Using Data Mining Tools for Breast Cancer Prediction and Analysis

Dr. S. N. Singh
Associate Professor,CS&E Department
ASET, Amity University
Noida, Uttar Pradesh
snsingh36@amity.edu

Shivani Thakral
M.Tech(CS&E)
ASET, Amity University
Noida, Uttar Pradesh
shivanithakral99@gmail.com

Abstract— **Breast Cancer is one of the most common disease that is responsible for high number of women's deaths every year. Despite the fact that cancer is treatable and healable in earliest stages, the huge number of patients are examined with cancer very late. Data mining process and classification are an efficient way to categorise the data particularly in medical fields, where those approaches are broadly used in diagnosis to make decision. This paper presents a performance comparison among the classifiers: Decision tree classifier (J4.8, Simple CART), Bayes classifier (NaiveBayes, Bayesian LogisticRegression) . The Wisconsin Breast Cancer(original) dataset is used here and is taken from UCI Machine learning Repository. The main goal is to classify data of both the algorithms in terms of accuracy. Our experimental result shows that among all the classifiers, decision tree classifier i.e. Simple CART (98.13%) gives higher accuracy.**

Keywords*— Breast Cancer, Data mining, Wisconsin breast cancer(WBCO), Naive Bayes, Bayesian Logistic Regression, Simple CART, J48.*

I. INTRODUCTION

Breast Cancer is second dominant reason of death among women(after lung cancer).Cancers are of two types i.e malignant and benign .If the cancer is benign and is diagnosed early ,then treatment of patients can be done. It is considered as one of the most prevailing cancer in women with 700,000 deaths happening universally and about 1.5 million cases occur worldwide yearly [1]. Currently , in the US(United States),almost one in eight women has a risk of breast cancer and its percentage is highest in age between 35-50 years. The study of the most latest data has shown that after 6 years of diagnosis, the survival rate is 87.8% and 80.3% after 12 years of diagnosis [2].The nations like Australia ,US Europe (western countries) confirm the high percentage rate of breast cancer patients. In several countries, its percentage expanded during 20th century which leads to rise in mammography and is widely affecting the overall modification in reproductive patterns[3]. For cancer care, ICT(Information and Communication Technologies) plays a vital role. Data Mining techniques, for example, are adapted by medical science because of great performance in outcome prediction, decreasing rates of medicines, supporting patient's health, generating decisions(real time) for people's lives and developing healthcare quality.

There are several techniques for prediction and classification of breast cancer. This paper presents the comparison between the accuracy of four classifiers: J4.8, Simple CART, NaiveBayes, Bayesian LogisticRegression. These are very leading data mining algorithms for research and comes under top ten algorithms of data mining.

The objective of this paper is to identify which classifier gives the most accurate result for WBCO dataset of breast cancer and for this we have used WEKA tool.

II. LITERATURE REVIEW

The authors[4] compared multi-layer perception(MLP), Decision tree(J48) and Sequential Minimal Optimization (SMO). Three dataset has been used i.e WBC, WDBC, WPBC on which 10-fold cross validation method is applied. Results shows that fusion of SMO, MLP, J48 is higher than other classifier in WPBC datast and the fusion of MLP and SMO is better in WDBC dataset.

In the paper [5], authors compared four free source data mining tools that is WEKA, Rapidminer , KNIME,Orange so hat analyst can use that single tool for rapid and good quality results. The authors took total 8 datasets and breast cancer was one among those datsets. Experiments were done on different tools and for Breast cancer, the best resulted from WEKA tool and decision tree with 85.96% accuracy.

S. Aruna, Dr. S.P. Rajagopalan and L.V. Nandakishore [6] compared Naive Bayes, uSpport Vector Machine, Radial Basis Neural Network,J48 and simple CART using WEKA. The authors used multiclass and binary dataset and compared all these classifiers on the basis of precision, sensitivity and specificity. Results shows that SVM-RBF kernel has high performance percentage when compared to other classifiers.

Zehra karapinar Senturk and Resul Kara[7] in their paper diagnosed breast cancer illnesss.The authors have collected the data of the patients who are in danger with the cancer. Predictions have been made through seven algorithms and their accuracies is found. For the prediction, RapidMiner tool is used with algorithms and found SVM algorithm is best suitable for predicting the cancer.

Subrata kumar Mandal [8] found the cancer features which can accurately classify the breast cancer as malignant and benign. Comparative study of various approaches of classifiers i.e. Naive Bayes(NB), Logistic Regression (LR) , Decision Tree on the basis of time complexity and accuracy. Experiments were conducted by the author and found that the Logistic Regression (LR) has maximum accuracy with minimum time complexity.

978-1-5386-6947-1/18/$31.00 ©2018 IEEE

III. DATASET DESCRIPTION

In this paper, we have taken WBCO (Wisconsin breast cancer original) dataset from UCI repository[9]. The dataset contains 697 instances out of which 456 are Malignant and 241 are Benign . There are 16 instances which has single missing attribute value and is denoted by a question mark '?'.

