training data point to do any generalization [17]. This means training phase is very fast, it keeps all the training data. All the training data is needed during the testing phase. To determine the k-instances in the training dataset are more similar to a new input a distance measured is used [16]. For real valued input variable the most popular Euclidian Distance is used.

Euclidean distance is calculated as,

$$\text{Euclidean distance}(x,x_i)=\sqrt{(\text{sum}(x_j-x_{ij})\text{^}2)} \quad (2)$$

When KNN is used for regression problems the prediction is based on the mean or median of the K-most similar instances. When KNN is used for classification can be calculated as class with highest frequency for k-most similar instances. Each instances in essence votes for their class and the class with the most votes is taken as prediction.

E. K-MEANS

It is a type of unsupervised learning which is used when you have unlabeled data. The goal of this algorithm is to find groups in the data with the number of groups represented by the variable k. The algorithm works iteratively to assign each data point to one of K-groups based on the features that are provided [3]. The centroids of K clusters which can be used as low labeled new data and labels for training data [43]. In this algorithm a group of feature vectors are given as the dataset to be clustered. Randomly selects the amount of seed by k to be the cluster centre. Assign the nearest data point to the cluster.

F. MULTILAYER PERCEPTRON

MLP have been widely used in the financial area for credit risk. Machine learning used back propagation algorithm for supervised learning. It contains input layer, output layer and one or more than hidden layer between them [6]. All the layers are fully connected to each other. The processing element of each layer except the input layer is called the nodes which behave like a neuron. Each node in the one layer connected with another node connected with a certain weights in the next layer. There are multiple layers of neuron with nonlinear activation function. These layers allow the network to learn relationship between input and output vectors.

G. EXTREME LEARNING MACHINE

ELM is developed by Huang is developed for generalized single hidden layer feed forward networks [6]. ELM randomly select the hidden node parameter after which the network can represents as a linear system and output weights can be computed analytically [12]. ELM tends to obtain a smallest training error and the smallest norm of weights that leads to good generalization. ELM is very fast in learning and provides good generalization performance on many artificial and real large applications [20], [21], [28], [36]-[38]. ELM is a novel training algorithm for single layer feed forward network and is very effective and efficient. Given N distinct training samples (xi , ti) ϵ Rn * Rm (i=1,2,…..,N), the output of a SLFN with N hidden nodes can be represented by,

$$O\,j = \sum_{i=1}^{N} \beta fi(xi) = \sum_{i=1}^{N} \beta f(x\,j; ai, bj) \quad j = 1,\ldots\ldots,N \quad (3)$$

Where, Oj is the output vector of SLFN with respect to input sample xi. ai = [ai1,ai2,……,ain]T and bi are the learning parameter generated randomly of the j th hidden node that is βi = [$\beta i1$, $\beta i2$,……,βim]T is the link connecting the j th hidden node and the output node. f (xj ; ai , bi) is the activation function of the original ELM.

H. SUPPORT VECTOR MACHINE

Support vector machine is the supervised learning with associated learning algorithm that examines the data used for classification and regression. SVM constructs is a hyper plane or set of hyper planes in a high or infinite dimensional space which can be used for classification, regression and other tasks [4], [5], [7], [8]. SVM is first proposed by vapnik in 1995 for machine learning and proved its performance in many fields. A good separation is achieved by that has largest distance to nearest training data point of any class (functional margin) [10], [29]. The larger is the margin the lower the generalization error of the classifier [14]. SVM are helpful in text and hypertext as their application can importantly reduce the need for labeled training instances in both the standard inductive and transductive [15], [30]. Classification of images can also be performed using SVM. Many experimental results showed that SVM give higher search accuracy than traditional query refinement [13]. Hand –written characters can be recognized using SVM. Let D be a training set by l pattern si , each pattern is (xi , yi) , where xi ϵ R1 and yi ϵ {-1,1} , where i = 1,…., l . The pattern with +1 output is positive patterns and the others are called negative patterns [25]. The decision function f(x) is given by,

$$f(x) = \text{sgn}(w.x+b) \quad (4)$$

To compute optimal hyper plane the optimized problem is to be solved, Minimization : $1 \div 2 \parallel w^2 \parallel$, Subject to : $y_i ((w.x_i) + b) - 1) \geq 0$ .The margin of hyper plane $= 2/ \|w\|$, equivalent optimization is ,

