COMPETENCY COMPARISON BETWEEN LOGISTIC CLASSIFIER AND PARTIAL DECISION TREE CLASSIFIER FOR CREDIT RISK PREDICTION

Lakshmi Devasena C

Dept. of Operations & Systems, ISB Hyderabad, IFHE University

ABSTRACT

Credit Risk prediction is a critical task of any Financial Industry like Banks. Discovering dodger before giving loan is a momentous and conflict-ridden task of the Banker. Classification techniques can be used to find the claimant, whether he/she is a cheat or an unpretentious customer. Determining the outstanding classifier is a precarious assignment for any industrialist like a banker. It leads to drill down efficient research works through evaluating different classifiers and finding out the best classifier for the credit risk approximation. This research work investigates the efficiency of Partial Decision Tree Classifier and Logistic Classifier for the credit risk prediction and compares their competence through various measures. To predict the classifier performance, German credit dataset has been taken, and open source machine learning tool is used.

1. INTRODUCTION

The enormous volume of transactions made information processing automation an invigorating factor for high quality standards, cost reduction, with high speed results. Data analysis automation and result of the relevant successes produced by state-of-the art computer algorithms have changed the opinions of many misanthropists. In the past, people thought that financial market analysis necessitates intuition, knowledge and experience and speculated how this job could be automated. Conversely, growth of scientific and technological advances, achieved the automation of financial market analysis. In recent days, credit defaulter prediction and credit risk evaluation have fascinated great deal of interests from regulators, practitioners, and theorists, in the financial industry. Since, the credit score of an applicant could be calculated from the past giant database and the demographic data, it needs automation. Automation of credit risk forecast can be achieved using classification techniques. Selecting the classifier, which envisages credit risk in an efficient manner, is an imperative and critical task. This work appraises the credit risk performance of two diverse classifiers, namely, Logistic Classifier and Partial Decision Tree Classifier and compares their accuracy of credit risk prediction.

2. LITERATURE REVIEW

There are many research works made to predict credit risk using wide-ranging computing techniques. In [1], a neural network based algorithm for automatic provisioning to credit risk scrutiny in a real world problem is presented. An assimilated back propagation neural network (BPNN) with the customary

discriminant analysis approach used to discover the performance of credit scoring is given in [2]. A comparative study of corporate credit rating analysis using back propagation neural network (BPNN) and support vector machines (SVM) is described in [3]. An uncorrelated maximization algorithm within a triple-phase neural network ensemble technique for credit risk evaluation to differentiate good creditors from bad ones are elucidated in [4]. An application of artificial neural network to credit risk assessment using two altered architectures are deliberated in [5]. Credit risk investigation using diverse Data Mining models like C4.5, NN, BP, RIPPER, LR and SMO is likened in [6]. The credit risk of a Tunisian bank through modeling the non-payment risk of its commercial loans is analyzed in [7]. Credit risk valuation using six stage neural network ensemble learning approach is argued in [8]. A modeling framework for credit calculation models is erected using different modeling procedures is explained and its performance is analyzed in [9]. Hybrid method for assessing credit risk using Kolmogorove-Smirnov test, Fuzzy Expert system and DEMATEL method is enlightened in [10]. An Artificial Neural Network centered methodology for Credit Risk supervision is proposed in [11]. Artificial neural networks using Feed-forward back propagation neural network and business rules to correctly determine credit defaulter is proposed in [12]. The performance comparison of Memory based classifiers for credit risk investigation is experimented and précised in [13]. The performance comparison between Instance Based and K Star Classifiers for Credit Risk Inspection is accomplished and pronounced in [14]. The performance comparison among Sequential Minimal Optimization and Logistic Classifiers for Credit Risk Calculation is specified in [15]. The performance comparison between Multilayer Perceptron and SMO Classifier for Credit Risk appraisal is described in [16]. The performance comparison between JRip and PART Classifier for Credit Risk Estimation is explored in [17]. This research work compares the efficiency of Logistic classifier and Partial Decision Tree Classifier for credit risk prediction.

3. DATASET USED

The German credit data is used to evaluate the performance of Logistic classifier and Partial Decision Tree Classifier for credit risk prediction. This data set contains 20 attributes, namely, Duration, Credit History, Checking Status, Purpose, Credit Amount, Employment, Installment Commitment, Saving Status, Personal Status, Other parties, Property magnitude, Age, resident since, Other payment plans, existing credits, job, Housing, No. of dependents, Foreign worker and Own Phone. The data set comprises 1000 instances of client credit data with class detail. It discriminates the records into two classes, namely, good and bad.

