ISSN:0975-9646

Performance Analysis of Classification Algorithms Using Healthcare Dataset

D. Rajeswara rao [*1], Vidyullata Pellakuri [#2], SathishTallam [#3], T. Ramya Harika [#4]

**1. Professor, Department of Computer Science, K L University, Guntur, Andhra Pradesh*
#2. Research Scholar, Department of Computer Science, K L University, Guntur, Andhra Pradesh
#3. Student, Department of Computer Science, K L University, Guntur, Andhra Pradesh
#4. Student, Department of Computer Science, K L University, Guntur, Andhra Pradesh

Abstract: **This paper presents a comparison among the classifiers FT, LMT, RandomForest, SimpleCart in terms of their accuracy by applying them on three completely different datasets of carcinoma, breast cancer and also the information associated with cardiovascular disease that vary greatly in their range of attributes in weka tool. The experimental results shows that there's a major distinction within the accuracy of a same algorithm once applied on three completely different datasets. The experiment shows that the accuracy differs for associate algorithm even on same dataset once the algorithm is applied for training dataset and also the testing dataset. The paper finally proposes ft algorithm as the best algorithm among the four algorithms (FT, LMT, RandomForest, SimpleCart) in terms of consistency of its accuracy for all the three datasets and that don't have much difference between the accuracy of its training dataset and also the testing dataset. But these results are solely confined to the WEKA tool only.**

Keywords: Breast cancer, Carcinoma, Cardiovascular disease, Classification, Testing Dataset, Training Dataset Weka.

I. INTRODUCTION

Data mining finds valuable information hidden in large masses of data. Data mining is the analysis of information and the role of software techniques for discovering patterns and regularities in sets of data. Data Mining, is an interdisciplinary study. It can be used in Machine Learning, High Performance Computing, Databases, Visualization,

Mathematics, Statistics etc. Data Mining Tools Used in 2005 are Analytic tools. For Enterprise-level the tools are Isaac, IBM, Insightful, KXEN, Oracle, SAS, and SPSS. For Department level, Angoss, CART/MARS/TreeNet/Random Forests, Equbits, GhostMiner, Gornik, Mineset, MATLAB, Megaputer, Microsoft SQL Server, Statsoft Statistica, Think Analytics and for Personal-level are Excel, See5 & open Free tools are C4.5, R, Weka, Xelopes.

Data mining has a wide set of applications. One of its major application area is the domain of healthcare. Data mining in healthcare is a promising new area of research. Data mining and machine learning essentially depends on classification. Data mining in healthcare can be used for various purposes. Many researches are going on various classifiers and feature selection techniques.

The classification techniques in the data mining can be applied to the healthcare dataset in order to make valuable predictions and important conclusions. In order to offer predictions and conclusions the accuracy in the results plays a very important role. But the accuracy may be varied depending upon various conditions such as size of the dataset, number of attributes, type of attributes, etc. The accuracy also depends on the classifier that is being used. This paper gives the accuracy of different classification algorithms when applied on the datasets with different number of attributes.

II. LITERATURE SURVEY

MonaliDey and Siddharth Swarup Rautaray [3] discussed the user oriented approach provided by data mining to novel and hidden information in the data. Valuable knowledge can be discovered from application of data mining techniques in healthcare system. Data mining applications can greatly benefit all parties involved in the healthcare industry. Divya Tomar et al [1] discussed the role of data mining for uncovering new trends in healthcare organization which in turn is helpful for all the parties associated with this field. They explored the utility of various Data Mining techniques such as classification,clustering,regression,association in health domain. They highlighted the challenges, applications and future issues of Data Mining in healthcare. Illhoi Yoo et al [2] suggested that data mining can help researchers to gain both novel and deep insights and can facilitate unprecedented understanding of large biomedical datasets. Data mining can expose new biomedical and healthcare knowledge for clinical and administrative decision making as well as generate scientific hypotheses from large clinical databases, experimental data, and/or biomedical literature. The successful application of data mining by health related organizations that has helped to predict health insurance fraud and under-diagnosed patients, recognise and classify at-risk people in terms of health with the goal of reducing healthcare cost. Vahid Rafe [4] discussed the widespread use of medical information systems and explosive growth of medical databases require traditional manual data analysis to be coupled with methods for efficient computer assisted analysis. Karina Gibert, Miquel Sànchez-Marrè and Víctor Codina [5] proposed a conceptual map of the most common data mining techniques. They identified the first main decisional criteria used by human experts in real decisions and the conceptual map is organized based on them. The proposal helps environmental data miners in the conceptual organization and rational understanding of the