$$\sum_{i=1}^{l} y_i \alpha_i = 0 \quad (5)$$

I. ARTIFICIAL NEURAL NETWORK

Artificial neural network is constituted of a group of neural network that connect with a weighted nodes [9], [11], [16]. Every node can replicate a neuron of creatures and the synaptic that connects among the neuron is equal to connection among these nodes. The neural network consists of three layers that is input layer, hidden layer and output layer and it is called multilayer perceptron [19], [23],[24]. In the MLP the network of layers connected as a layer of input units connected to layer of hidden units which are then connected to a layer of output unit as shown in Fig.1.

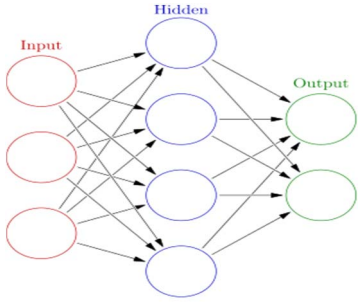

Fig.1- structure of ANN classifier
[https://en.wikipedia.org/wiki/Artificial_Neural_Network]

III. ENSEMBLE THE CLASSIFIER

An ensemble of classifier involves a group of base classifiers which are trained individually. To ensemble a classifier the decision is taken by base classifiers. The jointly decision for classification of new and unseen instances is taken by voting. The voting may be weighted or non-weighted. To ensemble the classifier the base classifiers are combined in a way so that higher performance is achieved in combined classifier than the alone one classifier. Some researchers have shown that by using the aggregating approach the classifiers can easily achieve improved accuracies on aggregation of the individual classifier in classification application as well as credit evaluation. Various different approaches of aggregation used for enhance the accuracies of classifier. The most aggregation approaches are bagging and boosting [25], [31]-[35].

A. Bagging

The Bagging is a fore standing ensemble strategy introduced by Breiman in 1996 [8], [13]. This is machine learning ensemble Meta algorithm which is designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It helps to reduce over fitting [18]. The bagging is special case of model averaging approach. In bagging mostly the similar kind of classifier are chosen as base classifier. Using bagging approach one can produce different decision structures by having different training [18] set of having same size and that is done by sampling the training set "with replacement".

B. Boosting

Boosting is a sequential ensemble method that is, try add new models that do well where previous models lack [25], [26], [27]. Boosting is aiming to decrease bias not the variance. It is suitable for low variance high bias models. Most of the boosting algorithms consist of weak classifier with respect to a distribution and adding them into strong classifier [16]. Boosting creates an ensemble of classifier by sampling again the training dataset which then combined by cost function or majority vote.

IV. DATASET USED

Mostly two credit datasets such as German credit and Australian credit dataset are used for the performance of machine learning algorithms and also used for ensemble learning. These two datasets are collected from UCI machinery repository (http://archive.ics.uci.edu/ml/). There are two classes in the dataset which is the "good" and "bad" reflection of the creditors to whom loan is approved and not approved.

Table.1 summary of credit dataset used

Dataset	Attributes	Instances	Classes
Australian	14	690	2
German	20	1000	2

V. COMPARISON & ANALYSIS

Here we have compared the classifiers accuracies of different algorithms using the German and Australian dataset.

Table.2 Accuracies of the classifiers

Algorithm	Dataset	Accuracy
Bayesian classifier	German	77.10
	Australian	86.96
Naive-bayes	German	77.20
	Australian	78.26
Decision tree	German	85.50
	Australian	90.72
KNN	German	72.20
	Australian	89.10
K-means	German	79.20
	Australian	80.40
MLP	German	73.00
	Australian	86.95
ELM	German	96.33
	Australian	96.32
SVM	German	78.40
	Australian	85.94
ANN	German	77.45
	Australian	82.56

VI. RESULT & DISCUSSION

The performance of these classifiers evaluates using the two credit risk dataset which is widely used i.e. German and

Australian datasets. Many more datasets are used for the classifier evaluation. Most of the evaluation is carried out in MATLAB software environment. Many- survey paper shows results that the ELM classifier gives better accuracies and is faster than any other classifiers. Now this classifier is used in the financial field for credit evaluation and referred as the best classifier in the credit evaluation environment.