4. METHODOLOGY USED

In this research work, two diverse classifiers namely, Partial Decision Tree Classifier and Logistic Classifier are compared for proficiency assessment of credit risk estimation.

4.1 Partial Decision Tree Classifier

Partial Decision Tree Classifier integrates the separate-and-conquer strategy of rule learning with the divide-and-conquer strategy to predict the new data. The generalized algorithm of this classifier is given below.

1. Construct a partial decision tree on the current set of instances

2. Generate a rule from the decision tree. i.e., the rule is made from the leaf with the largest coverage
3. Remove the decision tree
4. Eliminate the instances covered by the rule
5. Repeat from step one

4.2 Logistic Classifier

Logistic Classifier is a generalization of linear regression classifier [19]. It is mainly used for evaluating binary or multi-class reliant variables and the retort variable is discrete, it cannot be demonstrated directly by linear regression i.e. discrete variable transmuted into incessant value. Logistic classifier predominantly used to categorize low dimensional data having non-linear boundaries. It also affords the difference in the percentage of dependent variable and provides the rank of the individual variable according to its significance. So, the main dictum of Logistic classifier is to determine the result of each variable correctly. Logistic classifier is also known as logistic model or logit model that deliver categorical variable for the target variable with two classifications such as good and bad.

5. PERFORMANCE MEASURES USED

Various scales are used to gauge the performance of the classifiers.

Classification Accuracy

Any classifier could have an error rate and it may fail to categorize correctly. Classification accuracy is calculated as Correctly classified instances divided by Total number of instances multiplied by 100.

Mean Absolute Error

Mean absolute error is the average of the variance between predicted and actual value in all test cases. It is a good measure to gauge the performance.

Root Mean Square Error

Root mean squared error is used to scale dissimilarities between values actually perceived and the values predicted by the model. It is determined by taking the square root of the mean square error.

Confusion Matrix

A confusion matrix encompasses information about actual and predicted groupings done by a classification system

6. Results And Discussion

The performance of Partial Decision Tree Classifier and Logistic Classifier is experienced using

open source machine learning tool. The performance is checked using the Training set as well as using different Cross Validation methods. The class is attained by considering all 20 attributes of the dataset.

6.1 Performance of Partial Decision Tree Classifier

The overall assessment summary of Partial Decision Tree Classifier using training set and different cross validation methods is given in Table I. The performance of Partial Decision Tree Classifier in terms of Correctly Classified Instances and Classification Accuracy is shown in Fig. 1and Fig. 2. The confusion matrix for different test mode is given in Table II to Table VII. Partial Decision Tree Classifier gives 89.7% for the training data set. Various cross validation methods are used to check its actual performance. On an average, it gives around 70% of accuracy for credit risk estimation.

TABLE I
PARTIAL DECISION TREE CLASSIFIER OVERALL EVALUATION SUMMARY

Test Mode	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy	Mean Absolute Error	Root Mean Squared Error	Time Taken to Build Model (Sec)
Training Set	897	103	89.7%	0.1605	0.2833	3.48
5 Fold CV	688	312	68.8%	0.3348	0.5101	1.84
10 Fold CV	702	298	70.2%	0.3245	0.4974	0.72
15 Fold CV	726	274	72.6%	0.304	0.4828	1.2
20 Fold CV	696	304	69.6%	0.3253	0.499	0.69
50 Fold CV	706	294	70.6%	0.3164	0.4886	1.11

TABLE II
CONFUSION MATRIX – PARTIAL DECISION TREE CLASSIFIER (ON TRAINING DATASET)

	Good	Bad	Actual (Total)
Good	653	47	700
Bad	56	244	300
Predicted (Total)	709	291	1000

TABLE III
CONFUSION MATRIX – PARTIAL DECISION TREE CLASSIFIER (5 FOLD CROSS VALIDATION)

	Good	Bad	Actual (Total)
Good	548	152	700
Bad	160	140	300
Predicted (Total)	608	292	1000

TABLE IV
CONFUSION MATRIX – PARTIAL DECISION TREE CLASSIFIER (10 FOLD CROSS VALIDATION)

