ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Adeptness Comparison between Instance Based and K Star Classifiers for Credit Risk Scrutiny

C. Lakshmi Devasena[1]

Department of Operations and IT, IBS, Hyderabad, IFHE University, Hyderabad, Tamilnadu, India[1]

ABSTRACT —Banking Industry is a significant source of finance. In Banking, Credit Risk assessment is a crucial and decisive task to sanction loans. Automation of decision making for sanctioning loans by analyzing the credit risk of customers using best algorithms and classifiers is of important need today. This work evaluates and compares the adeptness between Instance based classifier and K Star Classifier for credit risk assessment. German credit data is taken as a sample data for adeptness estimation. The performances of both classifiers are analyzed using machine learning tool and a practical guideline for selecting well suited classifier for credit risk assessment is presented. In addition, some diplomatic criteria for evaluating and relating best classifier are discussed.

KEYWORDS—Credit Risk Analysis; IBk Classifier; K Star Classifier

I. INTRODUCTION

Important activity of banking industry is to lend money to those who are in need of money. In order to pay back the principal borrowed from depositors, this industry collect interests on the payments made by the borrower. The borrowers, who fail to make their payments, have defaulted on their guarantee of settlement. To minimize the defaulter rate banking industry need to evaluate the credit details of borrowers. This needs improved research in credit risk assessment. This paper evaluates the Adeptness of Instance Based Classifiers and K Star Classifier for Credit Risk Assessment of banking customers and compared their results. These classifiers fall in the family of Memory based learning. Memory-based learning compares new problem instances with training instances, which are already available in memory. It got its name because it constructs hypotheses directly from the training instances themselves. It has different advantages like, very fast training, learning complex target functions easily, No information loss, and the ability to adapt its model to previously unseen data. Motivated by the need of such requirement, in this work, adeptness evaluation of Instance based classification and K Star Classifier for their suitability on credit data analysis is examined and compared.

II. LITERATURE REVIEW

Many researchers have made the analysis of credit risk using various computing techniques at different stages. A neural network based system for automatic support to credit analysis in a real world problem is presented in [1] & [2]. Hybrid method for evaluating credit risk using Kolmogorove-Smirnov test, DEMATEL method and a Fuzzy Expert system is explained in [3]. The credit risk for a Tunisian bank through modeling the default risk of its commercial loans is analyzed in [4]. An integrated back propagation neural network with traditional discriminant analysis approach is used to explore the performance of credit scoring in [5]. An application of artificial neural network to credit risk assessment using two different architectures are discussed in [6]. A comparative study of corporate credit rating analysis using support vector machines (SVM) and back propagation neural network (BPNN) is analyzed in [7]. Modeling framework for credit assessment models is constructed by using different modeling procedures and performance is analyzed in [8]. A triple-phase neural network ensemble technique with an uncorrelation maximization algorithm is used in a credit risk evaluation system to discriminate good creditors from bad ones is explained in [9].Artificial neural networks using Feed-forward back propagation neural network and business rules to correctly determine credit defaulter is proposed in [10]. Credit risk analysis using different Data Mining models like C4.5, NN, BP, RIPPER, LR and SMO are compared in [11]. Credit risk assessment using six stage neural network ensemble learning approach is discussed in [12]. This research work compares the adeptness of Instance based classifier and K Star Classifier for the effective assessment of credit risk.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

III. DATA USED FOR ANALYSIS

The German credit data is taken for credit data analysis [13]. It consists of 20 attributes and 1000 instances. It has two classes, namely, good and bad. The class is obtained based on the values of all the 20 attributes. Table 1 lists the attributes of Credit-g dataset.

