Journal of Basic and Applied Engineering Research
p-ISSN: 2350-0077; e-ISSN: 2350-0255; Volume 4, Issue 3; April-June, 2017, pp. 180-184
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/Publication.html

Comparative Analysis of Various Classification Algorithms

Preeti Nair[1], Indu Kashyap[2] and Suresh Kumar[3]

[1,2,3]Manav Rachna International University, Faridabad
E-mail: [1]preeti.nair84@gmail.com, [2]indu.fet@mriu.edu.in, [3]suresh.fet@mriu.edu.in

Abstract—*The massive amount of data that carries a lot of useful insights which can benefits many domains such as health care, customer relation management, ecommerce and so on. The amount of unwanted data might be more than the useful ones, so the process which is used to refine such insights from a huge amount of data is called as DATA MINING. There are many techniques in data mining like clustering, classification, association etc. Here in this paper we have discussed about the data mining technique called classification. The rapid emergence of large amount of data made it impossible to study the behaviours and properties of the data, therefore a need has come to classify these data into certain groups such that each group contains similar form of data. In order to classify the data we use classification techniques. There are many classification techniques in data mining. Each model has its own level of accuracy. However, there is no model which can be said as too good or too bad, the performance of the model purely depends on the datasets used on the model. The aim of this study is to compare the accuracy of some classification techniques which were applied on certain datasets. The comparative study is done on four classification algorithms namely One R, Naive Bayesian, Decision tree and K- Nearest Neighbour on six datasets. We analysed that in this study of comparison the accuracy of Naive Bayesian is more than the other three algorithms.*

1. INTRODUCTION

Data mining is a task of finding interesting patterns from huge amount of data such that those interesting patterns can be benefited to certain fields such as research area, customer relationship management and many more.

Data mining is a multidisciplinary field which combines statistics, machine learning, and artificial intelligence and database technologies to predict future from large data repositories. The data mining techniques such as association, classification and clustering can be applied on various kinds of data such as database data, transactional data, and data warehouse. The data present in these data repositories hold rich hidden information that can be used for intelligent decision making.

Classification is a data mining technique in which a classification model also known as classifier predicts the class labels. The classification process has two steps; in the first step, a classification algorithm finds correlations between the values of target and the values of the predictors in a given dataset which is called a training data. This step is called the training or the learning step. The correlations are summarized to build a classifier model, which can then be applied to other datasets in which the class labels are unknown. The second step is the testing phase in which a classifier model is tested by comparing the actual target values and the classifier predicted values in a set of test data.

Classification has numerous applications, including fraud detection, target marketing, performance prediction, manufacturing, and medical diagnosis [5]. For building a classifier model there are many methods in classification. Some of the most popular methods such as Decision Tree, Naive Bayesian, K- nearest neighbour and One R are considered for this study.

2. LITERATURE SURVEY

Paul (2004) has applied four algorithms on medical datasets to compare and evaluate their performance. He has considered discriminant analysis (DA), regression models (multiple and logistic) tree-based algorithms (CART), artificial neural networks algorithms. The performance criteria for evaluating the classification models are accuracy, computational time, comprehensibility and ease of use. The paper reveals that there is no single bestclassification technique, but the performance of a classification algorithm will depend on the features of the dataset under consideration.

Christopher Sibona, Jonalan Brickey (2012) has compared eight different classification algorithms four base algorithms and four boosted versions of each algorithm. In both the versions of algorithms, they have compared two performance measures ROC curve and accuracy. They have taken four popular algorithms Bayes, logistic regression, J48 and Nearest Neighbour (NN). In the standard algorithms comparison study they analysed that the algorithms are significantly different from each other. J48 had the best accuracy and the boosted methodshaveimproved the accuracy of logistic regression.