	Good	Bad	Actual (Total)
Good	561	139	700
Bad	159	141	300
Predicted (Total)	720	280	1000

TABLE V
CONFUSION MATRIX – PARTIAL DECISION TREE CLASSIFIER (15 FOLD CROSS VALIDATION)

	Good	Bad	Actual (Total)
Good	577	123	700
Bad	151	149	300
Predicted (Total)	728	272	1000

TABLE VI
CONFUSION MATRIX – PARTIAL DECISION TREE CLASSIFIER (20 FOLD CROSS VALIDATION)

	Good	Bad	Actual (Total)
Good	562	138	700
Bad	166	134	300
Predicted (Total)	728	272	1000

TABLE VII
CONFUSION MATRIX – PARTIAL DECISION TREE CLASSIFIER (50 FOLD CROSS VALIDATION)

	Good	Bad	Actual (Total)
Good	560	140	700
Bad	154	146	300
Predicted (Total)	714	286	1000

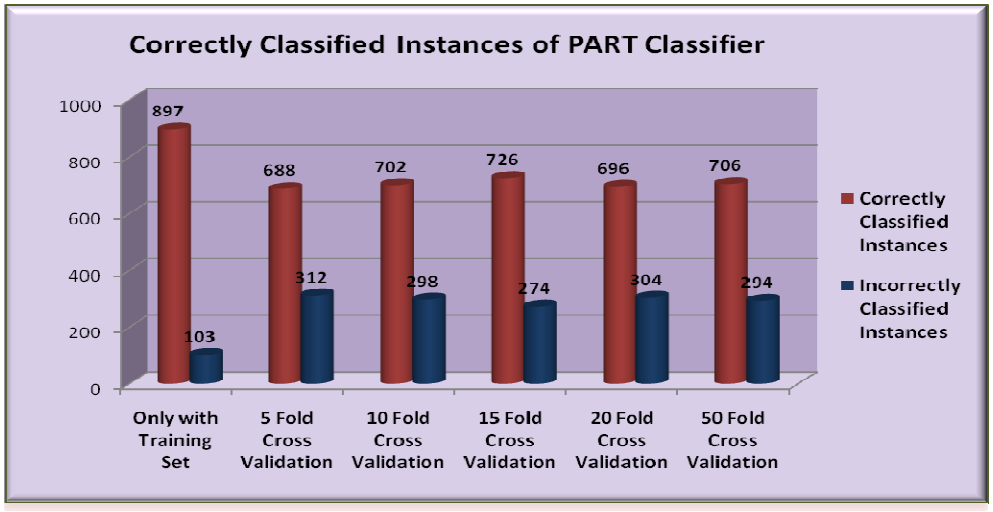

Fig. 1 Correctly Classified instances of Partial Decision Tree Classifier

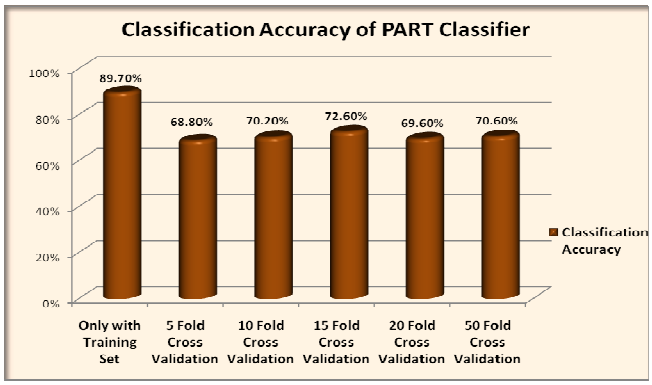

Fig. 2 Classification Accuracy of Partial Decision Tree Classifier

6.2. Performance of Logistic Classifier

The overall assessment summary of Logistic Classifier using training set and different cross validation methods is given in Table VIII. The performance of Logistic Classifier in terms of Correctly Classified Instances and Classification Accuracy is shown in Fig. 3and Fig. 4. The confusion matrix for different test mode is given in Table IX to Table XIV. Logistic Classifier gives 78.6% for the training data set. Various cross validation methods are used to check its actual performance. On an average, it gives around 75.4% of accuracy for credit risk estimation.