broad scope of data mining methods; also helps non-expert data miners to improve decisions in real applications. Qasem A. Al-Radaideh and Eman Al Nagi [6] proposed a classification model to predict the employee's production by working on the performance with many attributes. Shelly Gupta, Dharminder Kumar and Anand Sharma [7] have shown that different classification techniques behave differently on different datasets depending on the nature of their attributes and size. Jayanthi Ranjan [8] suggested that with data mining techniques, we could try to find alternative measures of relief, and endorse the drug in another way: on top of curing the disease in a standard way, with our drug you get some extras associated to the competitor. Milan Kumari, 2Sunila Godara [9] compared the classification techniques on basis of Sensitivity, Specificity, Error Rate, Accuracy, True Positive Rate and False Positive Rate. The study showed that Support Vector Machine model turned out to be best classifier for cardiovascular disease prediction. In this paper the results of the accuracy of an algorithm for a dataset depending upon the number of attributes of that dataset is discussed.

III. DATASET DESCRIPTION

This paper considers three datasets with different number of attributes. All these datasets encompasses the data about three different diseases which are the lung cancer, breast cancer and the heart disease. In order to apply classifiers on these datasets one need to have a clear understanding of the data that we are going to classify. The primary dataset that is being classified is about the lung cancer. Lung cancer is the most usual cancer in humans and the fifth most common in women, even producing more cancer-related deaths in women than breast cancer [11]. Lung cancer, too known as carcinoma of the lung or pulmonary carcinoma, is a malicious lung tumour characterized by uncontrolled cell growth in tissues of the lung. The vast majority (80–90%) of cases of lung cancer are due to long-term exposure to tobacco smoke. About 10–15% of cases occur in non-smokers. These cases are often caused by a combination of genetic factors. The dataset collected consists of the genetic codes of patients both with cancer and without cancer. The secondary dataset is about the breast cancer. The second leading reason of death in women next to lung cancer is the breast cancer. The data is about the different features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass [12]. They describe features of the cell nuclei present in the image which helps to know whether the cancer is benign or malignant. The last dataset is about the heart disease. This data comprises of the different features which helps to conclude whether they lead to a heart attack or not [13]. The data originally consists of 76 attributes out of which only 14 attributes are considered. All the three datasets are collecting from UCI machine learning repository.

Data mining methods in the medical domain are helping due to the increasing effectiveness of classifications that help the doctors especially in decision making. This paper proposes a classification algorithm which is more reliable for every type of datasets. This paper also proposes that the accuracy of an algorithm varies from one dataset to other especially depending on the number of attributes. The first dataset which is of lung cancer is a very huge dataset with 617 instances and 76 attributes. The dataset of breast cancer consists of 569 instances and 32 attributes. The third dataset of heart disease is also a very huge dataset which actually have 76 attributes but only 14 attributes are considered for the purpose of analysis [11].There are various tools that are available for the purpose of data mining. In this paper for the purpose of data mining and the analysis of performance of various algorithms, we use the tool called

WEKA.

The weka tool provides many classification algorithms. This paper considers four algorithms that are FT, LMT, Random Forest and Simple cart which are the tree classification algorithms [10]. *FT* is a classifier algorithm for constructing 'Functional Trees'. It could have logistic regression functions at the inner nodes/leaves. The algorithm can deal with numeric, nominal attributes, missing values, binary and multi class variables [4]. *LMT* is a classification model which combines both logistic regression and decision tree learning. It makes a tree with binary and multiclass variables, numeric and missing values. This technique uses logistic regression tree**.** *RandomForest* is an ensemble learning method for classification and regression which operated by building multitude of decision trees. It runs effectively on large data bases and handles thousands of input variables. *SimpleCART/CART* is defined as Classification and Regression Tree Algorithm which is developed by Leo Breiman. CART is used for data exploration and prediction. CART uses learning sample set of historical data set with pre assigned classes. Feature selection is the important aspect in the classification process. It is of a great advantage to limit the number of attributes for the classification in order to have good prediction and less computationally intensive models [5]. This paper infers that less number of attributes in the dataset leads to the less accuracy when compared to the accuracy of other datasets with more number of attributes. However the amount of accuracy also be contingent on two types of dataset that we are using to classify. Here two types of datasets which are the training dataset and the testing dataset. Training dataset means loading the full dataset whereas testing dataset means selecting a correct percentage of data to be tested. The calculation of accuracy on training dataset alone may not be sufficient since it tends to give more accuracy even when the algorithm over-fit the data. The accuracy on the testing data is more important since it shows how the algorithm generalise and perform with new data.