Congcong Li, Jie Wang, Lei Wang, Luanyun Hu and Peng Gong (2014) have compared fifteen classification algorithms on same Landsat Thematic Mapper (TM) dataset with same classification scheme. After the test, they found that when the algorithms are supplied with sufficient training data they performed very well. When there is a lack of training data, it leads to classification accuracy discrepancies.Some of the algorithms able to handle insufficient training samples than others and many algorithms improved the overall accuracy marginally.

Mustafa A (2016) has used data mining technique for predicting instructor performance. He has taken student evaluation questionnaires as a data set to evaluate the performance of the instructors. In his study four algorithms of classification has been considered for comparison using the performance matrix as accuracy, precision, recall and specificity. He found that although all algorithms were performing well, however C5.0 showed the best performance than all the algorithms.

3. CLASSIFICATION MODELS

Decision Tree

A decision tree algorithm aims to recursively split theobservations into mutually exclusive subgroups until there isno further split that makes a difference in terms of statistical or impurity measures. Among the impurity measures thatare used to the homogeneity of instances in a node of the tree, Information Gain, Gain Ratio, and Gini Index are the most well-known ones[4][5].

One R

One R predicts the most frequent class for the feature values. One R is also known as One Rule which indicates that only a single rule is used for the classification, i.e. the feature with the best performance is considered as the best predictor. One R algorithm has given good performance in many real world datasets. The algorithm starts by iterating over every value of every feature. For that value, count the number of samples from each class that have that feature value. Record the most frequent class for the feature value and the error of that prediction. The feature with the lowest error is chosen as the One Rule and then used to classify other instances. [7]

Naive Bayesian

This algorithm is based on the bayes' theorem with the assumption that the predictors are independent to each other. [5]

Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, X = (x1, x2,...,xn), depicting n measurements made on the tuple from n attributes, respectively, A1, A2,..., An. 2. Suppose that there are m classes, C1, C2... Cm. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naive Bayesian classifier predicts that tuple X belongs to the class Ci if and only if P (Ci |X) > P (Cj |X) for 1 ≤ j ≤ m,j 6= i. Thus, we maximize P (Ci |X). The class Ci for which P (Ci |X) is maximized is called the maximum posterior hypothesis. By Bayes' theorem [5]

$$P\ (Ci\ |X) = P\ (X|Ci)\ P\ (Ci)\ P(X)$$

K-Nearest Neighbour

Nearest-neighbour classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional space. In this way, all the training tuples are stored in an n-dimensional pattern space. When given an unknown tuple, a k-nearest-neighbour classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k "nearest neighbours" of the unknown tuple. "Closeness" is defined in terms of a distance metric, such as Euclidean distance. [5]

4. EVALUATING THE CLASSIFIERS

There are certain measures to evaluate the performance of the classification model. Here in this study we are using binary classification problem i.e. class with two variables positive and negative. True positives (TP): These refer to the positive tuples that were correctly labelledas positive by the classifier. True negatives (TN): These are the negative tuples that were correctly labelled as negative by the classifier. False positives (FP): These are the negative tuples that were incorrectly labeled as positive. False negatives (FN): These are the positive tuples that were mislabeled as negative. These terms are summarized in the Confusion Matrix. [5]

Table 1: Confusion Matrix for classification

Predicted Class

	Yes	No	Total
Yes	TP	FN	P
No	FP	TN	N
Total	P'	N'	P+N

Actual class

The evaluation of the classification model is done with the help of confusion matrix. The confusion matrix is a useful tool for analyzing how well your classifier can recognize tuples of different classes. True Positive and True Negative tell us when the classifier is getting things right, while FP and FN tell us when the classifier is getting things wrong (i.e., mislabeling)[5].

