International Conference on Advancements in Computing Technologies - ICACT 2018 ISSN: 2454-4248
Volume: 4 Issue: 2 55 – 58

Performance Analysis of Different Classification Methods in Data Mining

Maria Nithi[1], Jeya Priya[2]

P.G. Student, Department of Computer Science,Stella Maris College Chennai, India[1]

Associate Professor, Department of Computer Science, Stella Maris College Chennai, India[2]

Abstract:Classification is a major technique in data mining and widely used in various fields. Classification is a data mining function which assigns items in a collection to target categories or classes. Classification models predict categorical class labels**.** The goal of classification is to accurately predict the target class for each case in the data. This paper collects data from different dataset and six different classification algorithms applied to generate the accuracy of the algorithm and find the best algorithm. The algorithms are j48, random forest algorithm, ZeroR, Multilayer Perceptron, 1BK, naïve Bayes.

Keywords*:Data Mining, Decision Tree, Classification, j48, ZeroR, Multilayer Perceptron, 1BK, NavieBayes, Random forest algorithm.*

*****

I. INTRODUCTION

Classification is supervised learning .classification is used to predict categorical class labels like discrete or nominal.it uses labels of the training data to classify new data. There are two main steps in classification first step is we have to construct classification model based on training data and second is usage of model, we have to test its accuracy before using our model.

In this research work the classification algorithm like j48, random forest algorithm, ZeroR, IBK, multilayer perceptron, naïve bayes are used for classification. This paper uses four different dataset with different instance this paper uses the popular classification tool called WEKA .Weka is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Decision tree algorithm j48, ZeroR, Multilayer Perceptron, IBK, Naïve Bayes and random forest algorithm is used for the classification using WEKA tool.

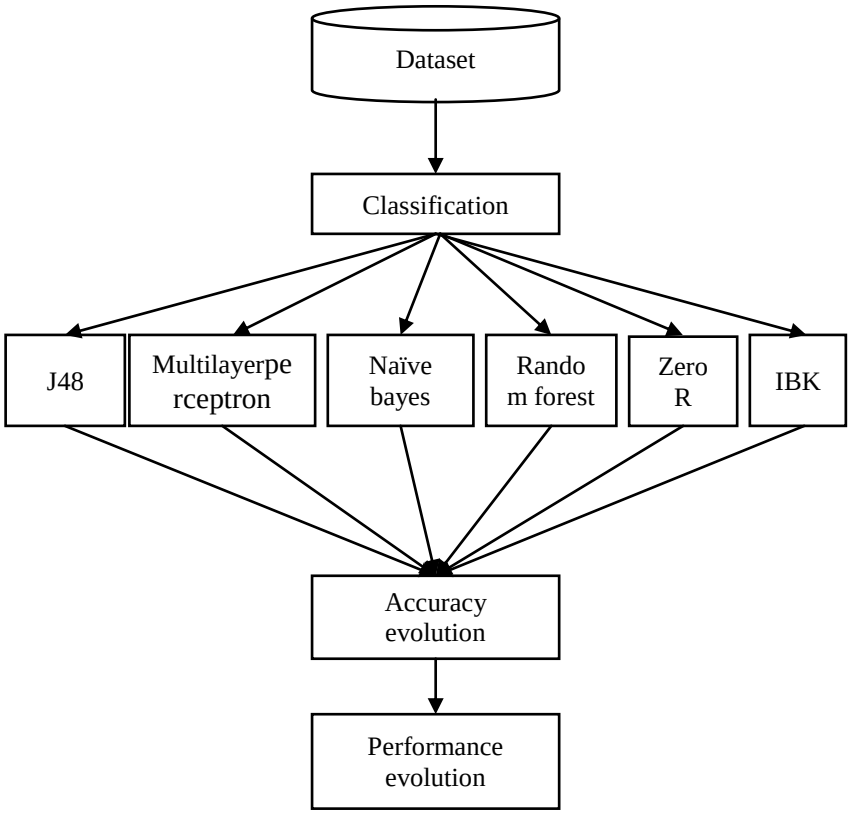
II. LITERATURE REVIEW

Liver disorder diagnosis using linear, nonlinear and decision tree classification algorithm [1] Liver disease consistent listed as one of the top ten fatal diseases around the globe costing millions of lives every year. Lack of timely diagnosis and appropriate treatment is visible with the registered cases of liver disorders in hospitals the study accordingly deployed a number of linear, nonlinear and decision tree classification Algorithms and presented a predictive model for liver disorder diagnosis. These algorithms include LDA, DLDA, QDA, DQDA, NB, ANN and CART, out of this entire algorithm CART is the best algorithm for classification accuracy prediction.Diagnosis of Hepatitis using Decision tree algorithm [2] they have used algorithms such as C5.0, PCL, J48, and fuzzy rule. In this Paper, C4.5 algorithm is used which is an efficient one than the existing algorithms. Since, the number of Attributes is lesser. The complexity of the decision tree is reduced. Analysis of classification algorithms for liver disease diagnosis [3] In terms of accuracy, precision, sensitivity and specificity K* algorithm was found to be superior because it had the lowest error rate with highest accuracy compared to NBC, Bagging, Logistic and Rep Tree with both AP and UCLA data sets. Therefore this algorithm (K star) is most suitable for liver disease diagnosis. An analysis of hepatitis c virus prediction using different data mining techniques [4] work is to provide a study of different data mining techniques that can be employed in Automated HCV infection prediction systems. The system extracts hidden knowledge from a historical HCV patient's database. They have been used three algorithm naïve Bayes, neural network and decision tree for the prediction among that algorithm decision tree gives the best prediction accuracy.

