Volume: 5 Issue: XII Month of publication: December 2017

DOI:

Performance Analysis of various rule based Classification Techniques

Alok Kumar[1], Deepak Sinwar[2]
[1]Research Scholar,[2]Assistant Professor
[1,2]Department of Computer Science & Engineering, BRCM College of Engineering & Technology, Bahal

Abstract: Classification of data is one of the biggest challenge now a day especially when the data set contains some hidden facts. The invention of such hidden patterns from the large amount of data can be done in numerous ways, but the classification is the good amongst many other due to its nature of prediction. Classification algorithms are of various natures, this paper deals with the rule based classification techniques. This paper has analyzed the performance of five rule based approaches viz. Decision Table, FURIA, JRip, NNge and OLM on three real datasets. Theoretical analysis and experimental results show that the performance of NNge approach is best (in terms of classification, kappa statistic, mean absolute error etc.) amongst all other rule based classification approaches

Keywords: Classification, dataset, Decision Table, FURIA, JRip, NNge, OLM

I. INTRODUCTION

Data Mining not only provides the detailed knowledge about the data but also now a day it can transform that knowledge into an understandable structure for further use. Aside from the unrefined examination step, it involves some data processing model and some inference considerations. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indexes. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system.

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Various data mining tasks to find different types of patterns include:

A. Characterization

It is a summarization of the general characteristics or features of a user-specified target class of data. The output of data characterization can be presented in various forms such as pie charts, bar charts, multidimensional tables and data cubes.

B. Discrimination

Data discrimination is a comparison of the general features of a user specified target class data objects with the general features of objects from one or a set of (user-specified) contrasting classes.

C. Association Analysis

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis

D. Classification and Regression

Classification is the process of finding a set of models that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data. While classification predicts a categorical value, regression is applied if the field being predicted comes from a real-valued domain or we can say that regression analysis is a statistical methodology that is most often used for numeric prediction. Common applications of classification include credit card fraud detection, insurance risk analysis, bank loan approval, etc.

E. Cluster Analysis

Objects in a database are clustered or grouped based on the principle of maximizing intra class similarity and minimizing interclass similarity. Unlike classification which has predefined labels, clustering must in essence automatically come up with the labels. Applications of clustering include demographic or market segmentation for identifying common traits of groups of people, discovering new types of stars in datasets of stellar objects, and so on.

The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm, which maps input data to a category. Some examples of traditional classifiers are: Linear classifiers (Fisher's linear discriminate, Logistic regression, Naive Bayes classifier, Perceptron), Support vector machines, Quadratic classifiers, k-nearest neighbour, Decision trees , Random forests, Neural networks etc. But in this paper we have diverted our focus on some rule based classifications which are listed in [10] as Decision Table, FURIA, JRip, NNge and OLM.

The rest of the paper is organized as follows. Section II will reviews some work related with classification algorithms. Section III will discuss about various approaches included in this paper. Experimental results are discussed in Section IV. Finally Section V concludes the study with summary and future work.

II. BACKGROUND

Most of the work in the area of classification is focused on prediction and analysis of data. Generally the problem of classifying the data is notrivial [8]. Regression based and Graphical based methods are most preferred in general, because human eyes can interpret them easily. Jingke Xi [9] classified outlier mining approaches in two classes: Classic Outlier approach and Spatial Outlier approach. The classic outlier approach deals with transactional data while spatial approach deals with spatial data. Same kind of classification has been given by J. Han and M. Kamber [8], they divided the computer based methods for outlier detection in four approaches: the statistical distribution based approach, the distance-based approach, the density-based approach, and the deviation-based approach.

As we know that the clustering is one of the famous techniques towards outliers' discovery. Jiang [16] generalize local outlier factor of object and propose a framework of clustering based outlier detection, which was effective enough. Another approach of this kind was also developed by Jiang [17], called a clustering-based outlier detection method (CBOD), which results in good scalability and adapts to large dataset. Zhang [21] also proposed a novel approach to detect outliers based on clustering, which combines probability with hierarchical agglomerative clustering.

Same kind of approach based on distance to k-neighbours has been presented by Yu et al. [5]. They proposed two algorithms based on local sparsity and local isolation coefficient. They showed in their experiments that we can achieve better outlier mining results if their algorithms are utilized instead of the conventional algorithms. Another outlier mining approach based on weighted attributes from data streams has been proposed by Yogita [22]. Such kind of outlier detection is a very challenging problem, because it is not possible to scan data streams multiple times.

They assign weights to attributes depending upon their respective relevance. Sometimes weighted attributes are helpful in reducing or removing the effect of noisy attributes in mining tasks. Yousri et al. [13] proposed fuzzy outlier analysis approach which can combine any outlier analysis approach with any clustering approach. They introduced the concept of universal clusters and outlier clusters along with their memberships.

