© 2018 IJSRST | Volume 4 | Issue 2 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Comparative Study and Analysis of Classification Algorithms In Data Mining Using Diabetic Dataset

R. S. Suryakirani*[1], R. Porkodi[2]
*[1]PG Scholar, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India
[2]Assistant Professor, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India

ABSTRACT

Classification is use to categorize each item in a set of data into one of predefined set of module or grouping. The data analysis task classification is where a model or classifier is constructed to predict categorical labels. The goal of the classification is to accurately predict the target class for each case in the data. The field of data mining due to its enormous success in terms of broad ranging application achievements and scientific progress, understanding. Many data mining application have been successfully implemented in various domains like healthcare, finance, retail, telecommunication, fraud detection and risk analysis etc. This paper presents the study and analysis of four classification algorithms namely J48, Random tree, Decision tree and Naive Bayes for Diabetic dataset and the performance are compared using the measures such as computing time, Correctly Classified Instances, Incorrectly Classified Instances, kappa statistics, Precision, Recall and F measure. The experimental result shows that J48 provides better accuracy than the Random tree, Decision tree and Naive Bayes.

Keywords : Classification Algorithms, Naive Bayes, Random Tree, Decision Tree, J48.

I. INTRODUCTION

Data mining refers to extract information from huge amount of data. Data mining is the process of sorting through large data sets to identify patterns and establish relationship to solve problems through data analysis. Data mining tools allow enterprise to predict future trends. In other words, we can say that data mining is the procedure of mining knowledge from data. Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources [1]. Data mining evaluation are data warehousing. Data warehousing is the process of constructing and using the data warehouse. A data warehouse is constructed by integrating the data from multiple heterogeneous sources. It supports analytical reporting, structured and/or ad hoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations. There are several techniques are there in data mining are classification, association, clustering, prediction, sequential pattern, decision tree. Data and information or knowledge has a significant role on human activities. Due to the importance of extract knowledge, information from the huge data repositories, Data mining has become an essential part in a variety of field of human life including business, education, education, Scientific etc. Classification is used to classify each item in a set of data into one of predefined set of classes or group. The data investigation job classification is where a model or classifier is construct to predict categorical labels. Classification is a data mining function that assigns item in a collection to target categories or classes. The goal of this paper is to accurately predict the target class for each case in the study and analysis of classification algorithm on diabetic dataset [2].

The section I discuss the introduction of data mining and the classification algorithm. Section II gives the brief analysis about literature survey. Section III explains the methods that are used classification algorithm. The result and discussion are explained in Section IV and Section V concludes this analysis work.

II. LITERATURE SURVEY REVIEW

M. Sujatha et.al (2013) [3] proposed a Classification model by using different classes according to specific constraints and provides a survey of numerous data mining classification techniques for innovative database applications. There are several major kinds of classification algorithms including Genetic algorithm C4.5, Naïve Bayes, SVM, KNN, decision tree and CART. Random forest is very high accuracy for gradient boosting.

Geraldin B. Dela Cruz, et.al (2014) [4] proposed an efficient data mining methodology based on PCA-GA is explored, presented and implemented to characterize agricultural crops in different agricultural datasets.

Abhale Babasaheb Annasaheb, Vijay Kumar Verma (2016) [5] proposed Heart Disease Prediction problem using Classification algorithm. They proposed efficient classification algorithm using heart disease prediction compared on basis of sensitivity, specificity, accuracy, error rate, true positive rate and false positive rate. Classification algorithm including Decision Tree Induction, Bayesian Classification, Support Vector Machines, Rule-based classification, Neural Network Classifier and K-Nearest Neighbor Classifier. Neural network gives the best accuracy when comparing the other algorithm.

P.Yasodha N.R. Ananthanarayanan (2014) [6] proposed study and analysis of different diabetic patient data that can be performed using classification algorithm. The authors used computing time, correctly /incorrectly classified instances, and kappa statistics, MAE, RMSE, RAE and RRSE. The classification techniques are J48, Random tree, REP, LAD for comparison. J48 is the better accuracy when comparing other algorithms.

Shelly Gupta Dharmindar Kumar Anand Sharma (2011) [7] proposed data mining classification techniques applies for breast cancer diagnosis and prognosis. The classification techniques analyzed here are Decision Trees, Support Vector Machine, Genetic Algorithms / Evolutionary Programming, Fuzzy Sets, Neural Networks, Rough Sets. The results of SVM and ANN prediction models were found comparatively more accurate than the other algorithms.

