Volume 4, Issue 5, May 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Competency Assessment between JRip and Partial Decision Tree Classifiers for Credit Risk Estimation

Lakshmi Devasena C*
Department of Operations & IT,
IBS Hyderabad, IFHE University, India

Abstract— Credit Risk estimation is an imperative task of any Banking Industry. Discovering defaulter before giving loan is a significant and divisive task of the Banker. Different Classification techniques can be used to find the applicant, whether he/she is a defaulter or a genuine customer. Determining the pre-eminent classifier is a critical assignment for any industrialist like banker. It leads to instill efficient research works through evaluating different classifiers and determining the best classifier for the credit risk estimation. This research work investigates the efficiency of JRip Classifier and PART Classifier for the credit risk estimation and compares their efficiency through various measures. German credit data is taken to predict the classifier performance using open source machine learning tool.

Keywords— Competency Assessment, Credit Risk Estimation, JRip Classifier, Partial Decision Tree, PART Classifier

I. INTRODUCTION

Banking Industry is the largest financial Industry. It earns through lending money to customers through loans. If the loans are not repaid, the jeopardy of customer makes the bank to massive loss or it is bankrupted. To prevent the bank from bankrupting, scrutinizing the applicant details and finding his/her genuineness is very much important. Biographic details of the applicant can be used to predict the defaulter in earlier stage through classification algorithms using similar past available training data. This reduces the burden of banker to scrutinize the applicant's credit risk. This research work evaluates the credit risk performance of two different classifiers, namely, JRip Classifier and PART Classifier and compares them to find which of them provide more accurate credit risk prediction.

II. LITERATURE REVIEW

Existing research literature has evidence for the credit risk prediction using varied computing techniques. A neural network based system for automatic support to credit risk analysis in a real world problem is presented in [1]. An integrated back propagation neural network with traditional discriminant analysis approach is used to explore the performance of credit scoring in [2]. A comparative study of corporate credit rating analysis using support vector machines (SVM) and back propagation neural network (BPNN) is analysed in [3]. 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 are explained in [4]. An application of artificial neural network to credit risk assessment using two different architectures are discussed in [5]. Credit risk analysis using different Data Mining models like C4.5, BP, NN, RIPPER, SMO and LR are compared in [6]. The credit risk for a Tunisian bank through modelling the default risk of its commercial loans is analysed in [7]. Credit risk assessment using six stage neural network ensemble learning approach is discussed in [8]. Modelling framework for credit assessment models is constructed by using different modelling procedures and performance is analysed in [9]. Hybrid method for evaluating credit risk using Kolmogorove-Smirnov test, DEMATEL method and a Fuzzy Expert system is explained in [10]. An Artificial Neural Network based approach for Credit Risk Management 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]. Adeptness evaluation of Memory based classifiers for credit risk analysis is experimented and summarized in [13]. Adeptness comparison of Instance Based and K Star Classifiers for Credit Risk Scrutiny is performed and described in [14]. Effectiveness Assessment between Sequential Minimal Optimization and Logistic Classifiers is compared for Credit Risk Prediction in [15]. Efficiency Comparison between Multilayer Perceptron and SMO Classifier is done for Credit Risk Prediction in [16]. This research work compares the competency of JRip classifier and Partial Decision Tree (PART) Classifier for credit risk estimation.

III. DATA SET USED

For experimentation and estimation of credit risk German credit data is used. It consists of twenty attributes, namely, Checking Status, Credit History, Duration, Purpose, Saving Status, Credit Amount, Employment, Instalment Commitment, Personal Status, Other parties, resident since, Property magnitude, Age, Housing, Other payment plans, existing credits, job, Own Phone, Foreign worker and Num dependents. It has two classes, namely, good and bad. The data set has 1000 instances of customer credit data with the class detail.

IV. METHODOLOGY USED

In this work, two different classifiers namely, JRip Classifier and PART Classifier are used for competency assessment of credit risk estimation.

A. JRip Classifier

JRip (RIPPER) is one of the fundamental and most trendy algorithms. It examines the classes in increasing size and generates an initial set of rules for the class using incremental reduced error. JRip (RIPPER) ensues by treating all the examples of a particular judgment in the training data as a class, and discovers a set of rules that cover all the members of that class. Subsequently it proceeds to the next class and does the same process; repeat this until all classes have been covered.

