Comparative Analysis Of Classification Algorithm Using Machine Learning Technique

[1]N.G.Sree Devi,[2]M.Jeyanthi
[1] Department of computer science, Aditanar College of Arts and Science, Tiruchendur, affiliated to Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India.
[2] Department of computer science, Aditanar College of Arts and Science, Tiruchendur, affiliated to Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India.
[1] devinair.93@gmail.com,[2] jeyanthieral88@gmail.com

Abstract - Data Mining is becoming one of the leading techniques applicable in a variety of area. Classification is a supervised learning approach of data mining and it is used to classify huge data. The dataset have been chosen from UCI Machine Learning Repository. Weka is a powerful machine learning tool that contains many built in algorithms to extract knowledge. In this paper we analyse the classification algorithms such as Naïve Bayes , K-Star, Random Forest, One-R. The multiple parameters taken into considerations for analytical purpose are Accuracy, True Positive Rate, False Positive Rate, Precision, Recall, F-Measure. Finally we observed the performance of the classification algorithm for the UCI Machine Learning data set.

Index Terms- Naïve Bayes, K-Star, Random Forest, One-R, Data Mining

I INTRODUCTION

Data mining is a technology that offers extracting or discovering new relations, hidden knowledge and important patterns from such data. It is also known as Knowledge Discovery in Databases (KDD). Data mining technique is important for analysis purpose. Data mining supports different techniques such as classification, clustering, association rule mining, outlier analysis etc. Data Mining (DM) discovers hidden relationships in data, in fact it is a part of wider process called “knowledge discovery”. Knowledge discovery describes the phases which must be done to ensure reaching meaningful results through research. The objective of DM process is to obtain information out of a dataset and converts it into a comprehensible outline. An understanding of algorithms is combined with detailed knowledge of the dataset An understanding of algorithms is combined with detailed knowledge of the datasets. Data mining must afford very complex and different situations to reach quality solutions. Therefore, data mining is a research field where many advances are being done to accommodate and solves emerging problems [1]. For present study purpose classification technique is investigated.

The organisation of the paper is as followed as : Section II shortly describes the Literature Review , Section III describes the UCI data set description, Section IV describes the classification Algorithms , Section V describes the methodology ,Section VI Results and Discussion, Section VII Concludes the paper.

II LITERATURE REVIEW

Mahesh Parmar (2018)[2] determines two classification algorithms are used for analyzing the datasets. This paper shows the comparative Analysis of decision tree (J48) and Back propagation classification algorithm using the tool of WEKA and finds out which technique is the most suitable for user working on different datasets.

Ismail Saritas et al.,(2017)[3] determines the performances of sixteen different classification methods are evaluated in terms of classification accuracy on Parkinson's Disease dataset. When comparing the performances of algorithms it’s been found that IB1 (96,4103%) has the highest accuracy compared to NaiveBayes (69,2308%).

Driyani Rajeshinigo et al.,(2017)[4] determines the classification algorithms C4.5, Random Forest, Naive Bayes, Multi Layer Perceptron and SVM classifiers are analysed on the students data set. WEKA tool is used to apply the classification algorithms on the selected data set for predicting the student’s semester results. The results are compared and found SVM classifier predicts the results with high accuracy of 81% and C4.5 found to be giving lower accuracy among the algorithms compared.

Sanaa Hassan Abou Elhamayed (2018)[5] determines the performance of the different classifiers is measured with different ratio of the testing and training dataset. Also, the performance of the classifiers is calculated with and without low variance filter. By applying the low variance filter the accuracy of the KNN classifier is enhanced with about 9% while the accuracy of the other classifier is decreased.

Ahmet Toprak et al.,(2017)[6] determines the comparison of different classification techniques on energy efficiency datasets. In this study ten different Data Mining methods namely Bagging, Decorate, Rotation Forest, J48, NNge, K-Star, Naïve Bayes, Dagging, Bayes Net and JRip classification methods are applied on energy efficiency dataset that are taken from UCI Machine Learning Repository. When comparing the performances of algorithms which have been found that Rotation Forest has highest accuracy where as Dagging has the worst accuracy.