IV. RESULTS AND DISCUSSIONS

This paper discusses concerning the performance of four algorithms as mentioned above for three datasets with completely different range of attributes. The results for each training set and also the testing set as they each vary considerably. The testing information are often created by specifying the proper share for the split.

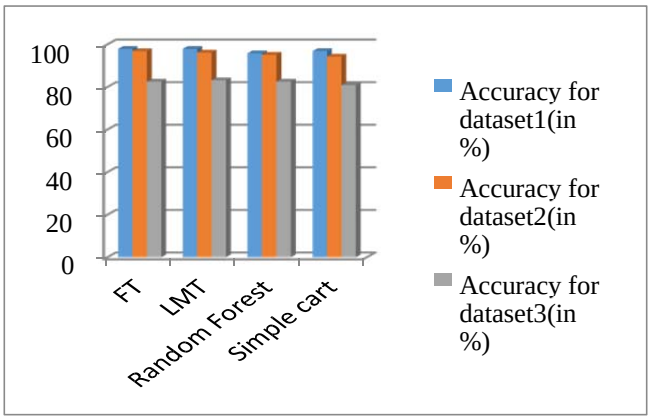

Fig:1 Accuracy for four classifiers on training datasets

Fig.1 shows the comparison of accuracies of four classifiers (FT, LMT, Random

Forest, simpleCart) based on tenfold cross validation as a test method. This is the result obtained from the training information. Here dataset1 refers to the carcinoma (lungcancer) dataset with highest range of attributes. Dataset2 refers to the data of the breast cancer with medium vary of attributes with thirty six attributes. Whereas the last dataset is that the cardiovascular disease dataset with lowest range of aattributes. .

Fig:2 shows the comparison of four classifiers on three datasets (testing datasets) based on tenfold cross validation as a test method. Fig: 3 shows a comparison of accuracy of training dataset and testing dataset for the dataset1 i.e., for the carcinoma dataset. The dataset1 have the highest range of attributes. We are able to see that the testing dataset offers the 100 percent accuracy with each share split. Though the distinction between the training dataset and also the testing dataset isn't much but for the crucial dataset like the medical dataset, even 0.1 percent of accuracy will result in amendment of things. Fig:4 shows the comparison of accuracies of four algorithms on the training dataset and the testing dataset for the dataset2 which is the data of the breast cancer. From fig: 4, we can see that the accuracy of testing data came down when compared to the accuracy of testing data of the first dataset. From the fig:4 it can be inferred that there isn't any difference between the accuracy of the training dataset and the testing dataset as the number of attributes came down. Fig:5 shows the comparison of accuracies of four algorithms on training dataset and the testing dataset for dataset3 i.e., the data of the heart disease with least number of attributes. The accuracy of the testing data came down gradually along with the number of attributes. From fig:3, fig:4, and fig:5 we can see that the first dataset is having the highest number of attributes and its accuracy on the testing dataset is as much as 100%. But as the number of attributes in the dataset2 came down to a medium number, the graph also came down to the level of the training data.

And finally, in the third dataset the testing data comes down to that of the training data. From the above results it is clear that the accuracy of a particular classifier definitely depends on the number of attributes. Less number of attributes may give good predictions with less computational efforts but also with less accuracy. By comparing the training set and the testing set of individual datasets for each algorithm, the results on testing dataset greatly varies from the results of training dataset depending on the number of attributes. Here in order to estimate the strength of the algorithm we need to consider the accuracy of the testing dataset since it shows how the algorithm generalize. The accuracy of training data alone could be miss-leading.