Sonali Agarwal et.al (2012) [8] proposed a Data Mining application in education using data classification and decision tree approach from a community college database and various classification approaches have been performed by comparative analysis. The classification algorithms used are Logistic, Multilayer Perceptron, LIBSVM, RBF, Network, Simple Logistic, SMO, Voted Perceptron, and Winnow. The study work also suggests that for the given data set LIBSVM with Radial base Kernel has been in use as a most excellent choice for data classification.

Mrs. M.S. Mythili, Dr. A.R.Mohamed Shanavas (2014) [9] analyzed and evaluated the school students' performance by Knowledge Discovery from Data (KDD) process for analyzing student performance. The various classification algorithms used are J48, Random Forest, Multilayer Perceptron, IB1 and Decision Table are used and observed that random forest algorithm provides better accuracy.

Dorina Kabakchieva (2012) [10] proposed a system using data mining application to analyse the students performance. The various classification algorithm are used to improve the students performance are One R

Rule Learner, Decision Tree, Neural Network and K-Nearest Neighbor and the highest accuracy is achieved for the Neural Network model 73.59%.

D.SindhujaR. Jemina Priyadarsini (2016) [11] proposed classification techniques in data mining for analyzing liver disease disorder. The classification algorithms used are c4.5, Naive Bayes Classifier, Naive Bayes Classifier, Back propagation neural network and Support vector machine. It is seen that c4.5 have better result compare to other algorithms.

Hlaudi Daniel Masethe, Mosima Anna Masethe (2014) [12] proposed a prediction system for heart disease for Data mining algorithms such as Bayes Net, J48, Naïve Bayes, Rep tree and Cart are used in their study. J48 shows the best accurate result when compare the other algorithms.

III. METHODOLOGY

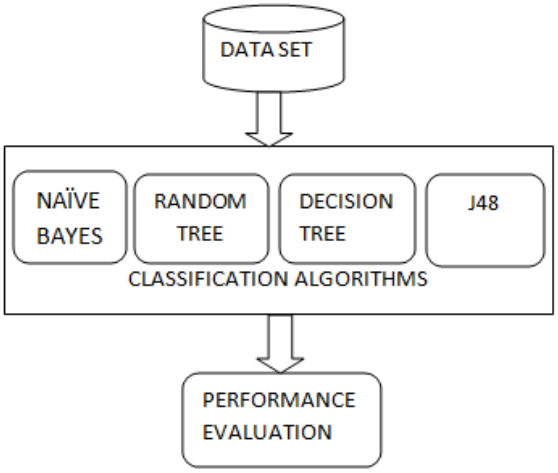

Figure 1. Methodology

The Fig.1 shows the methodology of the proposed work that consist of two phases namely classification phase and performance evaluation phase. The diabetic patient dataset is chosen as an experimental dataset. The classification phase uses 4 classification algorithms namely Naïve Bayes, Random tree, Decision tree and J48.These algorithms are analyzed and validated using the different performance evaluation metrics.

1. Classification Algorithms

Classification algorithm in data mining are capable of processing a large amount of data. It can be used to predict categorical class labels and classifies data based on training set and class labels and it can be used for classifying newly available data [13].

1.1 Naïve Bayes

This classifier is based on the Bayes rule of conditional probability. It uses all of the attributes contained in the data, and analyses them individually as though they are equally important and independent of each other. The Naïve Bayes classifier works on a simple, but comparatively intuitive concept. It makes use of the variables contained in the data sample, by observing them individually, independent of each other. It considers each of the attributes separately when classifying a new instance. It assumes that one attribute works independently of the other attributes contained by the sample [14].

1.2 Random tree

A Random tree is a tree built randomly from a set of probable tree having K random features at each and every node. In this context ‘‘at random’’ means that in the group of tree each has an equal possibility of being sampled. Random tree can be generated proficiently and the combination of large sets of random trees generally leads to accurate models. An extensive research in the current years over Random tree in the field of machine language is carried out [15].

1.3 Decision tree

Decision tree is a flow chart like tree structure Leaf nodes represent class labels or class distribution. Decision tree is a classifier in which each non-terminal node represents either a test or decision for the given data item. Which branch to be select next is depends upon the outcome of the test. To classify a given data item, need to from start at the root node and follow the assertions down until we reach a terminal node or leaf node. Decision is made when a

terminal node is approached. Decision trees use recursive data partitioning. The important things in decision tree are attribute selection measure. There is important parameter used for attribute selection. The attribute with highest information gain is used to be selected as a root [9].