Clustering based outlier detection approaches are common choices, same kind of approaches can be found in [6 & 3]. A Rough Set based approach to analyse outliers in high dimensional space has been proposed by Jin et al [20]. Their key concept behind the analysis is exceptional reduction algorithm (ERDA), which results in better understanding about the data. Comparison of various already developed outlier mining approaches has also been done by some researchers in [1 & 15]. Sometimes it becomes necessary to rank outliers according to their characteristics, such kind of ranking has been provided by Muller et al. [7]. This approach was very efficient in ranking outliers in high dimensional data. Zhou et al. [11] proposed a dissimilarity based approach to detect outliers called OMABD (Outlier Mining Algorithm Base on Dissimilarity). The key concept behind this approach is that they only check the objects in the dissimilarity matrix with the dissimilarity threshold. We can also categorize this concept in the class of clustering based outlier mining. There are numerous approaches of outlier mining; more detailed view about weighted frequent patterns based outlier mining, spatial outlier detection, entropy based, graph based and neural network based approaches can be found in [18, 14, 2, 4 & 19 respectively.

III. EXPERIMENTAL WORK

Performance of some rule based classification algorithms has been performed on three real world data sets. A comparative study between five rule based algorithms [10] viz. Decision Table, FURIA, JRip, NNge and OLM has been conducted. These experiments have been conducted on two real life data sets obtained from UCI Machine Learning Repository. All experiments were performed on Intel(R) Core(TM)2 Duo E7500 (each with 2.3 GHz clock) with 2 GB of main memory running on windows XP(32 bit) Service Pack 2. All algorithms were run on WEKA [10] version 3.7.9.

The characteristics of these databases are listed in Table 1 as follows:

TABLE 1: DATASETS USED

Relation Name	Number of Instances	Number of Attributes	Type of Data
Credit Approval	690	16	Nominal, Numeric
Eucalyptus Soil Conservation	736	20	Nominal, Numeric
German Credit	1000	21	Nomial, Numeric

Table 2: Performance analysis of different classification algorithms on Credit Approval dataset

	Decision Tabe	FURIA	JRip	NNge	OLM
Kappa Statistic	0.726	0.7881	0.7637	1	0.8751
Mean Absolute Error	0.2158	0.105	0.2012	0	0.0623
Root Mean Squared Error	0.327	0.2978	0.3172	0	0.2496
Classfication %	86.2319	89.5652	88.4058	100	93.7681

Description of the experimental results:As shown in table 2 the DecisionTable technique has correctly classified only 86.23 % of the total data (Credit Approval dataset), rest of the data has been declared as outliers by the algorithm. The kappa statistic measure of that algorithm on the same data set is 0.72. In this same manner we can see the results of rest four algorithms on the same dataset. It can be easily seen from Table 2 that the performance of NNge algorithm is best amongst all algorithms. It has the 100 % correct classification rate, which means the classification done by this algorithm is very accurate. If any algorithm is able to classify the whole data then divinity there will be no error of any type, hence the mean absolute error, root mean squared error are zero for this experiment. The performance of NNge algorithm on other two datasets can be seen from the Tables 3 and 4 respectively. Surprisingly the results are as expected. The algorithm has 100 % classification rate on second and third datasets. On the other hand the performance of OLM algorithm is also satisfactory, it was able to classify 94%, 87% and 82 % of the data of first, second and third dataset respectively.