B. PART Classifier

PART Classifier combines the divide-and-conquer strategy with separate-and-conquer strategy of rule learning. This classifier works as follows.

1. Build a partial decision tree on the current set of instances
2. Create a rule from the decision tree. i.e., the leaf with the largest coverage is made into a rule
3. Discarded the decision tree
4. Remove the instances covered by the rule
5. Go to step one.

V. MEASURES USED FOR PERFORMANCE EVALUATION

The following measures are used for performance evaluation.

A. Classification Accuracy

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

Accuracy = (Instances Correctly Classified / Total Number of Instances)*100 % (1)

B. Mean Absolute Error

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:

$$MAE = (|a1 - c1| + |a2 - c2| + \ldots + |an - cn|) / n \quad (2)$$

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

C. Root Mean Square Error

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

$$RMSE = [\sqrt{((a1 - c1)^2 + (a2 - c2)^2 + \ldots + (an - cn)^2)}]/n \quad (3)$$

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

D. Confusion Matrix

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

VI. RESULTS AND DISCUSSION

The performance of JRip and PART classifiers are tested out using open source machine learning tool. The performance is checked using the Training set itself and using different Cross Validation and Percentage Split methods. The class is obtained by considering the values of all the 20 attributes.

A. Performance of JRip Classifier

The overall evaluation summary of JRip Classifier using training set and different cross validation methods is given in Table I. The classification summary of JRip Classifier for different percentage split is given in Table VIII. The performance of JRip Classifier in terms of Correctly Classified Instances and Classification Accuracy is shown in Fig. 1, Fig. 2 and Fig. 3.The confusion matrix for each different test mode is given in Table III to Table VII and from Table IX to Table XII. JRip Classifier gives 74.3% for the training data set. But for evaluation testing with test data is essential. Therefore, various cross validation and percentage split methods are used to test its actual performance. On an average, it gives around 72% of classification accuracy for credit risk estimation.

TABLE I: JRIP CLASSIFIER OVERALL EVALUATION SUMMARY

Test Mode	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy	Kappa Statistic	Mean absolute error	Root Mean Squared Error	Time Taken to Build Model (Sec)
Training Set	743	257	74.3	0.3464	0.3666	0.4281	0.47
5 Fold CV	699	301	69.9	0.2153	0.386	0.4507	0.27
10 Fold CV	717	283	71.7	0.2153	0.3781	0.4472	0.75
15 Fold CV	726	274	72.6	0.2909	0.3717	0.4394	0.45
20 Fold CV	729	271	72.9	0.2876	0.3724	0.4394	0.23
50 Fold CV	731	269	73.1	0.306	0.3715	0.4387	0.34

TABLE II: CONFUSION MATRIX – JRIP CLASSIFIER (ON TRAINING DATASET)

	Good	**Bad**	**Actual (Total)**
Good	605	95	700
Bad	162	138	300
Predicted (Total)	767	233	1000

TABLE III: CONFUSION MATRIX – JRIP CLASSIFIER (5 FOLD CROSS VALIDATION)

	Good	**Bad**	**Actual (Total)**
Good	595	105	700
Bad	196	104	300
Predicted (Total)	791	209	1000

TABLE IV: CONFUSION MATRIX – JRIP CLASSIFIER (10 FOLD CROSS VALIDATION)

	Good	**Bad**	**Actual (Total)**
Good	611	89	700
Bad	194	106	300
Predicted (Total)	805	195	1000

TABLE V: CONFUSION MATRIX – JRIP CLASSIFIER (15 FOLD CROSS VALIDATION)

	Good	**Bad**	**Actual (Total)**
Good	605	95	700
Bad	179	121	300
Predicted (Total)	784	216	1000

TABLE VI: CONFUSION MATRIX – JRIP CLASSIFIER (20 FOLD CROSS VALIDATION)

	Good	**Bad**	**Actual (Total)**
Good	614	86	700
Bad	185	115	300
Predicted (Total)	799	201	1000