The information of the attribute in the dataset are as follows:

TABLE I : WBCO DATASET ATTRIBUTES

S.No	Attribute	Domain
1	SCN	Id no.
2	CT	1-10
3	Unif. of CS	1-10
4	Unif. of CS	1-10
5	M. Adh.	1-10
6	SECS	1-10
7	Bar Nuc.	1-10
8	Bland Chr.	1-10
9	NN	1-10
10	Mito.	1-10
11	Class Label	YES, NO

The description of the attributes are shown below in the table:

TABLE II : ATTRIBUTE DESCRIPTION

S.NO	Attributes	Description
1	SCN- Sample code number	ID Number
2	CT- Clump Thickness	Arranged in monolayers
3	Uniformity of Cell Size	Cancer units vary in size
4	Uniformity of Cell Shape	Cancer units vary in shape
5	Marginal Adhesion	Normal units stick together but cancer cells do not stick.
6	SECS - Single Epithelial Cell Size	Large cells are malignant cells
7	Bare Nuclei	Cancer nucleus are more in number
8	Bland Chromatin	Swollen in cancer units
9	NN - Normal Nucleoli	Cancer cell nucleus is rough.
10	Mitosis	Cell duplicates and divide
11	Class Label	YES for Benign and NO for Malignant.

The above parameters are very important for describing the nature of cells i.e whether it is cancerous cell or not.

IV. METHODOLOGY

In today's world, the databases are extremely vulnerable to missing, noisy and uncertain data because of their huge size and are originated from various misscellaneous sources. Therefore, pre-processing is very important phase which is followed by classification.

The pre-processing of dataset is done manually by taking the median of individual attribute values and replacing the missing values by that median value. In WBCO dataset, attribute bar nuclei contains all the 16 missing values. So it was very easy to preprocess it manually. All the missing values were replaced by the median. Now the dataset is free from all the missing values.

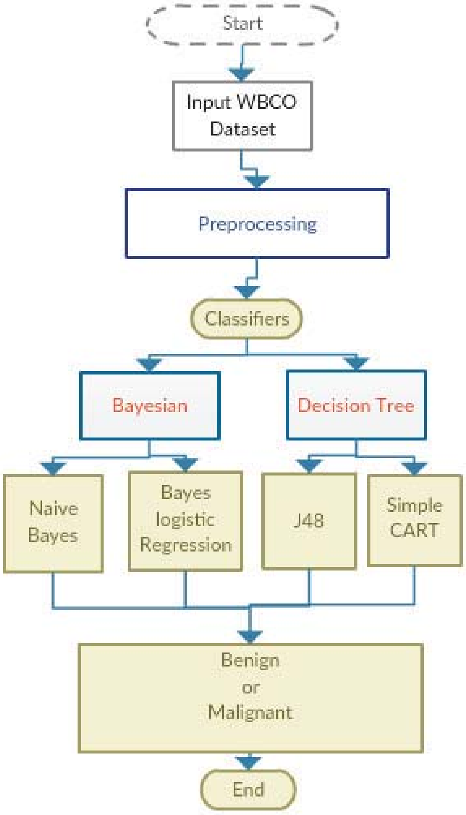

Fig. 1: Workflow diagram for breast cancer prediction

In this paper, we have considered four data mining algorithms: Bayes classifier (Naive Bayes, Bayesian Logistic Regression), Decision Tree (J48, simple CART). We have used these classifiers to analyze dataset of breast cancer. We have selected these classifiers and by using WEKA tool, we will find out that which of these classifier will show the most accurate result for breast cancer.

A. Naïve Bayes And Bayesian Logistic Regression

The Naive Bayes algorithm is based on the Bayesian approach, which is fast ,simple and clear classifier. It is called 'Naive' because it consider attributes that are mutually independent. But in practice, it is not true but can be obtainable by removing the dependent attributes in preprocessing step[10].

Bayes Logistic Regression also follows Bayesian approach. It is the technique which is taken from the machine learning. It is for the problems which have two class values and in this paper the WBCO dataset contains two class labels i.e. benign and malignant. It is also very efficient bayes classifier for those databases which uses two or more class values.

B. Decision Trees J48 And CART

Simple CART is abbreviated as Classification and regression tree. It is well known and widely used methodology[11]. This classifier is used for prediction. These classifiers uses historical data and forms binary decision tree. It has the ability to tackle the missing values in the dataset by just ignoring them[12]. In its implementation, the dataset is divided into 2 subgroups that are different from outcome and it is continued until minimum size subgroup is obtained.