TABLE I. CREDIT-G DATASET – LIST OF ATTRIBUTES

Attribute	Type
Checking account status	Nominal
Duration of credit in months	Continuous
Credit history	Nominal
Purpose of credit	Nominal
Credit amount	Continuous
Average balance in savings account Nominal	Present employment
Nominal	Installment rate as percentage of disposable income
Continuous	Personal status
Nominal	Other parties
Nominal	Present resident since (–) years
Continuous	Property magnitude
Nominal	Age in years
Continuous	Other payment plans
Nominal	Housing
Nominal	Number of existing credits at this bank
Continuous	Nature of job
Nominal	Number of people for whom liable to provide maintenance
Continuous	Applicant has phone in his or her name
Nominal	Foreign worker
Nominal	Class (Reject/Accept) Nominal

IV. METHODOLOGY USED

The Classifiers used for its adeptness to do credit risk evaluation are listed below.

A. IBk Classifier

IB_K is an implementation of the k-nearest-neighbor classifier. Each case is considered as a point in multi-dimensional space and classification is done based on the nearest neighbors. The value of 'k' for nearest neighbor can vary. This determines how many neighbors can be considered to decide how to classify an unknown instance. For example, for the 'German credit' data, IB_K would consider the 20 dimensional spaces for the 20 input variables. A new instance would be classified as belonging to the class of its closest neighbor using Euclidean distance measurement. If the value of 'k' is 6, then 6 closest neighbors are considered. The class of the new instance is considered to be the class of the majority of the instances. If 6 is used as the value of k and 4 of the closest neighbors are of type 'Good', then the class of the test instance would be assigned as 'Good'. The time taken to classify a test instance with nearest-neighbor classifier increases linearly with the number of training instances kept in the classifier. Its performance degrades rapidly with increasing

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

noise levels. It also performs poor, when different attributes affect the outcome to different extents. One parameter that can affect the performance of the IB_K algorithm is the number of nearest neighbors to be used. By default it uses just one nearest neighbor.

B. K Star Classifier

K-Star is a memory-based classifier that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. The use of entropy as a distance measure has several benefits. Amongst other things it provides a consistent approach to handling of symbolic attributes, real valued attributes and missing values. K* is an instance-based learner which uses such a measure [14].

1) Specification of K*

Let I be a (possibly infinite) set of instances and T a finite set of transformations on I. Each t □ T maps instances to instances: t: I → I. T contains a distinguished member □ (the stop symbol) which for completeness maps instances to themselves (□ (a) = a). Let P be the set of all prefix codes from T* which are terminated by σ. Members of T* (and so of P) uniquely define a transformation on I: t (a) = tn (tn-1 (... t1 (a) ...)) where t = t1...tn

A probability function p is defined on T*. It satisfies the following properties:

$$0 \le \frac{p(\bar{tu})}{p(\bar{t})} \le 1$$
$$\sum_u p(\bar{tu}) = p(\bar{t})$$
$$p(\Lambda) = 1 \qquad (1)$$

As a consequence it satisfies the following:

$$\sum_{\bar{t} \in P} p(\bar{t}) = 1 \qquad (2)$$

The probability function P* is defined as the probability of all paths from instance 'a' to instance 'b':

$$P^*(b|a) = \sum_{\bar{t} \in P : t(a) = b} p(\bar{t}) \qquad (3)$$

It is easily proven that P* satisfies the following properties:

$$\sum_b P^*(b|a) = 1$$
$$0 \le P^*(b|a) \le 1 \qquad (4)$$

The K* function is then defined as:

$$K^*(b|a) = -\log_2 P^*(b|a) \qquad (5)$$

K* is not strictly a distance function. For example, K*(a|a) is in general non-zero and the function (as emphasized by the | notation) is not symmetric. Although possibly counter-intuitive the lack of these properties does not interfere with the development of the K* algorithm below. The following properties are provable:

$$K^*(b|a) \ge 0$$
$$K^*(c|b) + K^*(b|a) \ge K^*(c|a) \qquad (6).$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

V. CRITERIA USED FOR CLASSIFICATION EVALUATION

The comparison of the results is made on the basis of the following criteria.