1.4 J48

J4.8 decision trees algorithm is an open source Java implementation of the C4.5. It grows a tree and uses divide-and-conquer algorithm. It 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. To classify a new item, it creates a decision tree based on the attribute values of the training data. When it encounters a set of items in a training set, it identifies the attribute that discriminates. It uses information gain to tell us most about the data instances so that it can classify them the best [14].

IV. RESULT AND DISCUSSION

1. Dataset Description

The Table 1 shows the Diabetic data set with 760 instances and 8 following attributes with numeric values are considered: preg, plas, pres, skin, insu, mass, pedi, age, class has been used for analyses diabetic patient data due to its proficiency of disease [15].

Table 1 Diabetic dataset

ATTRIBUTE	DESCRIPTION	POSSIBLE VALUES
Preg	Number of times pregnant	Numeric
Plas	Plasma glucose concentration of 2 hour in an oral glucose tolerance test	Numeric
Pres	Diastolic blood pressure(mm hg)	Numeric
Skin	Triceps skin fold thickness(mm)	Numeric
Insu	2-Hour serum insulin(mu U/ml)	Numeric
Mass	Body mass index(weight in kg/(height in m)^2)	Numeric
Prdi	Diabetes pedigree function	Numeric
Age	Age{years)	Numeric
Class	Class variable	1 or 0 (positive or negative)

The Table 2 shows the four classification models, generated with the selected data mining algorithms, are compared by using the following evaluation measures: % of correctly/incorrectly classified instances and Kappa Statistic. These are well known measures for evaluation of data mining models for classification. Kappa statistic is a measure of the degree of nonrandom agreement between observer and measurement of the same categorical variable. When comparing the kappa statistics in different classifiers are having that j48 gives the highest values. When comparing the time taken for different algorithm the Naive Bayes takes the least computing time. The Fig.2 shows that correctly and incorrectly classified instances from the result of the four classification instances and observed that j48 algorithms gives better classification result than the other algorithm.

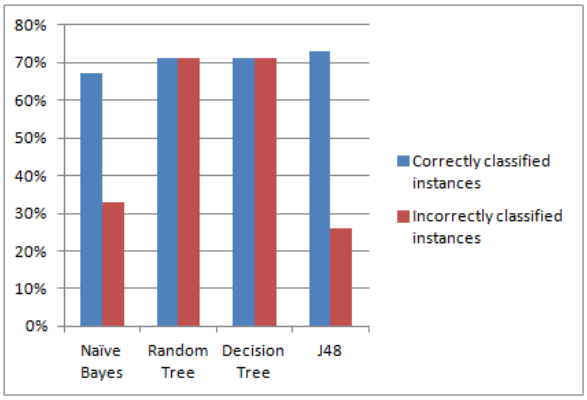

Figure 2. Correctly/Incorrectly classified instances

Table 2. Result of classifiers

EVALUATION CRETERIA	CLASSIFIERS			
	NAVIE BAYAS	RANDOM TREE	DECISION TREE	J48
Correctly Classified Instances	512	536	554	557
	67.3%	70.5%	70.5%	73.2 %
Incorrectly Classified Instances	248	224	206	203
	32.6%	29.4%	27.1%	26.7 %
Kappa Statistic	0.28	0.33	0.37	0.38
Time Taken	0.03Sec	0.14Sec	0.19Sec	0.3Sec
Accuracy	67.3%	70.5%	72.8%	73.2 %

Table 3. Performance Evaluation

CLASSIFIER	PRECISION	RECALL	F MEASURE
NAVIE BAYAS	0.67	0.67	0.67
RANDOM TREE	0.70	0.70	0.70
DECISION TREE	0.72	0.72	0.72
J48	0.72	0.73	0.72

The four classification algorithms have been validated using the important metrics such as precision, recall and f-measure and these validation results are shown in Table 3 and same is depicted in fig.3. The result shows that j48, decision tree classifier almost give better classification accuracy as 0.73 and 0.72 respectively than the random tree and naïve Bayes classifiers. The classification accuracy of Random tree and Naïve Bayes are 0.70 and 0.67 respectively.