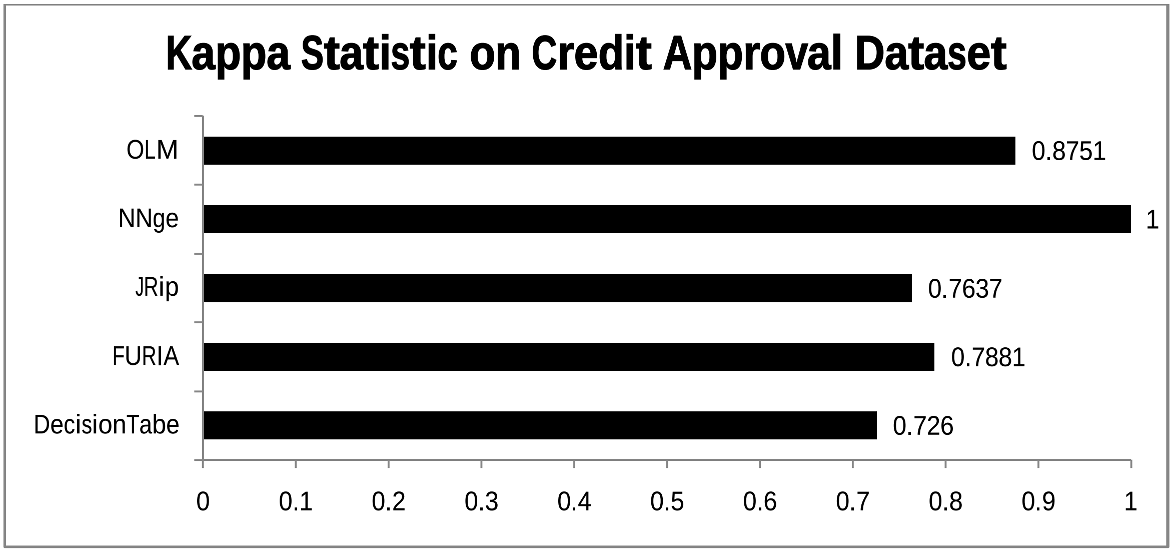
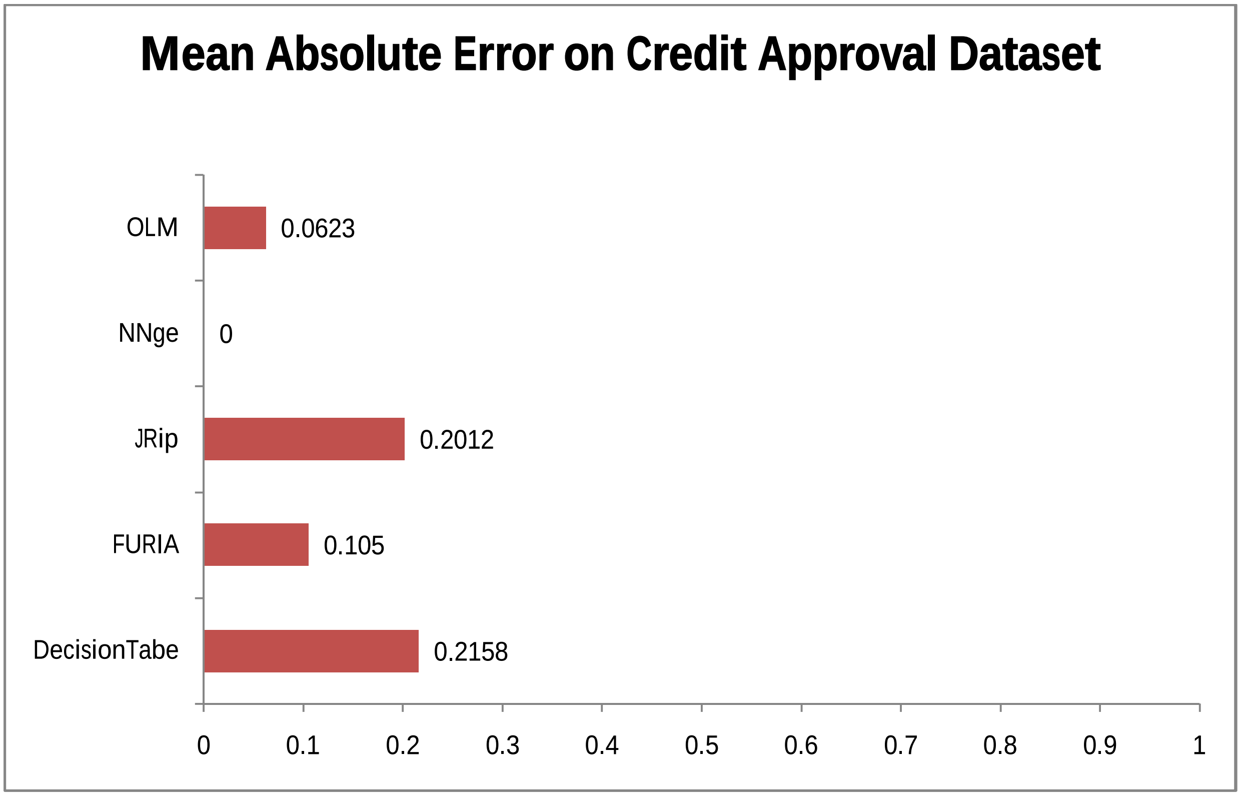


Figure 1: Kappa Statistic measure on first dataset (Credit Approval)

The performance analysis of these five algorithms can be easily observed from the figures 1-9. Figure 1 shows the kappa statistic measure of five rule based classification algorithms on first data set i.e. Credit Approval Dataset (obtained from UCI repository). In this figure we can see that NNge algorithm has the highest kappa value whereas DecisionTable algorithm has the least kappa value among all algorithms. On the other hand figure 2 shows Mean Absolute Error (MAE) generated on the first dataset by the five algorithms. As the name suggests the MAE is the error term used to show the error percentage during classification. In figure 2 it can be easily observed that NNge algorithm has no error value whereas DecisionTable has the highest error value among all algorithms for this experiment.



Figrue2: Mean Absolute Error on First Dataset (Credit Approval)