C. Accuracy Classification

All classification result could have an error rate and it may fail to classify correctly. So accuracy can be calculated as follows.

Accuracy = (Instances Correctly Classified / Total Number of Instances)*100 %
(7)

D. Mean Absolute Error (MAE)

MAE is the average of difference between predicted and actual value in all test cases. The formula for calculating MAE is given in equation shown below:

$$MAE = (|a1 - c1| + |a2 - c2| + \ldots + |an - cn|) / n \quad (8)$$

Here 'a' is the actual output and 'c' is the expected output.

E. Root Mean Square Error (RMSE)

RMSE is used to measure differences between values predicted by a model and the values actually observed. It is calculated by taking the square root of the mean square error as shown in equation given below:

$$RMSE = [\sqrt{((a_1 - c_1)^2 + (a_2 - c_2)^2 + \ldots + (a_n - c_n)^2)}]/n \quad (9)$$

Here 'a' is the actual output and c is the expected output. The mean-squared error is the commonly used measure for numeric prediction.

F. Confusion Matrix

A confusion matrix contains information about actual and predicted classifications done by a classification system.

The classification accuracy, mean absolute error, root mean squared error and confusion matrices were calculated using the machine learning tool.

VI. RESULTS AND DISCUSSION

This work is carried out using Open Access Machine learning tool to evaluate the adeptness of instance based classifier and K-Star classifier for credit risk assessment.

G. Performance of IBk Classifier

The performance of the IBk classifier for credit risk analysis is shown below. Table 2 summaries the overall performance of IBk classifier in terms of Correctly Classified Instances, Classification Accuracy, Kappa statistics, RMSE and MAE values, etc. Table 3 to Table 7 shows the Confusion matrix of IBk classifier for the training data set and other Cross Validation (CV) techniques used. Correctly Classified instances of IBk classifier is shown in Fig 1 and the Classification Accuracy obtained by IBk classifier is shown in Fig 2.

TABLE II. IBK CLASSIFIER OVERALL EVALUATION SUMMARY

	Training Set	5 Fold CV	10 Fold CV	15 Fold CV	20 Fold CV
Correctly Classified Instances	1000	705	720	715	723
Accuracy	100%	70.5%	72%	71.5%	72.3%
Kappa statistic	1	0.2805	0.3243	0.3076	0.3257
MAE	0.001	0.2955	0.2805	0.2855	0.2775

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

RMSE	0.001	0.5425	0.5286	0.5333	0.5258
RAE	0.2375 %	70.3264%	66.7546 %	67.9388 %	66.0375%
RRSE	0.2178 %	118.375 %	115.3422 %	116.3719 %	114.7293%

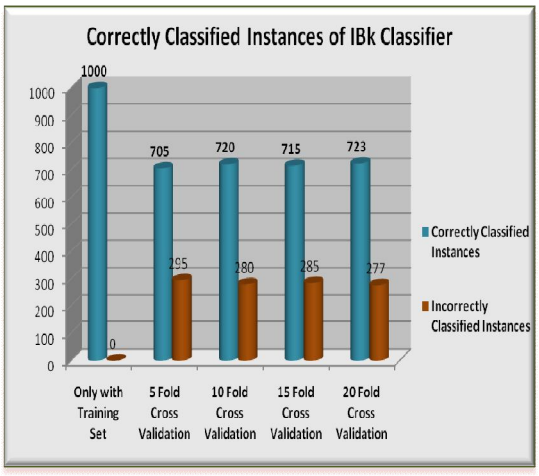

Fig.1. Correctly Classified Instances by IBk Classifier obtained by IBk classifier