The performance of the classification is obtained by calculating the measures such as accuracy, precision, recall and specificity. Accuracy measures the rate of total correct predictions to all predictions. Precision measures the correctness rate of the class predictions done as positive bythe classifier whereas recall measures the rate of positives correctly predicted as positive by the classifier. Likewise, specificity measures the rate of negatives correctly predicted as negative by the classifier [4][5].

$$\text{Accuracy} = \frac{TP+TN}{P+N}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{P}$$

$$\text{Specificity} = \frac{TN}{N}$$

5. EXPERIMENTAL ANALYSIS

In our experiment, we have taken randomly six data sets from the UCI data repository and each datasets had been executed on each classifier and the results are compared to see which classifier is giving the accurate value in most of the cases. A confusion matrix is built for each classifier against each dataset. With the help of the confusion matrix generated by each classifier, we have calculated and recorded each performance measures in the tables.

Table 2: Accuracy values of the algorithms applied on various datasets

Algorithms	**Datase t 1**	**Datase t 2**	**Datase t 3**	**Datase t 4**	**Datase t 5**	**Data set 6**
One R	.8	.715	.719	**.9**	**.815**	.657
Naive Bayesian	**.845**	**.763**	**.895**	.7	.780	.717
Decision Tree	.774	.719	.807	.6	**.815**	.685
K-NN	.806	.702	.825	.6	.813	**.724**

Following graph represents the classification accuracy in percentage form for each algorithm applied on the datasets. All the classifiers are performing well and the accuracy values obtained have very slight differences between them. But when we considered these values for comparison to find the best out of it, we observed that the naive Bayes has obtained highest accuracy values in most of the cases.

When we consider One R, we observed that outof six there are only two datasets in which One R is giving the highest accuracy values even though its values are very nearer to naive Bayes algorithm.Therefore, from this experiment we can say that naive Bayesis performing verywell and the second best is the One R. Likewise if we consider the algorithms Decision tree and K-NN, there is just one case out of six in which highest accuracy is obtained.

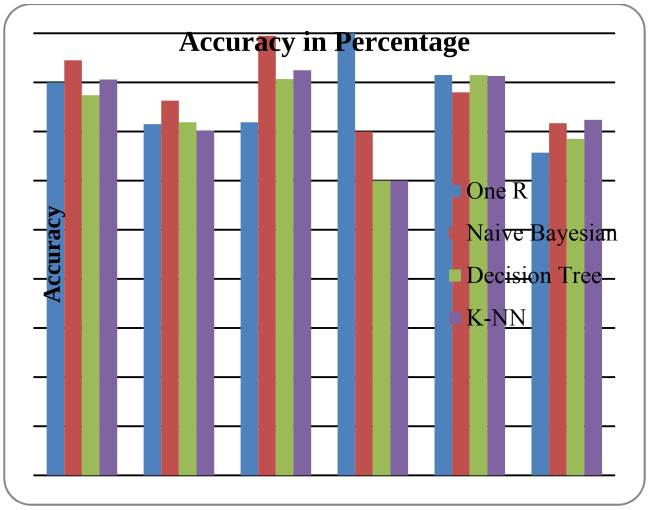

Fig. 1: Shows the accuracy percentage of each of the algorithms applied on various datasets.

Table 3: Precision Recall values of the algorithms applied on various datasets

Algorithms	*Dataset 1*	*Dataset 2*	*Dataset 3*	*Dataset 4*	*Dataset 5*	*Dataset 6*
One R	.529	.740	.643	**.833**	**.880**	.725
Naive Bayesian	**.611**	**.802**	.818	.750	.851	**.778**
Decision Tree	.286	.777	**.846**	.6	**.880**	.766
K-NN	.538	.759	.727	.6	.844	.756

The graph represents the precision values in percentage form. When each algorithm taken under consideration, we observed that, the naive Bayes classifier hasproduced the highest precision values in most of the cases, and the second best are the Decision Tree and One R algorithm.