TABLE VII: CONFUSION MATRIX – JRIP CLASSIFIER (50 FOLD CROSS VALIDATION)

	Good	**Bad**	**Actual (Total)**
Good	606	94	700
Bad	175	125	300
Predicted (Total)	781	219	1000

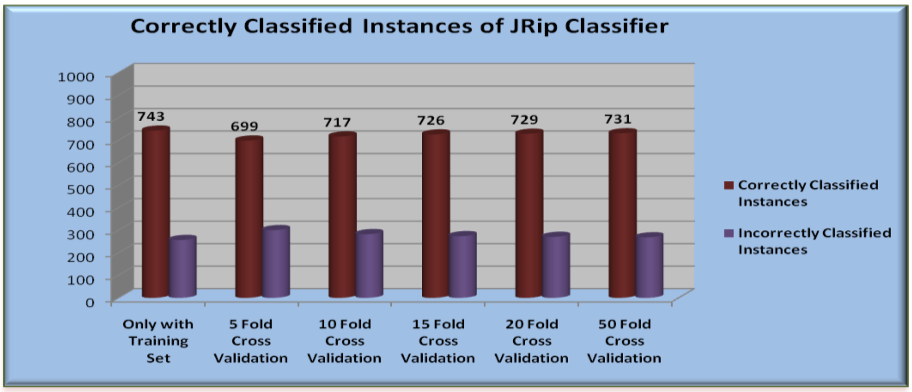

Fig. 1 Correctly Classified instances of JRip Classifier

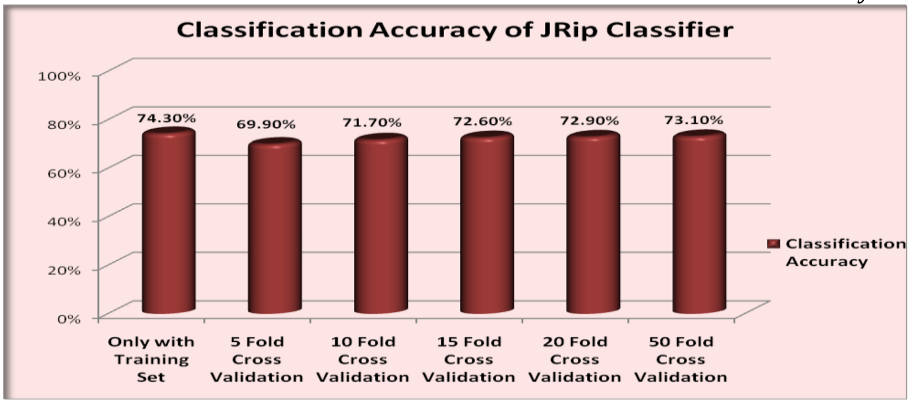

Fig. 2 Classification Accuracy of JRip Classifier

TABLE VIII
JRIP CLASSIFIER OVERALL EVALUATION SUMMARY

Test Mode	Total Test Instances	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy	Kappa Statistic	Mean absolute error	Root Mean Squared Error	Time Taken to Build Model (Sec)
33% Percentage Split	670	461	209	68.806%	0.1847	0.3958	0.4473	0.02
66% Percentage Split	340	238	102	70%	0.2068	0.3854	0.4345	1.91
75% Percentage Split	250	191	59	76.4%	0.338	0.352	0.4144	0.58
80% Percentage Split	200	147	53	73.5%	0.2596	0.371	0.4268	0.42

TABLE IX
CONFUSION MATRIX – JRIP CLASSIFIER (33% PERCENTAGE SPLIT)

	Good	Bad	Actual (Total)
Good	393	97	490
Bad	112	68	180
Predicted (Total)	505	165	670

TABLE X
CONFUSION MATRIX – JRIP CLASSIFIER (66% PERCENTAGE SPLIT)

	Good	Bad	Actual (Total)
Good	203	47	250
Bad	55	35	90
Predicted (Total)	258	82	340

TABLE XI
CONFUSION MATRIX – JRIP CLASSIFIER (75% PERCENTAGE SPLIT)

	Good	Bad	Actual (Total)
Good	163	21	184
Bad	38	28	66
Predicted (Total)	201	49	250