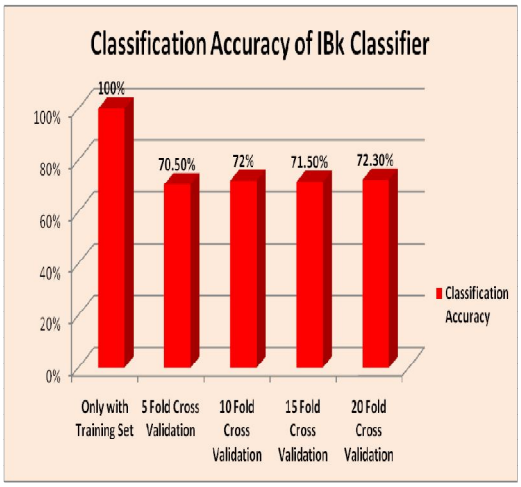

Fig. 2. Classification Accuracy

TABLE III. CONFUSION MATRIX – IBK CLASSIFIER (ON TRAINING DATA)

	Good	**Bad**
Good	700	0
Bad	0	300

TABLE IV. CONFUSION MATRIX – IBK CLASSIFIER (5 FOLD CROSS VALIDATION)

	Good	**Bad**
Good	565	135
Bad	160	140

TABLE V. CONFUSION MATRIX – IBK CLASSIFIER (10 FOLD CROSS VALIDATION)

	Good	**Bad**
Good	570	130
Bad	150	150

TABLE VI. CONFUSION MATRIX – IBK CLASSIFIER (15 FOLD CROSS VALIDATION)

	Good	**Bad**
Good	568	132
Bad	153	147

TABLE VII. CONFUSION MATRIX – IBK CLASSIFIER (20 FOLD CROSS VALIDATION)

	Good	**Bad**
Good	573	127
Bad	150	150

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

H. Performance of K Star Classifier

The performance of the K Star classifier for credit risk analysis is shown below. Table 8 to Table 12 shows the Confusion matrix of K Star classifier for the training data set and other Cross Validation (CV) techniques used. Table 13 summaries the overall performance of K Star classifier in terms of Correctly Classified Instances, Classification Accuracy, Kappa statistics, RMSE and MAE values, etc. Correctly Classified instances of K Star classifier is shown in Fig 3 and the Classification Accuracy obtained by K Star classifier is shown in Fig 4.

TABLE VIII. CONFUSION MATRIX – K STAR CLASSIFIER (ON TRAINING DATA)

	Good	Bad
Good	700	0
Bad	0	300

TABLE IX. CONFUSION MATRIX – K STAR CLASSIFIER (5 FOLD CROSS VALIDATION)

	Good	Bad
Good	573	127
Bad	171	129

TABLE X. CONFUSION MATRIX – K STAR CLASSIFIER (10 FOLD CROSS VALIDATION)

	Good	Bad
Good	569	131
Bad	175	125

TABLE XI. CONFUSION MATRIX – K STAR CLASSIFIER (15 FOLD CROSS VALIDATION)

	Good	Bad
Good	567	133
Bad	167	133

TABLE XII. CONFUSION MATRIX – K STAR CLASSIFIER (20 FOLD CROSS VALIDATION)

	Good	Bad
Good	576	124
Bad	173	127

TABLE XIII. IBK CLASSIFIER OVERALL EVALUATION SUMMARY

	Training Set	5 Fold CV	10 Fold CV	15 Fold CV	20 Fold CV
Correctly Classified Instances	1000	702	694	700	703
Accuracy	100%	70.2%	69.4%	70%	70.3%
Kappa statistic	1	0.2594	0.2396	0.2618	0.2582
MAE	0	0.3117	0.3148	0.3144	0.3123
RMSE	0.0009	0.4853	0.4884	0.4865	0.4854
RAE	0.0118 %	74.1765%	74.9091 %	74.8164 %	74.321%
RRSE	0.2065 %	105.8905 %	106.5831 %	106.1539 %	105.9214%