TABLE VIIII
LOGISTIC CLASSIFIER COMPLETE EVALUATION SUMMARY

Test Mode	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy	Mean absolute error	Root Mean Squared Error	Time Taken to Build Model (Sec)
Training Set	786	214	78.6%	0.2921	0.3823	0.58
5 Fold CV	757	243	75.7%	0.3067	0.4065	0.14
10 Fold CV	752	248	75.2%	0.3098	0.4087	0.13
15 Fold CV	757	243	75.7%	0.3103	0.4085	0.13
20 Fold CV	754	246	75.4%	0.3106	0.4086	0.14
50 Fold CV	752	248	75.2%	0.3116	0.4084	0.14

TABLE IX
CONFUSION MATRIX – LOGISTIC CLASSIFIER (ON TRAINING DATASET)

	Good	Bad	Actual (Total)
Good	626	74	700
Bad	140	160	300
Predicted (Total)	766	234	1000

TABLE X
CONFUSION MATRIX – LOGISTIC CLASSIFIER (5 FOLD CROSS VALIDATION)

	Good	Bad	Actual (Total)
Good	602	98	700
Bad	145	155	300
Predicted (Total)	747	253	1000

TABLE XI
CONFUSION MATRIX – LOGISTIC CLASSIFIER (10 FOLD CROSS VALIDATION)

	Good	Bad	Actual (Total)
Good	605	95	700
Bad	153	147	300
Predicted (Total)	758	242	1000

TABLE XII
CONFUSION MATRIX – LOGISTIC CLASSIFIER (15 FOLD CROSS VALIDATION)

	Good	Bad	Actual (Total)
Good	610	90	700
Bad	153	147	300
Predicted (Total)	763	237	1000

TABLE XIII
CONFUSION MATRIX – LOGISTIC CLASSIFIER (20 FOLD CROSS VALIDATION)

	Good	**Bad**	**Actual (Total)**
Good	605	95	700
Bad	151	149	300
Predicted (Total)	756	244	1000

TABLE XIV
CONFUSION MATRIX – LOGISTIC CLASSIFIER (50 FOLD CROSS VALIDATION)

	Good	Bad	Actual (Total)
Good	607	93	700
Bad	155	145	300
Predicted (Total)	762	238	1000