VII. CONCLUSION & FUTURE WORK

It is worth keep in mind that objective of the paper is to surveying on the different classifier which are used in the credit risk evaluation. In this paper different types of classifiers are discussed and also different types of ensemble classifiers are briefed. The dataset which are used in the classifier is discussed in the paper. We have analyzed and compare their accuracies using different types classifiers and from comparison table we found that the ELM classifier gives better accuracies compare to other classifiers that is ELM gives 96.33(%) in German dataset and 96.32(%) in Australian dataset.

This paper opens the doors for further research in the credit risk using the machine learning classifier and the ELM classifier will also evaluates the noisily dataset and will compare with the advanced classifier which are also introduced in the financial field.

References

[1] Bask, A., Merisalo-Ratanen, H., Tinnila, M. and Lauraeus, T., 'Towards e-banking: the evolution of business models in financial services', International Journal of Electronic Finance, Vol. 5(4), pp. 333–356,2011.

[2] Bekhet, H.A., Al-alak, B.A., 'Measuring e-statement quality impact on customer satisfaction and loyalty', International Journal of Electronic Finance, Vol. 5(4), pp.299–315, 2011.

[3] Curran, K., Orr, J., 'Integrating geolocation into electronic finance applications for additional security', International Journal of Electronic Finance, Vol. 5(3), pp.272–285,2011.

[4] Cawley, G., Talbot, N., 'Improved sparse least squares support vector machines, Neurocomputing, Vol. 48(1–4), pp. 1025–1031, 2002.

[5] Chorowski, J., Wang, J., Zurada, M.J., 'Review and comparison of SVM and ELM based classifiers', Neurocomputing, Vol. 128, pp. 506-516, 2014.

[6] Chen, M.C., Huang, S.H, 'Credit scoring and rejected instances reassigning through evolutionary computation techniques', Expert Systems with Applications, Vol. 24(4), pp. 433–441, 2003.

[7] Danenas, P., Garsva, G., Gudas, S., 'Credit risk evaluation using SVM classifier', International Conferences On Computational Science, pp.1699-1709, 2011.

[8] Dietterich, T.G., 'Experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization', Machine Learning, Vol. 40, pp.139–157, 2000.

[9] D. West,' Neural network credit scoring models', Computers and Operations Research, Vol. 27 (11/12), pp. 1131–1152, 2000.

[10] Danenas, P., Grasva, G., 'Selection Of Support Vector Machine Based Classifier For Credit Risk', Expert System With Application , Vol. 42, pp. 3194-3204, 2015.

[11] Dhaiya, S., Singh, N.P., 'Impact of Bagging on MLP classifier', International Conferences On Computing For Sustainable Global Development, pp. 3794-3799. 2016.

[12] Huang, B-G., Zhu, Q-Y., 'Extreme Learning Machine : Theory and Application', Neurocomputing, Vol. 70, pp. 489-501, 2006.

[13] Hui, Xiang., Yang, S.G., 'Using clustering-based bagging ensemble for credit scoring'. Business Management and Electronic Information (BMEI), 2011 International Conference on. Vol. 3. IEEE, 2011.

[14] Huang, J., Chen, H., Hsu, C.J., 'Credit rating analysis with SVM and neural network : A market comparative study', Decision Support System, Vol. 37, pp. 543-558, 2004.

[15] Hearst, M.A., Dumais, S.T., Osman, E., Platt, J., 'Support vector machines', IEEE Intelligent Systems, Vol. 13 (4), pp. 18– 28, 2008.

[16] Huang, G.B., Chen, L., Siew, C.K., ' Universal approximation using incremental networks with random hidden computation nodes', IEEE Trans. Neural Networks, Vol. 17(4), pp. 1243-1289, 2006.

[17] Islam, J.M., Wu, J.Q.M., Ahmadi, M., 'Investigating The Performance Of Naïve-Baysed and K- Nearest Neighbor Classifiers', International Conferences On Convergence Information Technology, pp. 1541-1546. 2007.