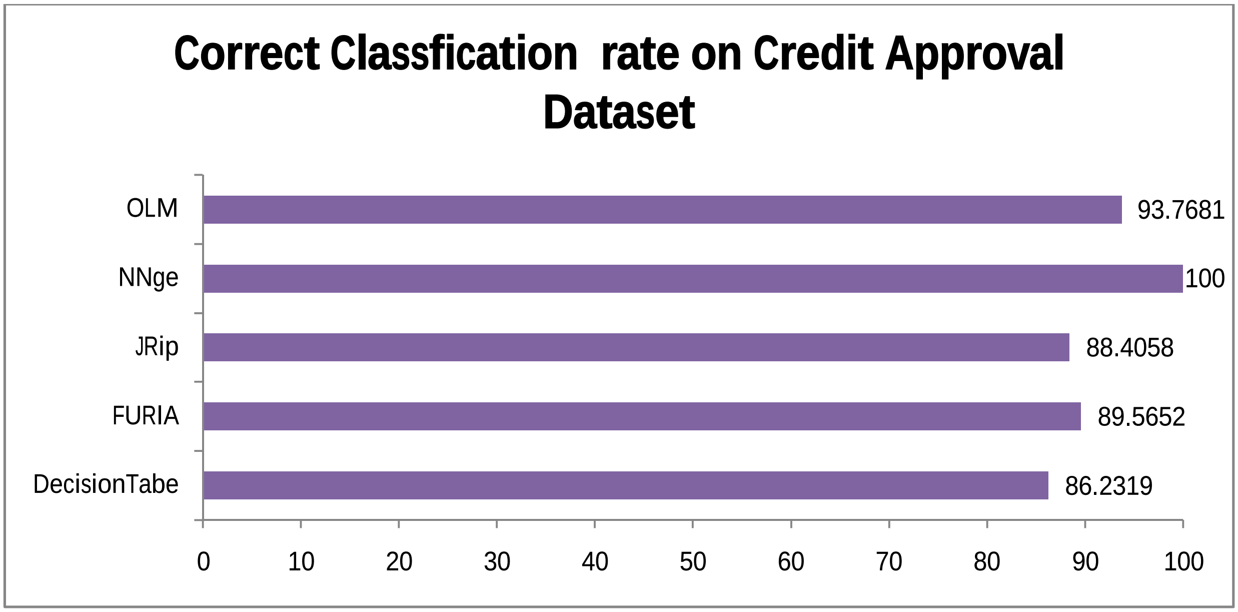


Figure 3: Correct Classification percentage on First Dataset (Credit Approval)

Figure 3 shows the correct classification rate of various rule based classification algorithms on first dataset. It has been observed from figure 3 that NNge has 100 % classification rate, which means the algorithm was able to classify the whole dataset without any errors, whereas the other algorithms were unable to classify the whole datasets. OLM algorithm has the second highest correct classification rate, whereas Decision Table has the least one for this experiment.

Table 3: Performance analysis of different classification algorithms on Eucalyptus dataset

	Decision Tabe	FURIA	JRip	NNge	OLM
Kappa Statistic	0.5555	0.6579	0.6636	1	0.0522
Mean Absolute Error	0.2012	0.1139	0.1556	0	0.2284
Root Mean Squared Error	0.3048	0.2893	0.2789	0	16.6568
Classfication %	65.7609	73.2337	74.1848	100	86.9565

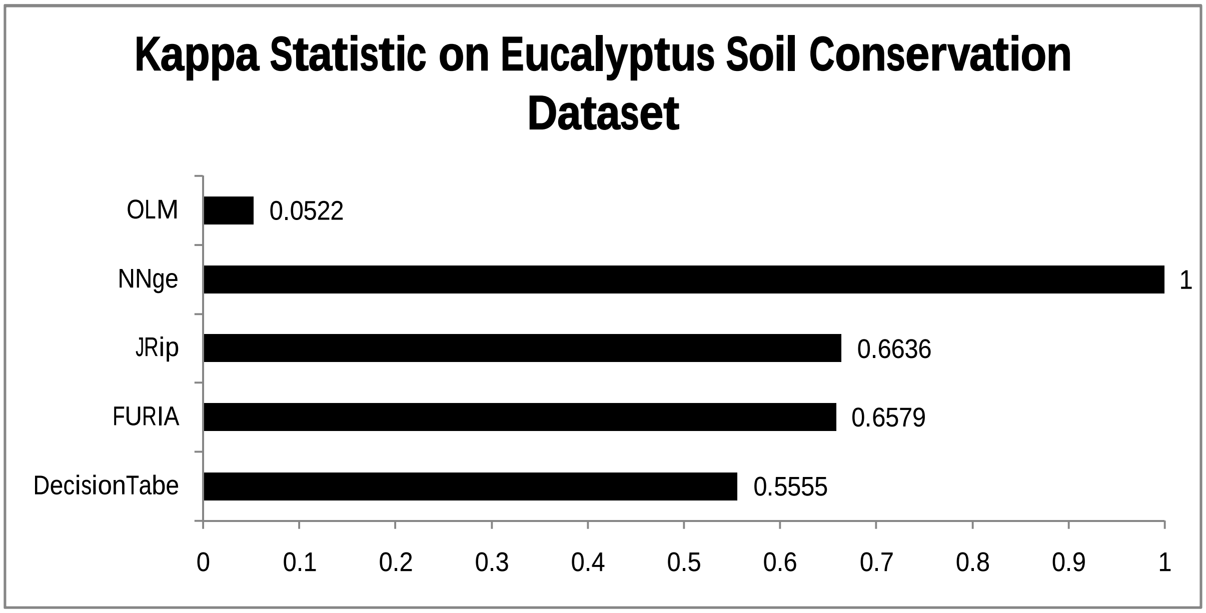


Figure 4: Kappa Statistic measure on first dataset (Eucalyptus Soil Conservation)