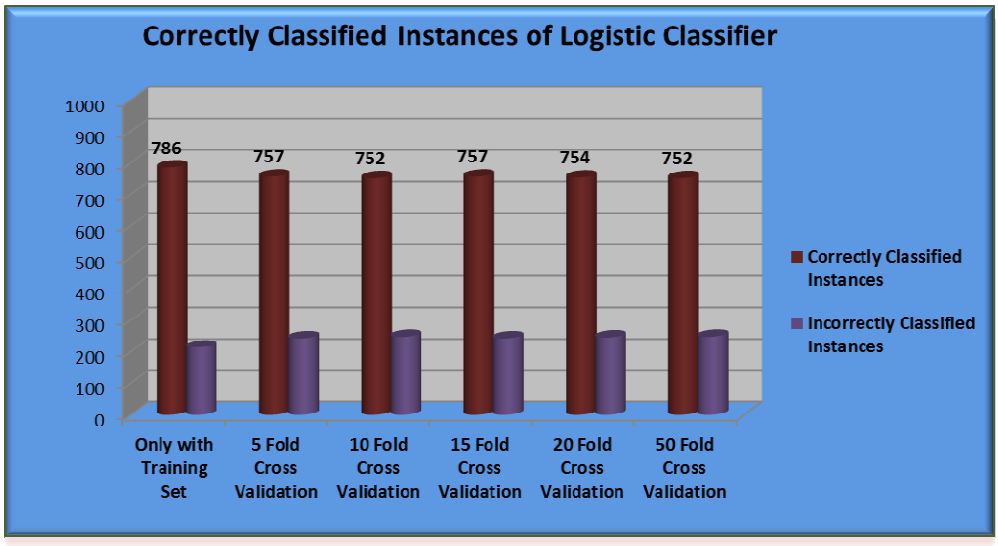

Fig. 3 Correctly Classified instances of Logistic Classifier

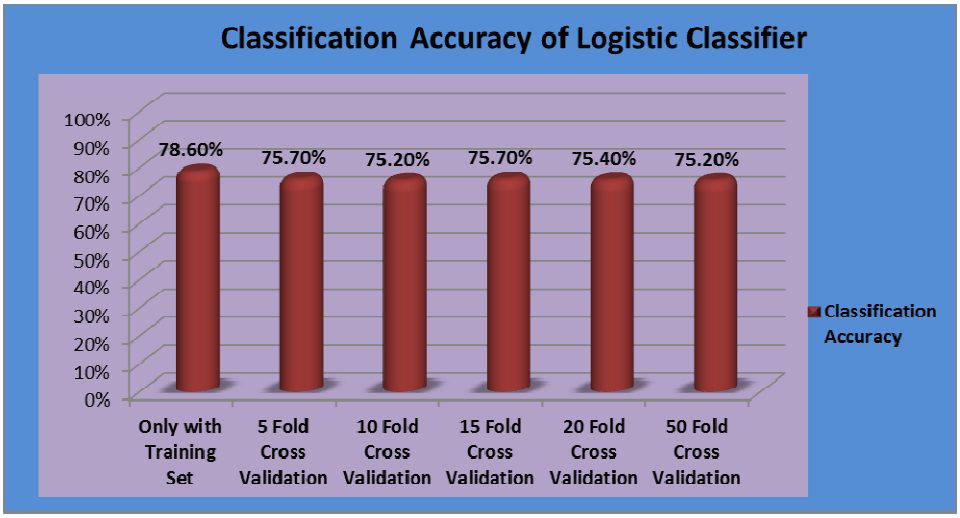

Fig. 4 Classification Accuracy of Logistic Classifier

6.3. Comparison of Partial Decision Tree Classifier and Logistic Classifier

The comparison of performance between Partial Decision Tree Classifier and Logistic Classifier is depicted in Fig 5, and Fig. 6 in terms of Correctly Classified Instances and Classification Accuracy. The complete ranking is prepared based on correctly classified instances, classification accuracy, MAE and RMSE values and other statistics found using Training Set result and Cross Validation Techniques. Consequently, it is perceived that Logistic classifier performs better than Partial Decision Tree Classifier.

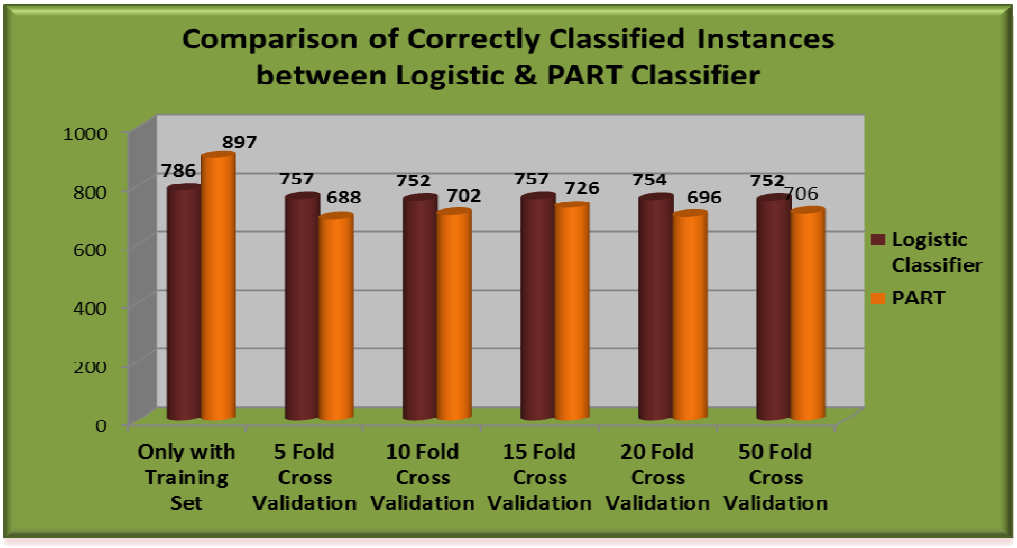

Fig. 5 Correctly Classified Instances Comparison between Partial Decision Tree Classifier and Logistic Classifier