Abdullah Caliskan et al.,(2017)[7] presents DNN based classifier is used to classify medical CAD data sets for the purpose of the diagnosis of CAD. The method is tested on the Cleveland, Hungarian, Long Beach and Switzerland data sets from the literature. Experimental results show that the proposed method offers the highest classification accuracy among the methods included in the experiments.

Ramesh Prasad Aharwal.,(2016)[8] determines the comparison of various classification methods using UCI machine learning dataset under WEKA. We have used three measuring factors which names are Accuracy, kappa statistics and mean absolute error for execution by each technique is observed during experiment. This work has been carried out to make a performance evolution of J48, Multilayerperceptron, Naïve Bayes and SMO classifier.

III DATASET DESCRIPTION

For experiments, data sets are taken from Data Mining Repository of University of California Irvine (UCI) [9]. These datasets are given in Table1.

No.	Datasets	Features	Instances	Classes
1	Arsenic Female Lung	5	559	2
2	DNA	181	3186	3
3	Popular kids	11	478	4
4	Zoo	18	101	2
5	Balance scale	5	625	3
6	Anneal	39	989	6

Table 1: Characteristics of Datasets

IV CLASSIFICATION ALGORITHM

A.Navie Bayes

When the dimensionality of the inputs is high, the Naïve Bayes Classifier technique is particularly suited. The problem with the Naïve Bayes Classifier is when it assumes all attributes are independent on each other which in general cannot be applied. Naive bayes is harder to debug and understandable [10]. Naive bayes used in robotics and computer vision. In naive bayes decision tree performs poorly.

Naive Bayes is a probabilistic based classifier which applies Bayes' theorem (or Bayes's rule) with strong independence (naive) assumptions.

$$P(\frac{H}{E})=P(\frac{E}{H})*P(H) \qquad (1)$$

Bayes's rule determines that the outcome of a hypothesis or an event can be predicted based on observations of some evidences. From Bayes's rule, we have

(1) A priori probability of H or P (H): This is the probability that an event occurres before the evidence are observed.

(2) A posterior probability of H or P (H/E): This is the probability that an event occurres after the evidence are observed.

Naive Bayes has an advantage that it requires small training data, while estimating parameters (means and variances of the variable) which are necessary for classification, this is because independent variables are assumed to be the variances of variables as each class needs to be determined and not the complete covariance matrix.

B. K STAR

The K* algorithm can be defined as a method of cluster analysis which mainly aims at the partition of 'n' observation into 'k' clusters in which each observation belongs to the cluster with the nearest mean. We can describe K* algorithm as an instance based learner which uses entropy as a distance measure. The benefits are that it provides a consistent approach to handling of real valued attributes, symbolic attributes and missing values. K* is a simple, instance based classifier, similar to K Nearest Neighbor (K-NN). New data instances, x, are assigned to the class that occurs most frequently amongst the k-nearest data points, y_j , where j = 1, 2…k. Entropic distance is then used to retrieve the most similar instances from the data set. By means of entropic distance as a metric has a number of benefits including handling of real valued attributes and missing values. The K* function can be calculated as:

$$K^* (y_i , x) = -\ln P^* (y_i , x) \qquad (2)$$

Where P* is the probability of all transformational paths from instance x to y. It can be useful to understand this as the probability that x will arrive at y via a random walk in IC feature space. It will performed optimization over the percent blending ratio parameter which is analogous to KNN 'sphere of influence', prior to assessment with other Machine Learning methods[11].

C. ONE R

One R short for "One Rule", is a simple, yet accurate, classification algorithm that generates one rule for each predictor in the data, and then selects the rule with the smallest total error as its "one rule". To create a rule for a predictors, we have to construct a frequency table for each predictor against the target. One R Algorithm for each predictors[12], For each values of that predictor, make rule as follows-

- Count how often each value of target(class)appears
- Find the most frequent class
- Make the rule assign that class to this value of the predictors
- Calculate the total error of the rules of each predictor
- Choose the predictor with the smallest total error.
- Find the best predictor which possess the smallest total error using One R algorithm

D. RANDOM FOREST

The Random Forests algorithm is able to classify large amounts of data with accuracy. Random Forests are an ensemble learning method for classification and regression that construct a number of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. Random Forests are a combination of tree predictors where each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest. The basic principle is that a group of "weak learners" can come together to form a "strong learner". Random Forests are a wonderful tool for making predictions considering they do not overfit because of the law of large numbers. Random Forests grows many classification trees[13]. Each tree is grown as follows:

Algorithm	Sen	Spe	Prec	FMe	Mcc	ROC	PRC	Acc
Naïve Bayes	0.95	0.04	0.95	0.95	0.92	0.99	0.99	0.95
K-Star	0.76	0.10	0.79	0.76	0.63	0.93	0.93	0.76
OneR	0.63	0.19	0.77	0.78	0.55	0.72	0.72	0.63
RandomForest	0.94	0.03	0.94	0.95	0.91	0.99	0.99	0.94

1. If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number mM is specified such that at each node, m variables are selected at random out of the M and the best split on this m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning.

V METHODOLOGY

The Waikato Environment for Knowledge Analysis (Weka) is a machine learning toolkit introduced by Waikato University, New Zealand. It is open source software written in Java. It contains collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes[14]. Advantages of Weka tool:

i. Available freely under the GNU General Public License.

ii. It is portable, as it is implemented in the Java programming language and thus runs on almost any platform.

iii. It is easy to use due to its graphical user interfaces.

VI RESULTS AND DISCUSSION

The experiment is performed using the Machine Learning UCI dataset. In this study we used six datasets. To compare the performance of the classification algorithms using WEKA data mining tool. All the experiments were carried out using a ten-fold cross validation approach.The result of the paper shows which algorithm is more convenient for a particular dataset. Each datasets are classified by four classification algorithms In Table 2 to Table 7 shows the results. Figure 1 to Figure 7 shows the classification performance of the Naïve Bayes, KStar, OneR and Random Forest classifier. Based on the measures Naive Bayes classifier shows the highest performance for all the datasets.

Algorithm	Sen	Spe	Prec	FMe	Mcc	ROC	PRC	Acc
Naïve Bayes	0.99	0.30	0.99	0.99	0.82	0.85	0.98	0.99
K-Star	0.99	0.30	0.99	0.99	0.82	0.85	0.98	0.99
OneR	0.96	0.92	0.94	0.95	0.10	0.52	0.93	0.96
RandomForest	0.96	0.97	0.93	0.94	0.10	0.82	0.97	0.96

Table 2: Comparison of different classifiers for the Arsenic Female Lung data set using 10-fold cross-validation

Table 3: Comparison of different classifiers for the DNA data set using 10-fold cross-validation

Table 4: Comparison of different classifiers for the Popular kids data set using 10-fold cross-validation

Algorithm	sen	Spe	Prec	FMe	Mcc	ROC	PRC	Acc
Naïve Bayes	0.98	0.01	0.98	0.98	0.98	1.00	0.99	0.98
K-Star	0.80	0.15	0.79	0.78	0.68	0.93	0.86	0.80
OneR	0.64	0.28	0.68	0.81	0.60	0.68	0.48	0.64
RandomForest	0.96	0.03	0.96	0.95	0.94	0.99	0.99	0.96

Table 5: Comparison of different classifiers for the Zoo data set using 10-fold cross-validation

Algorithm	sen	Spe	Prec	FMe	Mcc	ROC	PRC	Acc
Naïve Bayes	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
K-Star	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
OneR	0.42	0.39	0.76	0.28	0.12	0.52	0.53	0.43
RandomForest	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 6: Comparison of different classifiers for the Balance scale data set using 10-fold cross-validation

Algorithm	sen	Spe	Prec	FMe	Mcc	ROC	PRC	Acc
Naïve Bayes	0.96	0.06	0.96	0.96	0.89	0.99	0.99	0.96
K-Star	0.78	0.05	0.97	0.86	0.66	0.96	0.95	0.78
OneR	0.83	0.52	0.82	0.90	0.83	0.50	0.72	0.84
RandomForest	0.94	0.16	0.94	0.94	0.84	0.99	0.99	0.94

Table 7: Comparison of different classifiers for the Anneal data set using 10-fold cross-validation

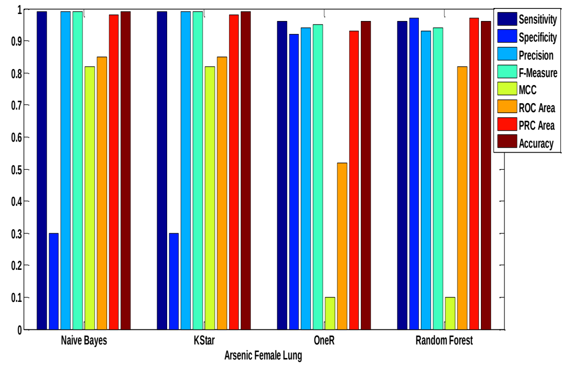

Figure 1: Comparative Analysis for classification Algorithm for UCI Machine Learning DataSet Arsenic Female Lung.