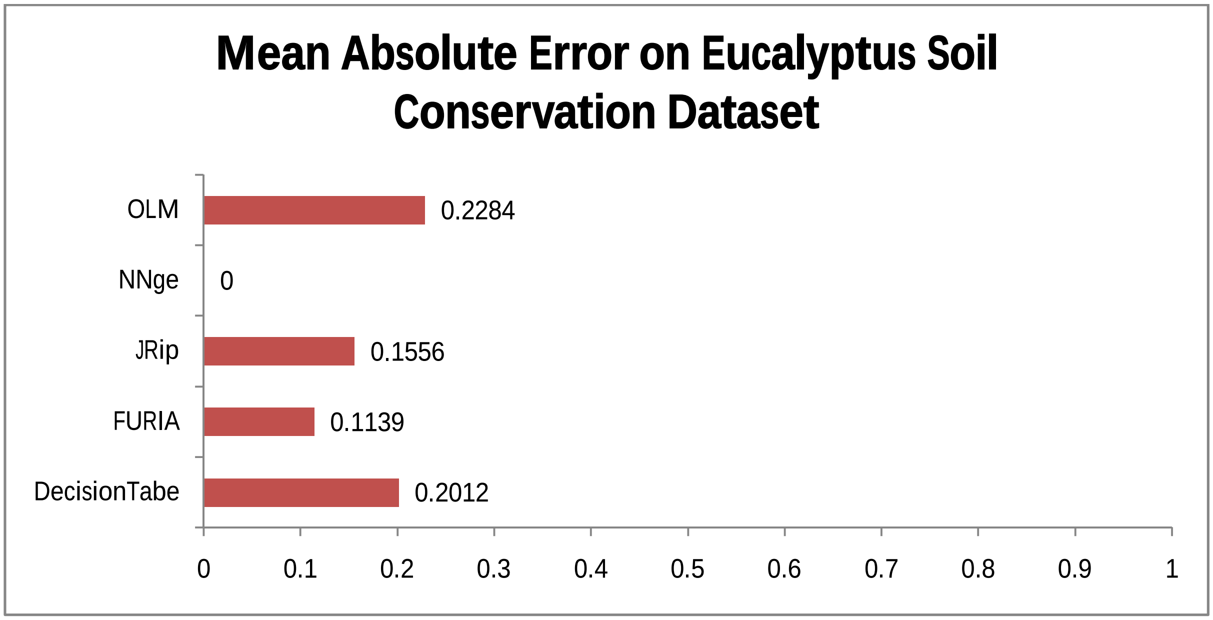


Figure 5: Mean Absolute Error on Eucalyptus Soil Conservation dataset

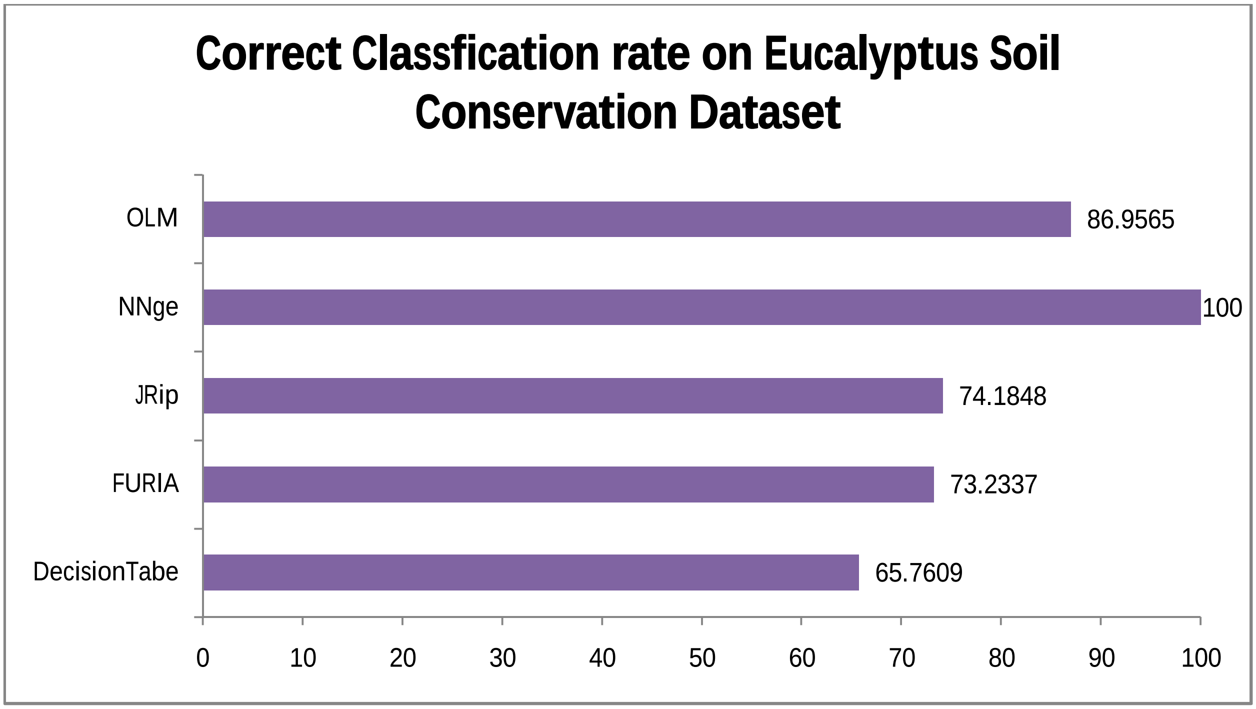


Figure 6: Correct Classfication rate on Eucalyptus Soil Conservation Dataset

Table 4: Performance analysis of different classification algorithms on German Credit dataset