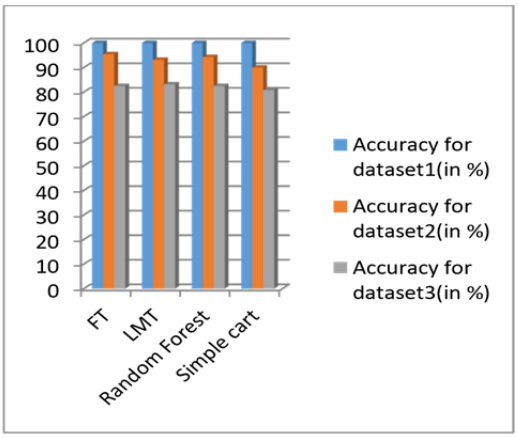

Fig.2 Accuracy of four classifiers on testing dataset

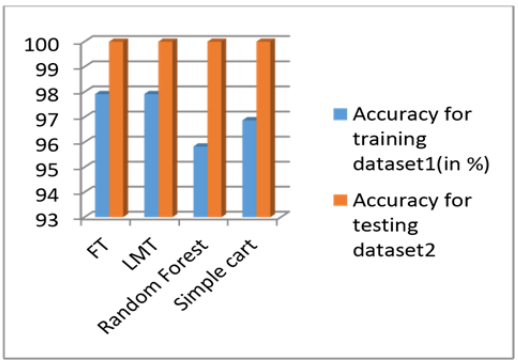

Fig:3 Accuracy of four algorithms on training dataset and testing dataset for dataset1

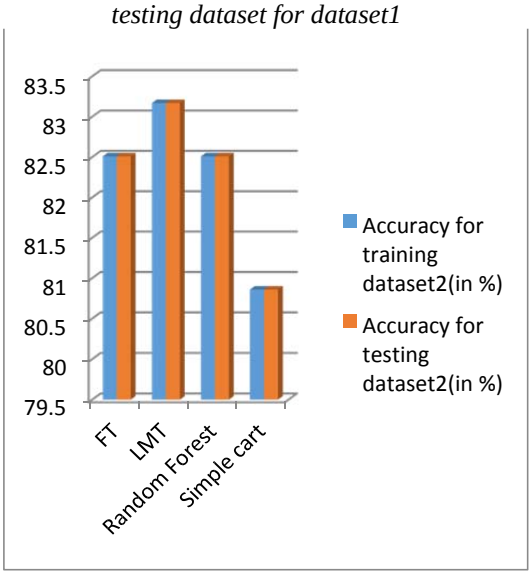

Fig:4 Accuracy of four algorithms on the training dataset and the testing dataset for dataset2

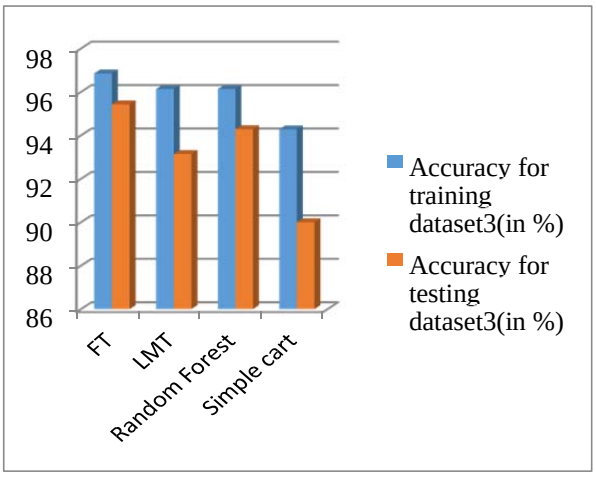

Fig: 5 Accuracy of four algorithms on the training dataset and the testing dataset for dataset3

By studying the accuracies of testing data on all the three datasets we conclude that FT algorithm is a best algorithm in all the three cases since it is the most consistent one among the four and also do not have much difference between the accuracy of training dataset and the testing dataset. If the training dataset alone is considered, there is a huge difference between the first two datasets and the third dataset. The results of FT algorithm is same in the first two datasets and so for the others. However it can also inferred that LMT also gives the best results in almost all the cases. In fact for the dataset2 it gives the highest accuracy for both the training dataset and the testing dataset. But the main problem with the LMT algorithm is that it consumes significantly more time when compared to the FT algorithm and also when it comes to the dataset with the least number of attributes, there is a lot of difference between the accuracy of training dataset and the testing dataset. Therefore the FT algorithm is the best algorithm among the considered four algorithms which can be applied to any dataset with both more number of attributes and less number of attributes and also on both training dataset and the testing dataset since it is consistent and gives the accurate results in less time.