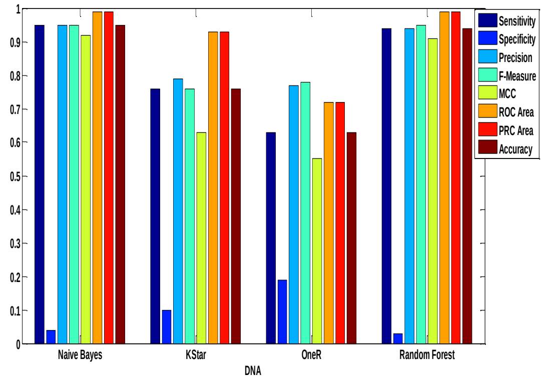

Figure 2: Comparative Analysis for classification Algorithm for UCI Machine Learning DataSet DNA.

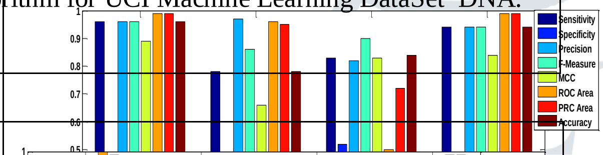

Figure 6: Comparative Analysis for classification Algorithm for UCI Machine Learning DataSet Anneal

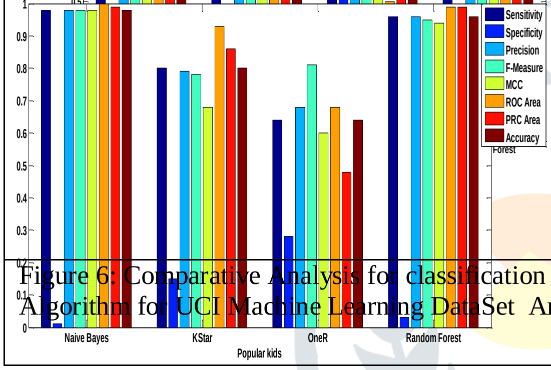

Figure 3: Comparative Analysis for classification Algorithm for UCI Machine Learning DataSet Popular kids.

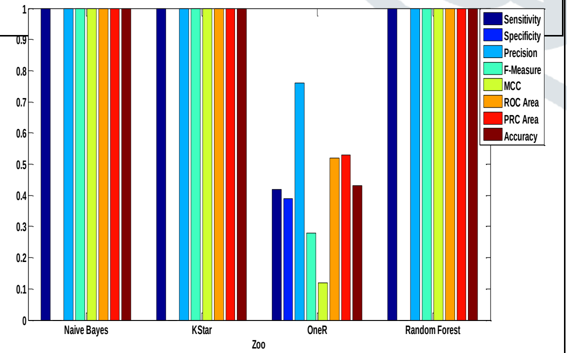

Figure 4: Comparative Analysis for classification Algorithm for UCI Machine Learning DataSet Zoo.

Algorithm	sen	Spe	Prec	FMe	Mcc	ROC	PRC	Acc
Naïve Bayes	0.91	0.08	0.91	0.95	0.91	0.96	0.92	0.91
K-Star	0.88	0.10	0.88	0.91	0.83	0.92	0.90	0.88
OneR	0.56	0.37	0.56	0.59	0.20	0.60	0.49	0.56
RandomForest	0.78	0.19	0.78	0.81	0.65	0.86	0.84	0.78

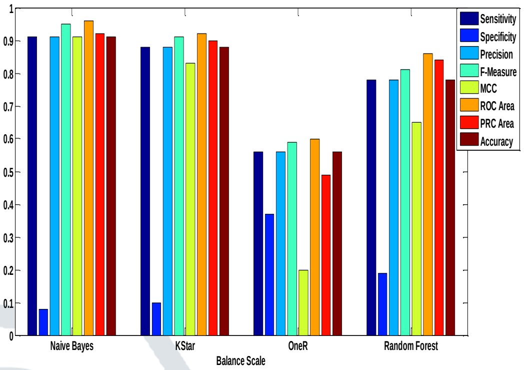

Figure 5: Comparative Analysis for classification Algorithm for UCI Machine Learning DataSet Balance Scale.