III.METHODOLOGY

The objective of this paper paper is to find out the performance of the six different classification algorithms and find the accuracy of algorithm which is working better. Classification is the most commonly applied data mining technique. The proposed work of this paper is first fetching data's from the four different dataset with different instance and then classify it by classification algorithms like ZeroR, Multilayer Perceptron,1BK,Naïve Bayes, random forest algorithm and j48 algorithm and finally find out the accuracy of the algorithm and performance of the six classification algorithm.

 ISSN: 2454-4248
55 – 58



1. J48

A decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The internal nodes of a decision tree denote the different attributes, the branches between the nodes tell us the possible values that these attributes can have in the observed samples, while the terminal nodes tell us the final valueofthedependentvariable.

2. Random forest algorithm

Random forest algorithm is a supervised classification algorithm.This algorithm creates the forest with a number of trees. The more trees inthe forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the highaccuracy results.

3. Naïve Bayes Algorithm

In machine learning, naiveBayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers

4. ZeroR

ZeroR is the simplest classification method which relies in the target and ignores all predictors.
ZeroR classifiers simply predict the majority category (class).although there is no predictability power in ZeroR, itis useful for determining a baseline performance as a benchmark for other classification methods.

5. Multilayer Perceptron

A multilayerperceptron (MLP) is a class of feed forward artificial neural network. An MLP consists of at least three layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable. Multilayer perceptron are sometimes colloquially referred to as "vanilla" neural networks, especially when they have a single hidden layer.

6. 1BK

The *k*-nearest neighbor's algorithm is also known as IBK.In pattern recognition, the *k*-nearest neighbor's algorithm (*k*-NN) is a non-parametric method used for classification and regression. In both cases, the input

consists of the *k*-closest training examples in the feature space. The output depends on whether *k*-NN is used for classification or regression:

- ✓ In *k-NN* classification, the output is a class membership. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its *k* nearest neighbours (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbour.
- ✓ In k-NN regression, the output is the property value for the object. This value is the average of the values of its *k* nearest neighbours.

Algorithms	Splitting criteria	Attribute type	Pruning strategy	Technique	Tree type	Outlier detection
J48	gain ratio	Only Nominal	Pruning is done	greedy technique	classification	-
ZeroR	-	Nominal	-	-	-	
Multilayer Perceptron	-	Nominal or categorical	-	supervised learning technique	classification or regression	pattern based detection of outliers
IBK	-	Nominal or categorical	Pruning is done	Non parametric method	Classification and regression	Anomaly detection.
Naïve Bayes	-	continuous attribute	-	baseline method	classification	Beta-divergence method
Random Forest	information gain or the Gini impurity	Nominal or categorical	-	ensemble learning method	classification or regression	Local Outlier Factor

IV. EXPERIMENTAL RESULTS

1. WEKA Tool

Weka is a 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.

2. Result Analysis

Different datasets used to analysis the performance of the classifications algorithms. The result of the analysis shows that random forest algorithm gives the best classification accuracy than other classification algorithms.

Dataset	Instances	Features	Algorithms	Accuracy
BUPA liver disorders	345	7	J48 ZeroR Multilayer Perceptron 1BK Naïve Bayes Random Forest	68.6957 % 57.971 % 71.5942 % 62.8986 % 55.3623 % **73.0435 %**
Thyroid disease dataset	3772	30	J48 ZeroR Multilayer Perceptron 1BK Naïve Bayes Random Forest	**99.5758%** 92.2853 % 94.1676 % 91.5164 % 95.281 % 99.3107 %

 ISSN: 2454-4248
55 – 58

Nursery Database	12960	9	J48 ZeroR Multilayer Perceptron 1BK Naïve Bayes Random Forest	97.0525 % 33.3333 % **99.7299 %** 98.3796 % 90.3241 % 99.0664 %
Letter Image Recognition Data	20000	16	J48 ZeroR Multilayer Perceptron 1BK Naïve Bayes Random Forest	87.98 % 4.065 % 82.085 % 96.03 % 64.115 % **96.505 %**

V. CONCLUSION

The proposed system is shows that the performance of different classifiers with different dataset with different number of instances. After analyzing the six classifiers the best one is random forest algorithm and its performance accuracy is better comparing with other algorithms.

REFERENCES

[1] Aman Singh, Babita Pandey Liver disorder diagnosis using linear, Nonlinear and decision tree classification Algorithms Department of Computer Science and Engineering, Lovely Professional University, Jalandhar, Punjab – 144411, India Vol 8 No 5 Oct-Nov 2016

[2] V.Shankar sowmien , V.Sugumaran , C.P.Karthikeyan, T.R.Vijayaram Diagnosis of Hepatitis using Decision tree Algorithm School of Mechanical and building Sciences, VIT University, Chennai, India Vol 8 No 3 Jun-Jul 2016

[3] Shapla Rani Ghosh and Sajjad Waheed Analysis of classification algorithms for liver disease diagnosis Dept. of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Tangail-1902, Bangladesh. Volume 05,2017

[4] ahmed a. a. radwan, tarekabd-el-hafeez & heba mamdouh an analysis of hepatitis c virus prediction using different data mining techniques International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR)ISSN 2249-6831 Vol. 3, Issue 4, Oct 2013.

[5] tapas rajan baitharu,subhendu kumar pani,analysis of data mining for health care decision support system using liver disorder dataset ,Orissa engineering college,2016.