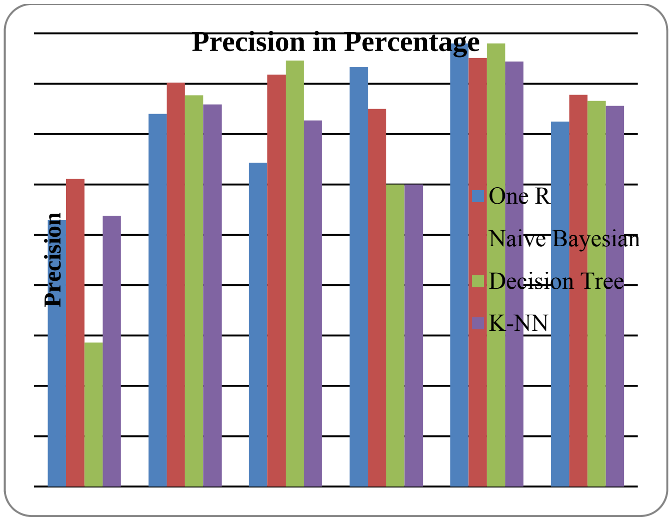

Fig. 2: Shows the precision in percentage of each of the algorithms applied on various datasets.

Table 4: Recall values of the algorithms applied on various datasets

Algorithms	**Dataset 1**	**Dataset 2**	**Dataset 3**	**Dataset 4**	**Dataset 5**	**Dataset 6**
One R	.280	**.866**	.450	**1.00**	.819	.826
Naive Bayesian	**.688**	.844	**.900**	.6	.789	.836
Decision Tree	.063	.796	.550	.6	.819	.796
K-NN	.438	.794	.800	.6	**.862**	**.896**

In following graph, we observed the recall percentage values; we found that Naive Bayes and One R are performing well. K- Nearest Neighbour also scored well and the decision tree algorithm was the worst in terms of recall.

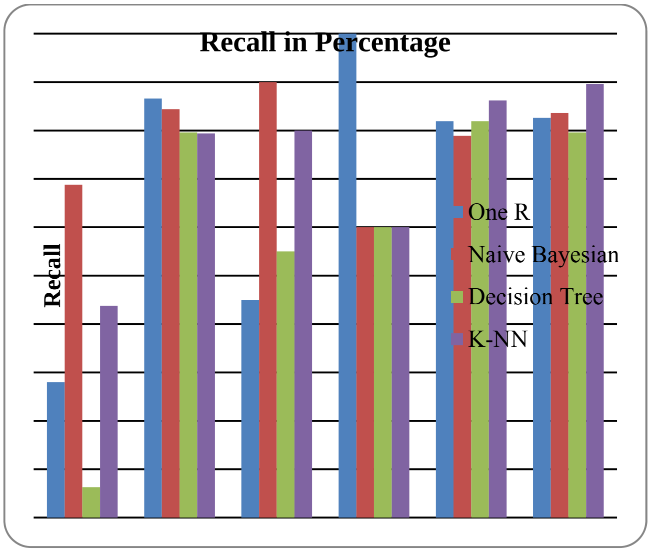

Fig. 3: Recall values in percentage of each of the algorithms applied on various datasets.

Table 5: Specificity Recall values of the algorithms applied on various datasets.

Algorithms	**Dataset 1**	**Dataset 2**	**Dataset 3**	**Dataset 4**	**Dataset 5**	**Dataset 6**
One R	.935	.433	.865	.8	**.809**	.256
Naive Bayesian	.886	**.612**	.892	**.8**	.765	**.435**
Decision Tree	**.959**	.575	**.946**	.6	**.809**	.424
K-NN	.902	.530	.838	.6	.728	.318

In terms of specificity, we observed that naive bayes and Decision tree have scored well and the second best is the One R and the worst specificity values are of K- NN Algorithm in this study.