	DecisionTabe	FURIA	JRip	NNge	OLM
Kappa Statistic	0.3622	0.3014	0.3464	1	0.514
Mean Absolute Error	0.3374	0.2495	0.3666	0	0.18
Root Mean Squared Error	0.403	0.4594	0.4281	0	0.4243
Classfication %	76.3	75.8	74.3	100	82

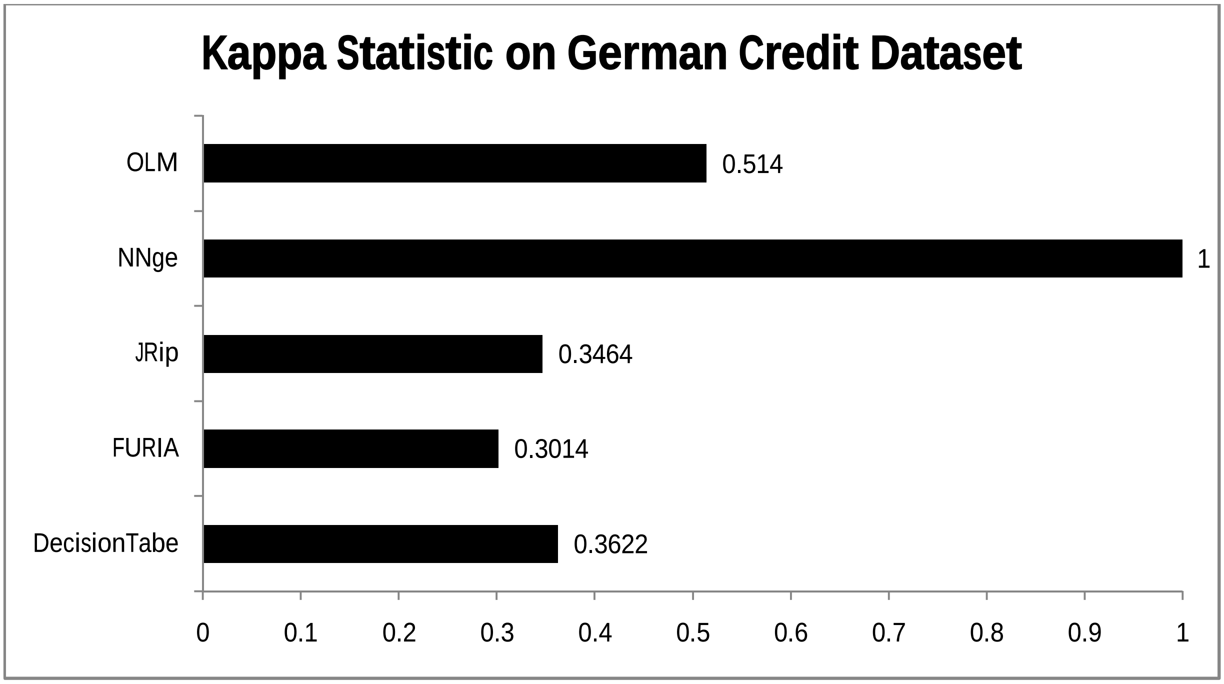


Figure 7: Kappa Statistic on German Credit Dataset

Table 3, 4 and Figures 4-9 shows that NNge algorithm has the highest performance amongst all other algorithms in terms of all the parameters such as kappa statistic, mean absolute error, root mean squared error and correct classification rate.