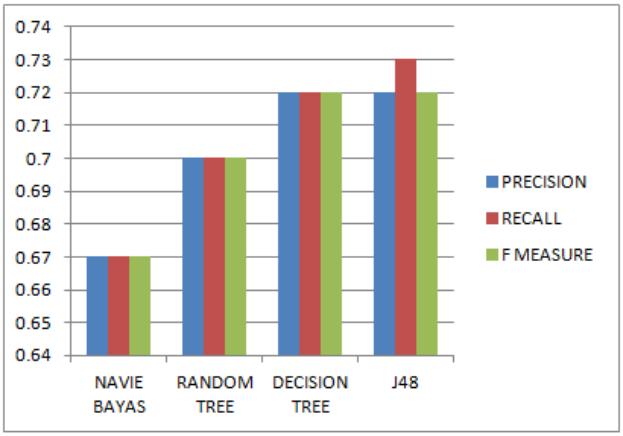

Figure 3. Performance Evaluation

V. CONCLUSION AND FUTURE WORK

This paper conducted the study and analysis of four classification algorithms and experimental result shows that J48 classifier gives better accuracy 73.28% which gives 0.3 seconds for training.The second best algorithm is the decision tree which gives the accuracy 72% and takes 0.19 seconds. The third best algorithm is random tree which gives the accuracy 70.5% and takes 0.14 seconds. Finally Naïve Bayes is the least accuracy 67.3% and takes 0.03 seconds.

The future work will be focused on using the other classification algorithms of data mining. It is a known fact that the performance of an algorithm is dependent on the domain and the type of the data set. Hence, the usage of other classification algorithms like machine learning will be explore in future.

VI. REFERENCES

[1]. http://en.wikipedia.org/wiki/Data_mining
[2]. http://www.flatworldsolutionscom/data-management/articles/data-mining -future-trends. php.
[3]. D.Sindhuja R. Jemina Priyadarsini, International Journal of Computer Science and Mobile Computing" A Survey On Classification

Techniques In Data Mining for Analysing Liver Disease Disorder".

[4]. Geraldin B. Dela Cruz, Member, IACSIT, Bobby D. Gerardo, and Bartolome T. Tanguilig III" Agricultural Crops Classification Models Based on PCA-GA Implementation in Data Mining"

[5]. Abhale Babasaheb Annasaheb, Vijay kumarverma (2006)," Data Mining Classification Techniques: A Recent Survey "International Journal of Emerging Technologies in Engineering Research (IJETER).

[6]. P.Yasodha -Pachiyappa's college for women, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya, N.R. Ananthanarayanan Pachiyappa's college for women,Sri Chandra sekharendraSaraswathiViswaMahavidyalaya " Comparative Study of Diabetic Patient Data's Using Classification Algorithm in WEKA Tool.

[7]. Shelly Gupta,DharmindarKumar,Anand Sharma," Indian Journal of Computer Science and Engineering (IJCSE)", Data Mining Classification Techniques Applied.For Breast Cancer Diagnosis And Prognosis.

[8]. Sonali Agarwal, G. N. Pandey, and M. D. Tiwari, International Journal of e-Education, e-Business, e-Management and e-Learning," Data Mining in; Education: Data Classification and Decision Tree Approach".

[9]. Mrs. M.S. Mythili1, Dr. A.R.Mohamed Shanavas2 IOSR Journal of Computer Engineering (IOSR-JCE)" An Analysis of students' performance using classification algorithms"

[10]. DorinaKabakchieva, International Journal of Computer Science and Management Research," Students Performance Prediction By Using Data Mining Classification Algorithm".

[11]. M. Sujatha,S. Prabhakar,Dr. G. Lavanya Devi , International Journal of Innovations in Engineering and Technology (IJIET "A Survey Of Classification Techniques In Data Mining".

[12]. Hlaudi Daniel Masethe, Mosima Anna Masethe, Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II " Prediction of Heart Disease using Classification Algorithms".

[13]. Sagar S.Nikam,(),An international research journal of computer science and technology,"A comparative study of classification techniques in data mining algorithms".

[14]. http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/diabetes.arff.

[15]. R.Ranjani Rani,P.Manikandan,D.Ramya chitre,International Journal of Advanced Research in Computer Science, "An Emprical Analysis of Classification Tree Algorithm for Protein Datasets".

[16]. Mats Jontell, Oral medicine, Sahlgrenska Academy,Göteborg University (1998) "A Computerised Teaching Aid in Oral Medicine and Oral Pathology. " OlofTorgersson, department of Computing Science, Chalmers University of Technology, Göteborg.

[17]. T. Mitchell, "Decision Tree Learning", in T. Mitchell, Machine Learning (1997) the McGraw- Hill Companies, Inc., pp. 52-78.

[18]. Witten Ian H., E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Ch. 8, © 2000 Morgan Kaufmann Publishers

[19]. http://grb.mnsu.edu/grbts/doc/manual/J48_Decision_Trees.html, accessed.

[20]. Jiawei Han and Micheline KamberData Mining: Concepts and Techniques, 2ndedition.

[21]. Baik, S. Bala, J. (2004), A Decision Tree Algorithm For Distributed Data Mining.