VI CONCLUSION

In this paper, we mainly focused on the performance of four classification algorithms such as Naïve Bayes, KStar, OneR and Random Forest using six machine learning UCI Datasets. The dataset efficiency is evaluated by means of classification accuracy, Sensitivity, Specificity, Precision ,F-Measure ,MCC ,ROC Area ,PRC Area using WEKA tool. Ten fold Cross validation testing used for the experiments. Results are shown in the Table 2 to 7. From the results it is evident that Naïve Bayes produces the best classification accuracy which is compared to other classification algorithms. The result of the paper show which algorithm is more convenient for a particular dataset. Finally we propose Naïve bayes classification algorithm for UCI Machine Learning dataset.

ACKNOWLEDGEMENT

The author would like to thank to Data Mining Repository of University of California Irvine (UCI) for providing Machine Learning UCI Dataset.

REFERENCES

[1] Shivangi Gupta, Neeta Verma,”Comparative Analysis of classification Algorithms using WEKA tool**”,** International Journal of Scientific & Engineering Research, Volume 7, Issue 8, August-2016.

[2] Mahesh Parmar,” Comparative Analysis of Classification Techniques using WEKA on Different Datasets**”,** International Journal of Latest Engineering and Management Research (IJLEMR) Volume 03 - Issue 06 ǁ June 2018 ǁ PP. 01-05.

[3] Ismail Saritas, Murat Koklu, Kemal Tutuncu,” Performance of classification Techniques on Parkinson’s Disease”,International Journal of Advances in Science Engineering and Technology, ISSN: 2321-9009, Vol-5, Iss-1, Spl. Issue-2 Feb.-2017.

[4] Driyani Rajeshinigo, J. Patricia Annie Jebamalar ,” Educational Mining: A Comparative Study of Classification Algorithms Using WEKA” International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 3, March 2017.

[5] Sanaa Hassan Abou Elhamayed,” Comparative Study on Different Classification Techniques for Spam Dataset” **,**International Journal of Computer and Communication Engineering,Volume 7, Number 4, October 2018.

[6] Ahmet TOPRAK, Nigmet KOKLU, Aysegul TOPRAK, Recai OZCAN,” Comparison of Classification Techniques on Energy Efficiency Dataset” *International Journal of Intelligent* Systems and Applications in Engineering,june 2017.

[7] Abdullah Caliskan and Mehmet Emin Yuksel,” Classification of coronary artery disease datasets by using a deep neural network”, volume 1 Issue 4 | October 2017.

[8] Ramesh Prasad Aharwal,”Evolution of various classification techniques of Weka using different datasets”, *Vol-2 Issue-2 2016.*

[9] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz, UCI Repository of machine learning databases, University California Irvine, Department of Information and Computer Science,1998.

[10] Meenakshi, Geetika,” Survey on Classification Methods using WEKA”, International Journal of Computer Applications (0975 – 8887) Volume 86 – No 18, January 2014.

[10] Ms.S. Vijayarani , Ms. M. Muthulakshmi,” Comparative Analysis of Bayes and Lazy Classification Algorithms”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 8, August 2013.

[11] Chitra Nasa, Suman “Evaluation of Different Classification Techniques for WEB Data”, International Journal of Computer Applications (0975 – 8887) Volume 52– No.9, August 2012.

[12] Vrushali Bhuyar,” Comparative Analysis of Classification Techniques on Soil Data to Predict FertilityRate for Aurangabad District”, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 2, March – April 2014.

[13] M. Purnachary, B. Srinivasa S P Kumar, Humera Shaziya,” Performance Analysis of Bayes

[14] Classification Algorithms in WEKA Tool using Bank Marketing Dataset”, *International* Journal of Engineering Research in Computer Science and Engineering (IJERCSE)Vol 5, Issue 2, February 2018.