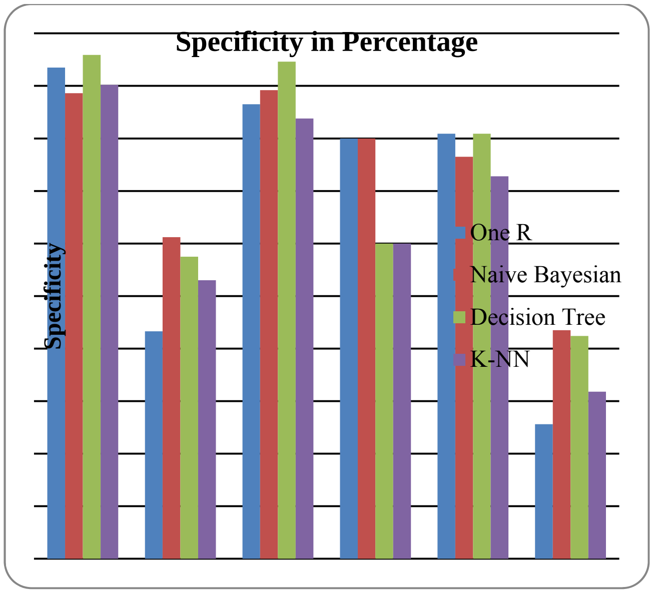

Fig. 4: Specificity values in percentage of each of the algorithms applied on various datasets.

6. CONCLUSION

Our motivation was to check which algorithm performs well on most cases. In this study, we found that no algorithm is too good or too bad in performance. The Performance depends on the size and features of datasets. We took some random datasets with different features and sizes, and applied four popular algorithms on them. All the algorithms performed well and there were very minute differences in the scores of performance measures. In order to calculate the overall best performing algorithm in our study,a tableon each performance measures against each datasets and classification algorithms was created along with the graphical representation of it in percentage form for more clear understanding of the scores obtained by each classifiers.

When we observed the overall scores we found that the Naive Bayesian classification model attains more number of highest values according to the performance measures i.e. accuracy, precision, recall and specificity. Then One R algorithm attains the second best performing model followed by K- Nearest Neighbour and Decision Tree.

REFERENCES

[1] Congcong Li , Jie Wang 2, Lei Wang , Luanyun Hu and Peng Gong, “Comparison of Classification Algorithms and Training Sample Sizes in Urban Land Classification with Landsat Thematic Mapper Imagery”,*Remote Sens.* 2014, *6*, 964-983; doi:10.3390/rs6020964

[2] Christopher Sibona and Jon Brickey, "A Statistical Comparison of Classification Algorithms on a Single Data Set" (July29, 2012).*AMCIS 2012 Proceedings*. Paper 2.http://aisel.aisnet.org/amcis2012/proceedings/ResearchMethods/2

[3] Paul R. Harper, “A review and comparison of classification algorithmsfor medical decision making” , Elsevier Ireland Ltd, Health Policy 71 (2005) 315–331,

[4] Mustafa Agaoglu,“Predicting Instructor Performance Using DataMining Techniques in Higher Education”,IEEE Access 2016

[5] J. Han, M. Kamber, and J. Pei, “Data Mining: Concepts and Techniques”, Waltham, MA, USA: Morgan Kaufmann, 2012.

[6] Minakshi Sharma and Suresh Kumar Sharma, “Generalized K-Nearest Neighbour Algorithm- A Predicting Tool”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, November 2013.

[7] Robert Layton, “Learning Data mining with Python”, Packt publishing Ltd, open source community experience distilled, Birmingham-UK, 2015.

[8] Quinlan, J. Ross, "Induction of decision trees", Machine learning 1, no. 1 (1986): 81-106

[9] ShahrukhTeli, Prashasti Kanikar, “A Survey on Decision Tree Based Approaches in Data Mining”, Volume 5, Issue 4, 2015 International Journal of Advanced Research in Computer Science and Software Engineering.

[10] Fahd SabryEsmail, M. Badr Senousy, Mohamed Ragaie, “Predication Model for Leukaemia Diseases Based on Data Mining Classification Algorithms with Best Accuracy”, World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:10, No:5, 2016.

[11] Chung-Chian Hsu a, Yan-Ping Huang, Keng-Wei Chang, “Extended Naive Bayes classifier for mixed data”, Expert Systems with Applications 35 (2008) 1080–1083