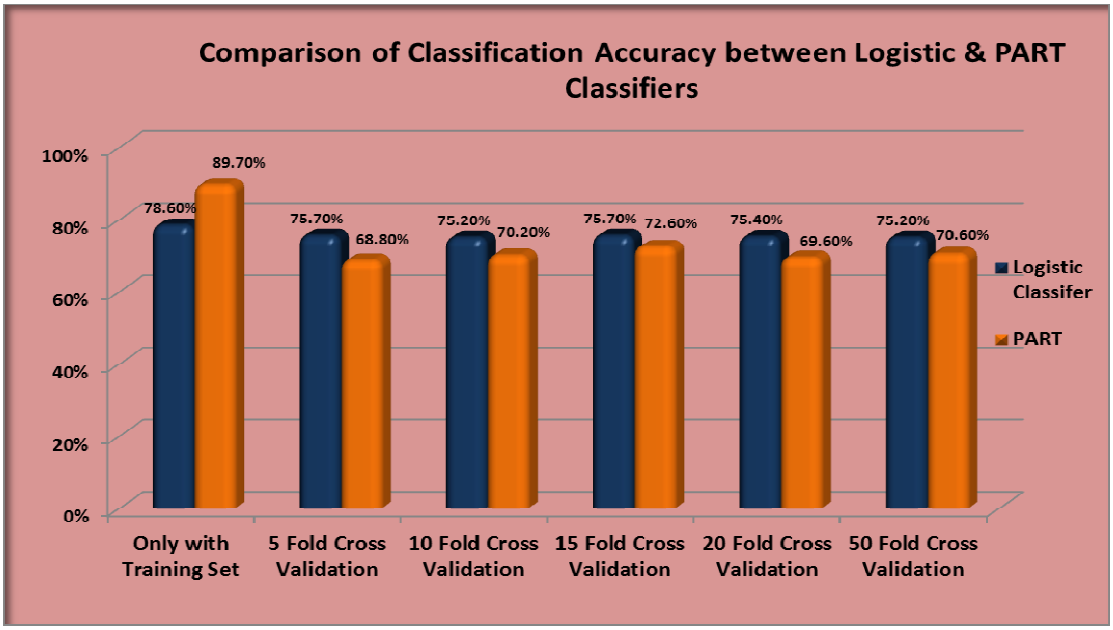

Fig. 5 Classification Accuracy Comparison between Partial Decision Tree Classifier and Logistic Classifier

7. CONCLUSION

This work investigated the efficiency of two different classifiers namely, Partial Decision Tree Classifier and Logistic Classifier for credit risk prediction. Testing is accomplished using the open source machine learning tool. Also, effectiveness comparison of both the classifiers has been done in view of different scales of performance evaluation. At last, it is observed that Logistic Classifier performs better than Partial Decision Tree Classifier for credit risk prediction.

ACKNOWLEDGMENT

The author expresses her gratitude to the Management of IBS Hyderabad, IFHE University and Operations & IT Department of IBS Hyderabad for constant support and motivation.

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.

[2] 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, pp. 245–254, 2002.

[3] 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, pp. 543– 558, 2004.

[4] 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, Springer LNCS 4132, pp. 682 – 690, 2006.

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

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

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

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

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

[10] 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.

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

[12] A.R.Ghatge and 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.

[13] Lakshmi Devasena, C., "Adeptness Evaluation of Memory Based Classifiers for Credit Risk Analysis," Proc. of International Conference on Intelligent Computing Applications - ICICA 2014, 978-1-4799-3966-4/14 (IEEE Explore), 6-7 March 2014, pp. 143-147, 2014.

[14] Lakshmi Devasena, C., "Adeptness Comparison between Instance Based and K Star Classifiers for Credit Risk Scrutiny," International Journal of Innovative Research in Computer and Communication Engineering, Vol.2, Special Issue 1, March 2014.

[15] Lakshmi Devasena, C., "Effectiveness Assessment between Sequential Minimal Optimization and Logistic Classifiers for Credit Risk Prediction," International Journal of Application or Innovation in Engineering & Management, Volume3, Issue 4, April 2014.

[16] Lakshmi Devasena, C., "Efficiency Comparison of Multilayer Perceptron and SMO Classifier for Credit Risk Prediction," International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 4, 2014.

[17] Lakshmi Devasena, C. 2014. Competency Assessment between JRip and Partial Decision Tree Classifiers for Credit Risk Estimation. International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4 (5), May – 2014, pp. 164-173.

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

[19] De Mantaras and Armengol E,"Machine learning from example: Inductive and Lazy methods", Data & Knowledge Engineering 25: 99-123, 1998.