[18] Jain, A., Kumar, A.M., ' Hybrid neural network models for hydrologic time series Forecasting', Applied Soft Computing, Vol. 7 (2), pp. 585–592, 2007.

[19] Kim, Y. S., Sohn, S. Y., ' Managing loan customers using misclassification patterns of credit scoring model'. Expert Systems with Applications, 26(4), pp. 567–573, 2004.

[20] Kruppa, J., Schwrz, A., Arminger, G. and Zeigler, A., 'Consumer credit risk : Individual probability estimate using machine learning', Expert System with Application, Vol. 40, pp. 5125-5131, 2013.

[21] Li, F.C., Wang, P.K., Wang, G.E., 'Comparison of primitive classifier with ELM for credit scoring', Procceding of IEEE IEEM, pp. 685-688, 2009.

[22] Lee, T-S., Chiu, C-C., Chou, Y-C., Lu, C-J., 'Mining the customer credit using classification and regression tree and multivariate adaptive regression splines', Computational Statistics and Data Analysis, Vol. 50, pp.1113–1130, 2006.

[23] Malhotra, R., Malhotra, D. K, ' Evaluating consumer loans using neural networks', Omega, Vol. 31, pp. 83–96, 2003.

[24] Olafsson, S., Li, X., Wu, S., ' Operations research and data mining', European Journal of Operational Research, Vol. 87, pp. 1429–1448, 2008.

[25] Pandey, T.N., Jagadev, A.K., Choudhury, D. and Dehuri, S., 'Machine learning-based classifiers ensemble for credit risk assessment', Int. J. Electronic Finance, Vol. 7(3/4), pp.227–249, 2013.

[26] Shih, K-H., Hung, H-F., Lin, B., 'Construction of classification models for credit policies in banks', Int. J. of Electronic Finance, Vol. 4(1), pp.1–18, 2010.

[27] Sokolova, M., Lapalme, G., 'A systematic analysis of performance measures for classification tasks', Journal of Information Processing and Management, Vol. 45, pp.427–437, 2009.

[28] Tsai, C-F., Chen, M-L., 'Credit rating by hybrid machine learning technique', Applied Soft Computing, Vol. 10, pp. 374-380, 2010.

[29] Wang, G., Ma, J., 'A hybrid ensemble approach for enterprise credit risk assessment based on SVM', Expert System with Application, Vol. 39, pp. 5325-5331, 2012.

[30] Wang, Y., Wang, S., Lai, K.K., 'A New Fuzzy SVM To Evaluate Credit Risk', IEEE transaction on fuzzy system , Vol. 13, pp. 820.831, 2005.

[31] Wang, G., Hao, J., Ma, J.,'A comparative assessment of ensemble learning for credit scoring', Expert System with Application, Vol. 38, pp. 223-230, 2011.

[32] West, D., Dellana, S. and Qian, J., 'Neural network ensemble strategies for financial decision applications', Computers and Operations Research, Vol. 32, pp.2543–2559, 2005.

[33] West, D, 'Neural network credit scoring models', Computers & Operations Research, Vol. 27(11–12), pp.1131–1152, 2000.

[34] Yang, L., Xiaohvi, Y., Jimmy, X.H. and Aijun, A., 'Combining integrated sampling with SVM ensembles for learning from imbalanced dataset', Journal of Information Processing and Management, Vol. 47, pp.617–631,2011.

[35] Yu, L., Yue, W., Wang, S., Lai, K.K., 'SVM based multiagent ensemble learning for credit risk evaluation', Expert System with Application, Vol. 37, pp.1351-1360, 2010.

[36] Zhang, Z., Gao, G., Shi, Y., 'Credit risk analysis using multi-criteria optimization classifier with kernel, fuzzyfication and penalty factor', European Journal of Operational Research, Vol. 237, pp.335-348, 2014.

[37] Zhong, H., Miao C., Shen, Z., Feng, Y., 'Comparing the learning effectiveness of BP, ELM, I-ELM and SVM for corporate credit rating', Neurocomputing, Vol. 128, 285-295, 2014.

[38] Zhou, H., Lan, Y., Soh, Y.C., Huang, G.B., 'Credit risk evaluation using Extreme Learning Machine', IEEE International Conferences on System, Man and Cybernetics, pp.1064-1069, 2012.