TABLE XII

CONFUSION MATRIX – JRIP CLASSIFIER (80% PERCENTAGE SPLIT)

	Good	Bad	Actual (Total)
Good	127	22	149
Bad	31	20	51
Predicted (Total)	158	42	200

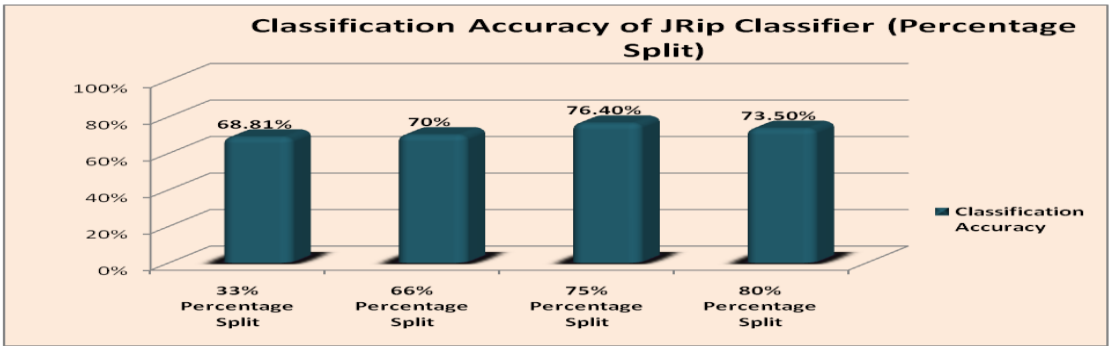

Fig. 3 Classification Accuracy of JRip Classifier (Percentage Split)

B. Performance of PART Classifier

The overall evaluation summary of PART Classifier using training set and different cross validation methods is given in Table XIII. The performance of PART Classifier in terms of Correctly Classified Instances and Classification Accuracy is shown in Fig. 4, Fig. 5 and Fig. 6.The classification summary of PART Classifier for different percentage split is given in Table XX. The confusion matrix for each different test mode is given in Table XIV to Table XIX and from XXI to XXIV. PART Classifier gives 78.6% for the training data set. Various cross validation and percentage split methods are used to test its actual performance. On an average, it gives around 70% of classification accuracy for credit risk estimation.

TABLEXIIII

PART CLASSIFIER OVERALL EVALUATION SUMMARY

Test Mode	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy	Kappa Statistic	Mean absolute error	Root Mean Squared Error	Time Taken to Build Model (Sec)
Training Set	897	103	89.7%	0.7526	0.1605	0.2833	3.48
5 Fold CV	688	312	68.8%	0.2514	0.3348	0.5101	1.84
10 Fold CV	702	298	70.2%	0.2767	0.3245	0.4974	0.72
15 Fold CV	726	274	72.6%	0.3297	0.304	0.4828	1.2
20 Fold CV	696	304	69.6%	0.2564	0.3253	0.499	0.69
50 Fold CV	706	294	70.6%	0.2905	0.3164	0.4886	1.11

TABLE XIV

CONFUSION MATRIX – PART CLASSIFIER (ON TRAINING DATASET)

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

TABLE XV

CONFUSION MATRIX – PART CLASSIFIER (5 FOLD CROSS VALIDATION)

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

TABLE XVI
CONFUSION MATRIX – PART CLASSIFIER (10 FOLD CROSS VALIDATION)

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

TABLE XVII
CONFUSION MATRIX – PART CLASSIFIER (15 FOLD CROSS VALIDATION)

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

TABLE XVIII
CONFUSION MATRIX – PART CLASSIFIER (20 FOLD CROSS VALIDATION)

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

TABLE XIX
CONFUSION MATRIX – PART CLASSIFIER (50 FOLD CROSS VALIDATION)