V. Conclusion

The experimental results on the three datasets shows that the FT algorithm is the best classifier among the opposite algorithms that are LMT, RandomForest and SimpleCart. But these results are confined to the weka tool solely. From the experiments it may be concluded that the accuracy of associate algorithm depends upon the number of attributes of that dataset. The results might vary greatly once a similar datasets are classified on different tools like tanegra, rapid mining etc., that are latest tools with in the data mining. This experiment can be extended by applying additional range of classification algorithms on additional range of datasets of various domains.

References

[1] Divya Tomar and Sonali Agarwal(2013), *"A survey on Data Mining approaches for Healthcare"*. International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241-266

[2] IIlhoi, Patricia Alafaireet, et at (2012), **"***Data Mining in Healthcare and Biomedicine: A Survey of the Literature"*. Journal of medical sciences Volume 36, Issue 4, pp 2431-2448

[3] Monali Dey and Siddharth Swarup Rautaray, *"Study and Analysis of Data mining Algorithms for Healthcare Decision Support System"*. International Journal of Computer Science and Information Technologies, Vol. 5 (1), 2014, pp 470-477

[4] Parvathi and Siddharth Rautaray, *"Survey on Data Mining Techniques for the Diagnosis of Diseases in Medical Domain"*. International Journal of Computer Science and Information Technologies, Vol. 5 (1), 2014, pp 838-846.

[5] Karina Gibert, Miquel Sànchez-Marrè and Víctor Codina(2010). "*Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation"*. Proceeding at International Environmental Modelling and Software Society (iEMSs) 2010, Fifth Biennial Meeting, Ottawa, Canada

[6] Qasem A. Al-Radaideh and Eman Al Nagi. "*Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance"*. International Journal of Advanced Computer Science and Applications, Vol. 3, No. 2, 2012, pp 144-151

[7] Dharminder Kumar and Anand Sharma. "*Performance analysis of various data mining classification techniques on healthcare data"*. International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 4, pp 155-169, August 2011.

[8] Jayanthi Ranjan. "*Applications of data mining techniques in pharmaceutical industry"* . Journal of Theoretical and Applied Information Technology (20052007).,pp(61-67)

[9] *Milan Kumari, 2Sunila Godara, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction"*. IJCST Vol. 2, ISSN : 22294333(Print) | ISSN : 0976-8491(Online) Issue 2, June 2011.

[10] D.Lavanya, Dr.K.Usha Rani,..,*" Analysis of feature selection with classification: Breast cancer datasets"*, Indian Journal of Computer Science and Engineering (IJCSE),vol.88-no.11,pp 28-33, October 2011.

[11] A. Choudhary, 2011. Poster: A lung cancer mortality risk calculator based on SEER data. IEEE proceedings of the 1st International Conference on Computational Advances in Bio and Medical Sciences, Feb. 3-5, IEEE Xplore Press, Orlando, FL., pp: 233-233. DOI:10.1109/ICCABS.2011.5729887

[12] Abdelaal, A.M.M., H.A. Sena, M.W. Farouq and A.M. Salem, 2010. Using data mining for assessing diagnosis of breast cancer. Proceedings of the International Multiconference on Computer Science and Information Technology, Oct. 18-20, IEEE Xplore Press, Wisla, pp: 11-17. DOI:10.1109/IMCSIT.2010.5679647

[13] Narendra Kohli and Nishchal K. Verma. "*Arrhythmia classification using SVM with selected features" International Journal of Engineering, Science and Technology Vol. 3, No. 8, 2011, pp. 122-131*

[14] Dr. S.Vijayarani and S.Sudha. "*Comparative Analysis of Classification Function Techniques for Heart Disease Prediction*". International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 3, ISSN (Print) : 2320 – 9798 ISSN (Online): 2320 – 9801 May 2013

[15] http://en.wikipedia.org/wiki/Logistic_.

[16] http://www.datasciencecentral.com

[17] http://www.ijarcsse.com/docs/papers/

[18] http://en.wikipedia.org/wiki/Weka_