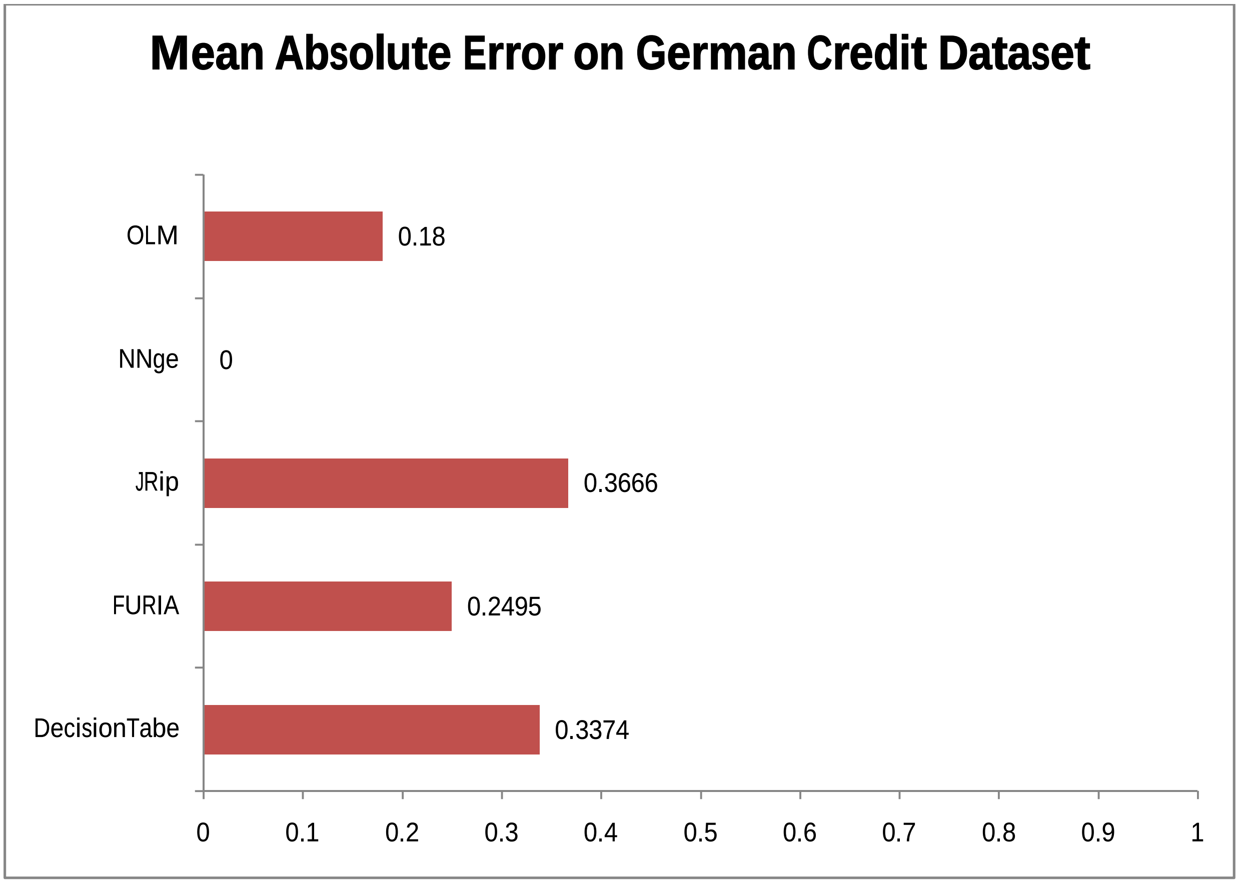


Figure 8: Mean Absolute Error on on German Credit Dataset

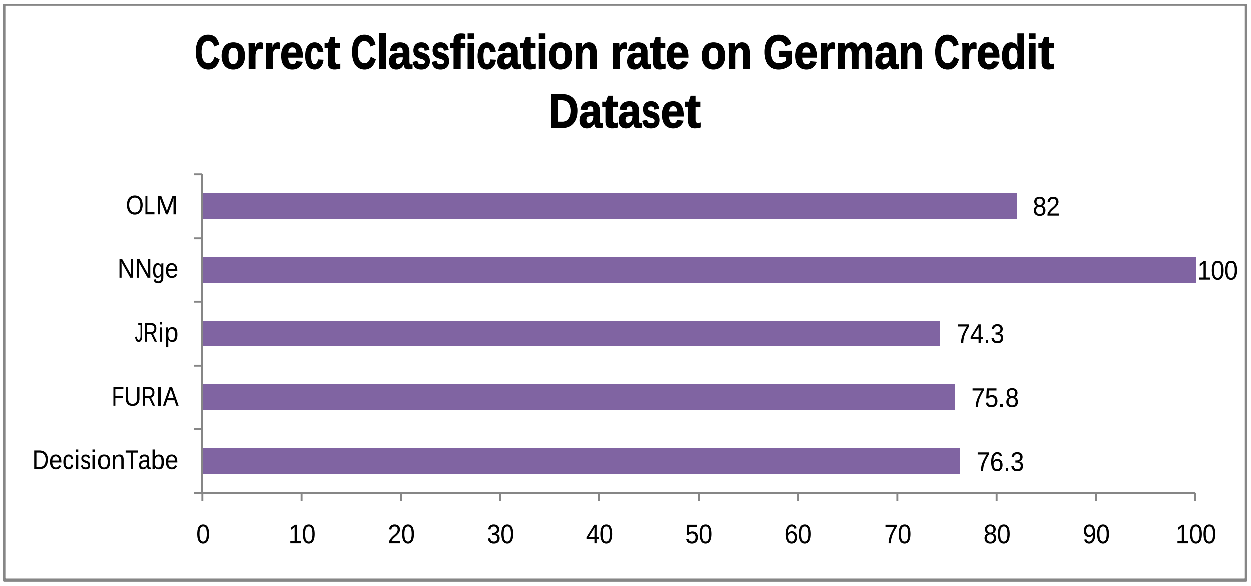


Figure 9: Correct classification rate on German Credit Dataset

IV. CONCLUSION

Currently the use of classification is very much due its nature. This paper has focused some rule based classification approaches. Theoretical analysis and experimental results shown that the NNge algorithm has outperformed other approaches on kappa statistic measure, mean absolute error, root mean squared error and correct classification rate. Experimenting some other algorithms of this kind with some modifications on some large datasets can be the future work of this study.