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

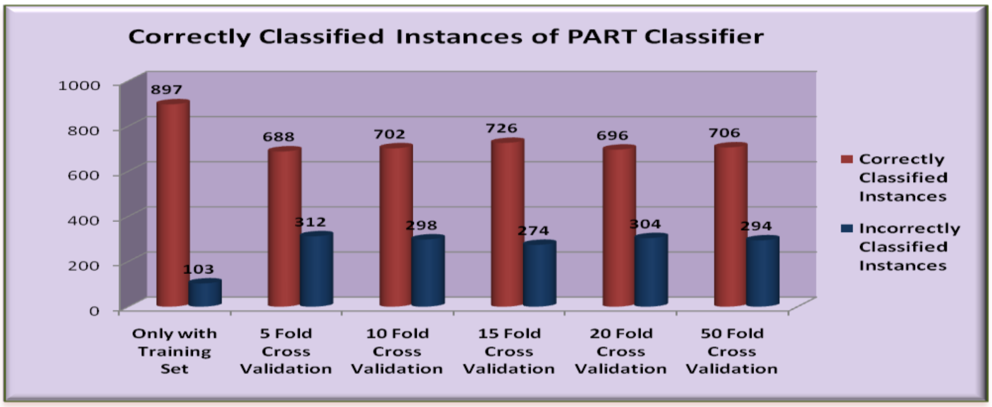

Fig. 4 Correctly Classified instances of PART Classifier

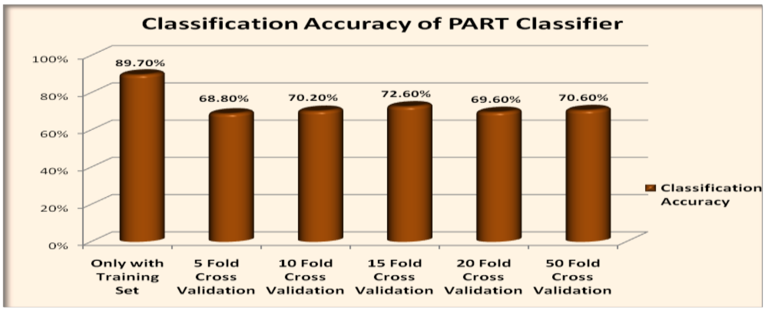

Fig. 5 Classification Accuracy of PART Classifier

TABLE XX

PART CLASSIFIER OVERALL EVALUATION SUMMARY

Test Mode	Total Test Instances	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy	Kappa Statistic	Mean absolute error	Root Mean Squared Error	Time Taken to Build Model (Sec)
33% Percentage Split	670	456	214	68.0597 %	0.2201	0.3486	0.5015	0.34
66% Percentage Split	340	240	100	70.5882 %	0.2653	0.3099	0.4939	1.88
75% Percentage Split	250	184	66	73.6%	0.3272	0.2944	0.4675	2.53
80% Percentage Split	200	150	50	75%	0.342	0.2827	0.4684	3.88

TABLE XXI

CONFUSION MATRIX – PART CLASSIFIER (33% PERCENTAGE SPLIT)

	Good	**Bad**	**Actual (Total)**
Good	371	119	490
Bad	95	85	180
Predicted (Total)	466	204	670

TABLE XXII

CONFUSION MATRIX – PART CLASSIFIER (66% PERCENTAGE SPLIT)

	Good	**Bad**	**Actual (Total)**
Good	196	54	250
Bad	46	44	90
Predicted (Total)	242	98	340

TABLE XXIII

CONFUSION MATRIX – PART CLASSIFIER (75% PERCENTAGE SPLIT)

	Good	**Bad**	**Actual (Total)**
Good	150	34	184
Bad	32	34	66
Predicted (Total)	182	68	250

TABLE XXIV

CONFUSION MATRIX – PART CLASSIFIER (80% PERCENTAGE SPLIT)

	Good	**Bad**	**Actual (Total)**
Good	124	25	149
Bad	25	26	51
Predicted (Total)	149	51	200