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

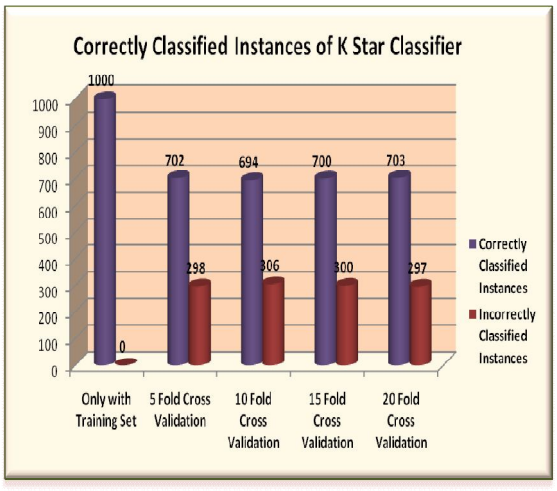

Fig.3. Correctly Classified Instances by K Star Classifier

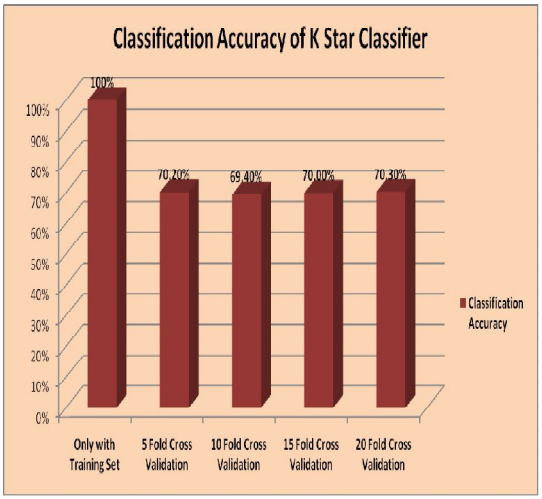

Fig. 4. Classification Accuracy obtained by K Star classifier

The comparison between IBk classifier and K Star classifier are shown in Fig 5 and Fig 6 in terms of classification accuracy and Correctly Classified Instances. The overall ranking is done based on the classification accuracy, MAE and RMSE values and other statistics found using Training Set results and Cross Validation Techniques. Based on that, it is concluded that IBk classifier performs better than K Star.

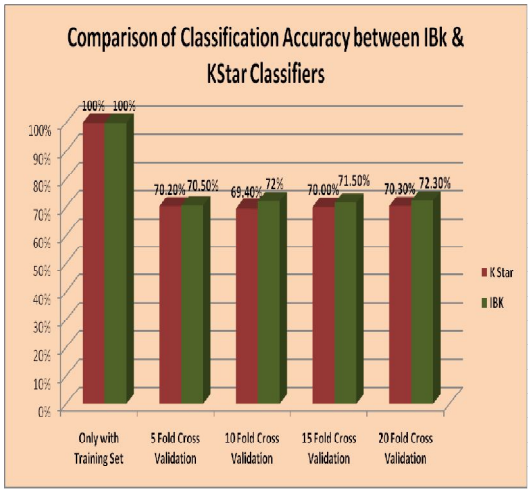

Fig. 5. Comparison of Classification Accuracy b/n IBk & K Star Instances b/n IBk & K Star

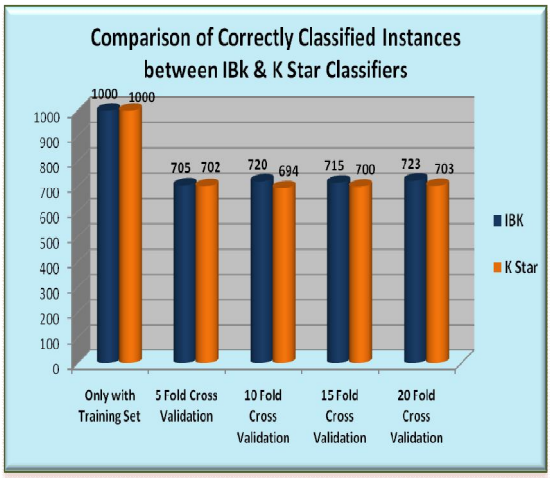

Fig. 6.Comparison of Correctly Classified

VII. CONCLUSION

This research work investigated the adeptness of Instance Based classifier and K Star Classifier for credit risk assessment. Experimentation is done using the open source machine learning tool. Adeptness evaluation of both classifiers has been done and a practical guideline for selecting the renowned and more suited algorithm for credit risk