REFERENCES

[1] A. D. Bella, L. Fortuna, S. Grazianil, G. Napoli and M.G. Xibilia, “A Comparative Analysis of the Influence of Methods for Outliers Detection on the Performance of Data Driven Models”, Instrumentation and Measurement Technology Conference - IMTC 2007, Warsaw, Poland, pp. 1-5.

[2] A. Daneshpazhouh and A. Sami, “Entropy-based outlier detection using semi-supervised approach with few positive examples”, Pattern Recognition Letters (Elsevier), 49, 2014, pp. 77–84.

[3] A. Fawzy, H. M. O. Mokhtar and O. Hegazy, “Outliers detection and classification in wireless sensor networks”, Egyptian Informatics Journal (Elsevier), 14, 2013, 157–164

[4] A. Rahmani, S. Afra, O. Zarour, O. Addam, N. Koochakzadeh, K. Kianmehr, R. Alhajj and J. Rokne, “Graph-based approach for outlier detection in sequential data and its application on stock market and weather data”, Knowledge-Based Systems (Elsevier), 61, 2014, pp. 89–97.

[5] B. Yu, M. Song and L. Wang, “Local Isolation Coefficient-Based Outlier Mining Algorithm”, International Conference on Information Technology and Computer Science” IEEE, 2009, pp. 448-51.

[6] C. Cassisi, A. Ferro, R. Giugno, G. Pigola and A. Pulvirenti, “Enhancing density- based clustering: Parameter reduction and outlier detection”, Information Systems (Elsevier), 38, 2013, pp. 317–30.

[7] E. Muller, I. Assent, U. Steinhausen and T. Seidl, “OutRank: ranking outliers in high dimensional data”, ICDE Workshop, IEEE 2008, pp. 600-603.

[8] J. Han and M. Kamber, “Data Mining Concepts and Techniques”, Second Edition, Morgan Kaufmann Publishers – An Imprint of Elsevier, ISBN: 978-81-312-0535-8.

[9] J. Xi, “Outlier Detection Algorithms in Data Mining”, Second International Symposium on Intelligent Information Technology Application”, IEEE, 2008, pp. 94-97.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, “The WEKA Data Mining Software: An Update”, SIGKDD Explorations, Volume 11, Issue 1.

[11] M. J. Zhou and X. J. Chen, “An Outlier Mining Algorithm Based on Dissimilarity”, International Conference on Environmental Science and Engineering (ICESE 2011)”, Procedia Environmental Sciences (Elsevier), 12, 2012, pp. 810-14.

[12] M. T. Hagan, H. B. Demuth and M. Beale, “ Neural Network Design”, PWS Publishing Company- a division of Thomson Learning, United States of America, ISBN: 7-111-10841-8

[13] N. A. Yousri, M. A. Ismail and M. S. Kamel, “Fuzzy Outlier Analysis A Combined Clustering - Outlier Detection Approach”, IEEE, 2007, pp. 412-18.

[14] Q. Cai, H. He, and H. Mana, “Spatial outlier detection based on iterative self-organizing learning model”, Neurocomputing (Elsevier), 117, 2013, pp. 161–72.

[15] S. H. Wu, D. Drmanacand Li-C. Wang, “A Study of Outlier Analysis Techniques for Delay Testing”, INTERNATIONAL TEST CONFERENCE”, IEEE, 2008, pp. 1-10.

[16] S. Jiang and A. Yang, “Framework of Clustering-Based Outlier Detection”, Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009, pp. 475-79.

[17] S. Jiang and Q. An, “Clustering-Based Outlier Detection Method”, Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008, pp. 429-33.

[18] U. Yun, H. Shin, K. H. Ryu and E. Yoon, “An efficient mining algorithm for maximal weighted frequent patterns in transactional databases”, Knowledge-Based Systems (Elsevier), 33, 2012, pp. 53–64.

[19] X. Zhang and Y. Zhang, “Outlier detection based on the neural network for tensor estimation”, Biomedical Signal Processing and Control, 13, 2014, pp. 148–156.

[20] Y. Jin, Q. Zhu and Y. Xing, “An Exceptional Reduction Algorithm for Outliers Analyzing in High-Dimension Space”, 6th World Congress on Intelligent Control and Automation, 2006, Dalian, China, pp. 5911-14.

[21] Y. Zhang, J. Liu and H. Li, “An Outlier Detection Algorithm based on Clustering Analysis”, First International Conference on Pervasive Computing, Signal Processing and Applications, 2010, pp. 1126-28.

[22] Yogita and D. Toshniwal, “A Framework for Outlier Detection in Evolving Data Streams by Weighting Attributes in Clustering”, 2nd International Conference on Communication, Computing & Security (ICCCS-2012), Procedia Technology (Elsevier), 6, 2012, pp. 214-22.

[23] D. Sinwar and V.S. Dhaka, “Outlier Detection from Multidimensional Space using Multilayer Perceptron, RBFNetwork and Pattern Clustering Techniques”, IEEE Intl. conf ICACEA-2015