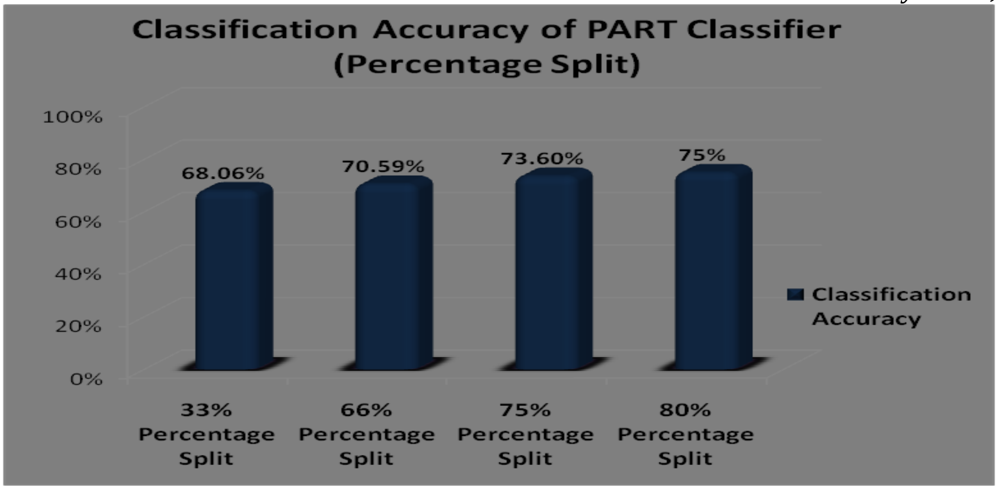

Fig. 6 Classification Accuracy of PART Classifier (Percentage Split)

C. Comparison of JRip and PART Classifiers

The comparison between JRip classifier and PART classifier is shown in Fig 7, Fig. 8 and Fig. 9 in terms of Correctly Classified Instances and Classification Accuracy. The overall ranking is done based on classification accuracy, correctly classified instances, MAE and RMSE values and other statistics found using Training Set results, Percentage Split and Cross Validation Techniques. Based on that, it is observed that JRip classifier performs better than PART Classifier.

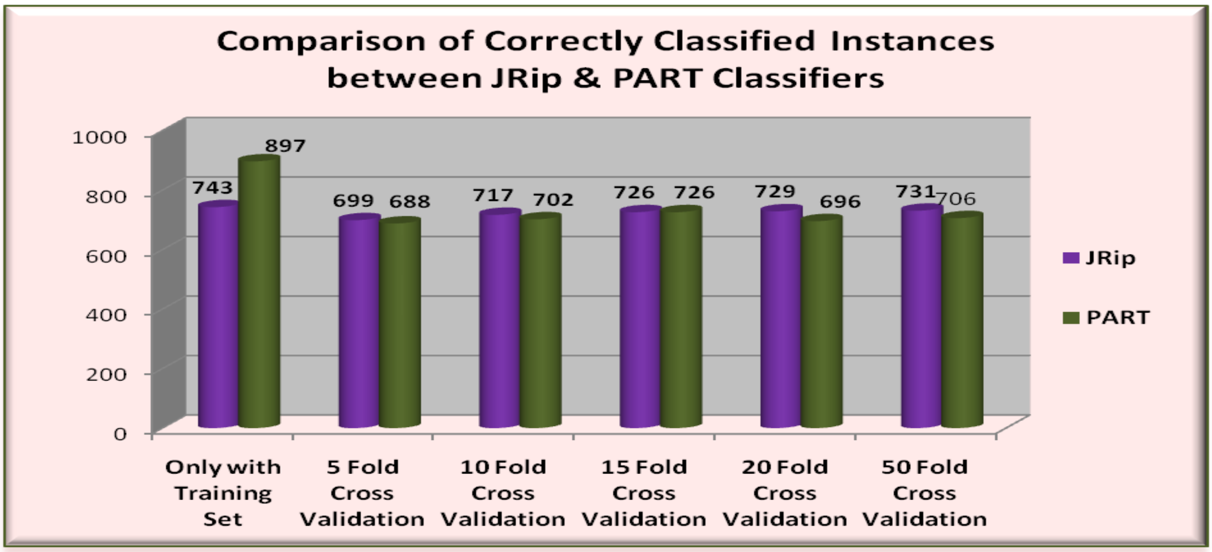

Fig. 7 Correctly Classified Instances Comparison between JRip and PART Classifier