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

assessment is presented. After experimentation it is concluded that IBk Classifier performs better than K star Classifier in Credit Risk Assessment.

ACKNOWLEDGMENT

The author acknowledges the Management of IBS, Hyderabad for providing motivation and support.

REFERENCES

[1] Germano C. Vasconcelos, Paulo J. L. Adeodato and Domingos S. M. P. Monteiro, "A Neural Network Based Solution for the Credit Risk Assessment Problem", Proceedings of the IV Brazilian Conference on Neural Networks - IV Congresso Brasileiro de Redes Neurais pp. 269-274, July 20-22, 1999 - ITA, São José dos Campos - SP – Brazil.

[2] Vincenzo Pacelli and Michele Azzollini, "An Artificial Neural Network Approach for Credit Risk Management", Journal of Intelligent Learning Systems and Applications, 2011, 3, 103-112.

[3] Sanaz Pourdarab, Ahmad Nadali and Hamid Eslami Nosratabadi, "A Hybrid Method for Credit Risk Assessment of Bank Customers" , *International Journal of Trade, Economics and Finance, Vol. 2, No. 2, April 2011.*

[4] Hamadi Matoussi and Aida Krichene, "Credit risk assessment using Multilayer Neural Network Models - Case of a Tunisian bank" 2007.

[5] Tian-Shyug Lee, Chih-Chou Chiu, Chi-Jie Lu and I-Fei Chen, "Credit scoring using the hybrid neural discriminant technique", Expert Systems with Applications (Elsevier) 23 (2002), pp. 245–254.

[6] Eliana Angelini, Giacomo di Tollo, and Andrea Roli "A Neural Network Approach for Credit Risk Evaluation", Kluwer Academic Publishers, 2006, pp. 1 – 22.

[7] Zan Huang, Hsinchun Chena, Chia-Jung Hsu, Wun-Hwa Chen and Soushan Wu, "Credit rating analysis with support vector machines and neural networks: a market comparative study", Decision Support Systems (Elsevier) 37 (2004) pp. 543– 558.

[8] Arnar Ingi Einarsson, "Credit Risk Modeling", Ph.D Thesis, Technical University of Denmark, 2008.

[9] Kin Keung Lai, Lean Yu, Shouyang Wang, and Ligang Zhou, "Credit Risk Analysis Using a Reliability-Based Neural Network Ensemble Model", S. Kollias et al. (Eds.): ICANN 2006, Part II, LNCS 4132, pp. 682 – 690, 2006.© Springer-Verlag Berlin Heidelberg.

[10] A.R.Ghatge, P.P.Halkarnikar, " Ensemble Neural Network Strategy for Predicting Credit Default Evaluation" International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 7, January 2013 pp. 223 – 225.

[11] S. Kotsiantis, "Credit risk analysis using a hybrid data mining model", Int. J. Intelligent Systems Technologies and Applications, Vol. 2, No. 4, 2007, pp. 345 – 356.

[12] Lean Yu, Shouyang Wang, Kin Keung Lai, "Credit risk assessment with a multistage neural network ensemble learning approach", Expert Systems with Applications (Elsevier) 34 (2008) 1434–1444.

[13] UCI Machine Learning Data Repository – http://archive.ics.uci.edu/ml/datasets.

[14] John G. Cleary, "K*: An Instance-based Learner Using an Entropic Distance Measure".