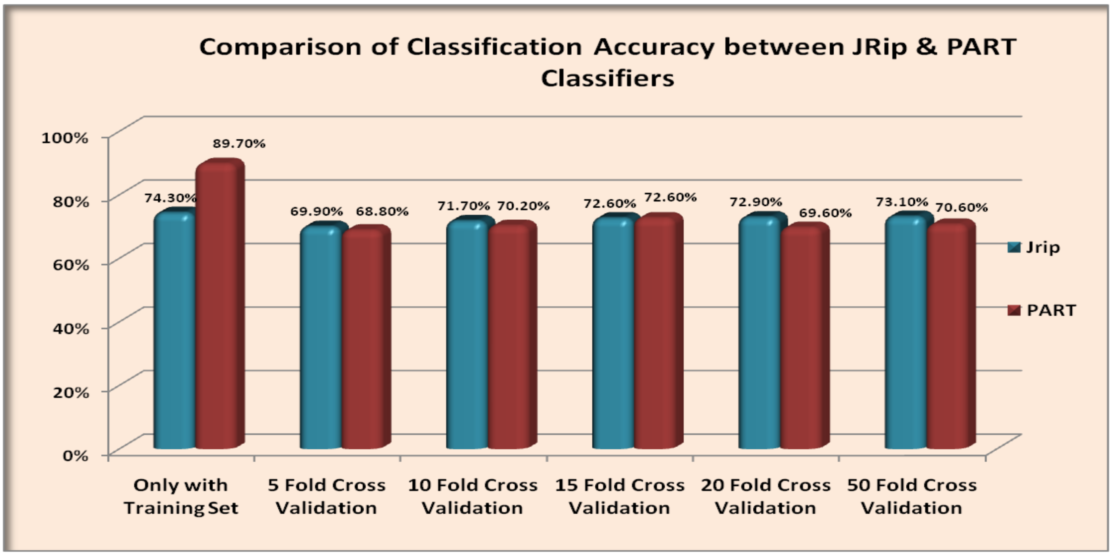

Fig. 8 Correctly Classified Instances Comparison between JRip and PART Classifier

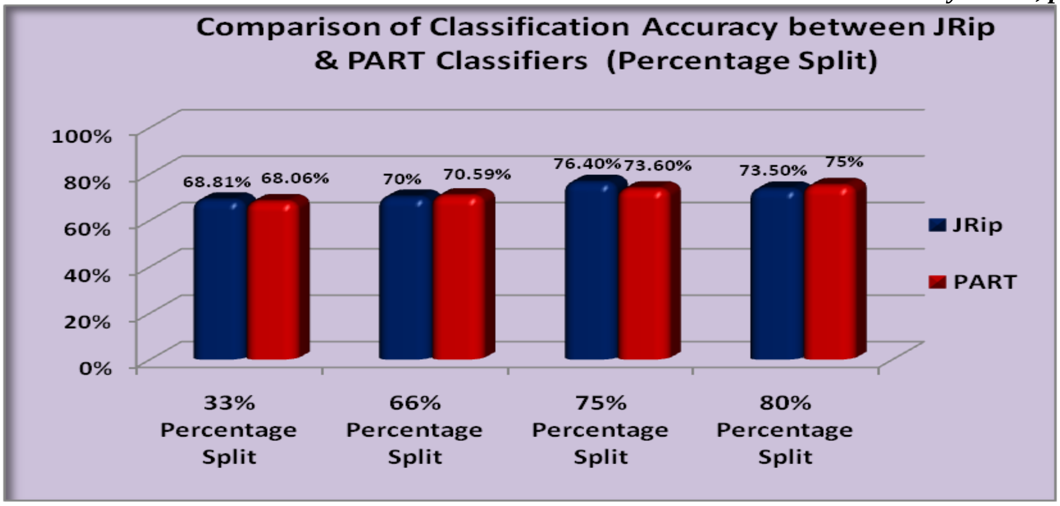

Fig. 9 Correctly Classified Instances Comparison between JRip and PART Classifier for Percentage Split

VII. CONCLUSIONS

This work investigated the efficiency of two different classifiers namely, JRip Classifier and PART Classifier for credit risk estimation. Experiment is done using the open source machine learning tool. Competency comparison of both the classifiers has been done by considering different measures of performance evaluation. After experiment, it is observed that JRip Classifier performs better than PART Classifier for credit risk estimation.

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