Decision tree J48 is used for building unpruned or pruned tree from labelled training datasets. In this decision tree algorithm, every attribute of dataset can make decision by just splitting datasets into very small subsets. The splitting stops when all the instances which are in the subsets belong to same or a single class. J48 also creates decision node in the tree by guessing the expected value of the class.

V. EXPERIMENTAL RESULTS AND COMPARISONS

In this paper we have done a broad study on different classifiers and compared the accuracy and execution time(time complexity) of each classifier which is obtained by applying dataset on WEKA tool.

A. Performance Criteria

Performance involves accuracy which is defined as the percentage of correct predictions. It is calculated using the formula :

Accuracy = true_negative + true_positive/ true_negative + true_positive + false_positive + false_negative.

Below table shows the classification accuracies of different classifiers :

TABLE III : ACCURACY PERCENTAGE FOR WBCO DATASET

S.No	Algorithm	Classification Accuracy(%)
1	Naive Bayes	95.2654
2	Bayesian logistic Regression	65.4232
3	Simple CART	98.1349
4	J48	97.274

TABLE IV : EXECUTION TIME OF ALGORITHMS

S.No	Algorithm	Execution Time(seconds)
1	Naive Bayes	0.02
2	Bayesian logistic Regression	0.06
3	Simple CART	0.29
4	J48	0.20

In Table IV, we got the execution times of all the four classification algorithms. The execution time for Naive Bayes , Bayesian logistic regression, Simple CART and J48 are O.O2,0.06,0.29,0.20 respectively.

TABLE V : ERROR COMPARISON OF ALGORITHMS

	Naive Bayes	Bayesian logistic Regression	J48	Simple CART
Kappa statitics	0.8967	0.3458	0.9396	0.9588
Mean Absolute error	0.0467	0.588	0.0484	0.0351
Root mean squared error	0.2125	76.4024	0.1556	0.1325
Relative absolute error(%)	10.309	76.4024	10.6954	7.7604
Root relative squared error(%)	44.6759	123.6328	32.7087	27.8617

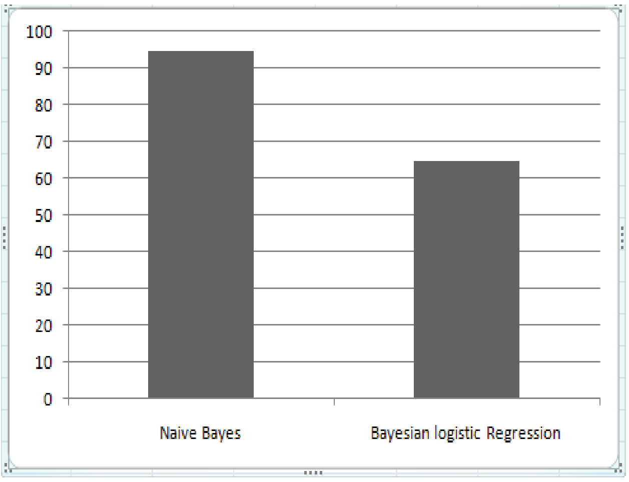

Fig. 2: Comparison of accuracies between bayesian classifier

The above figure 2 shows that among the Bayesian classifier, Naive Bayes has the best accuracy of 95.26% when compared with Bayesian logistic regression.

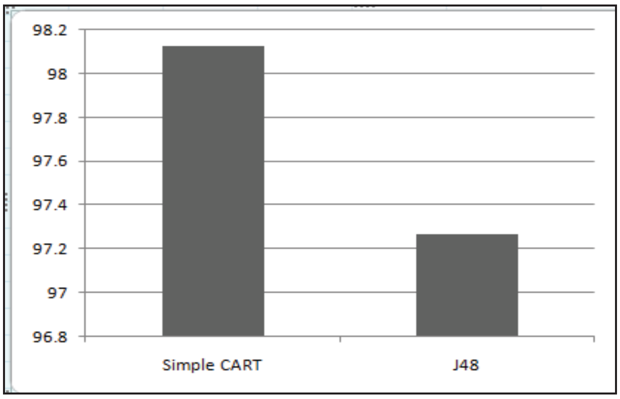

Fig. 3: Comparison of accuracies between decision tree classifier

The above figure 3 shows that among decision tree classifier, Simple CART has the best accuracy of 98.13% when compared with J48 classifier.

VI. CONCLUSION

In this paper, we have compared accuracy and time complexity of different classifier like Naive Bayes , Bayesian Logistic Regression, Simple CART ,J48. The objective of this research is to find which is the best classifier among all. Classifier used here classifies dataset of breast cancer namely WBCO. All the experiments were done in WEKA. The results obtained are compared and it is found that Simple CART decision tree algorithm is the best classifier in terms of accuracy among all the classifiers used here in the research. The time complexity of simple CART is more. So in future we can work on that so more accurate results can be produced in less time.

REFERENCES

[1] Lyon IAfRoC: World Cancer Report. *International Agency for Research on Cancer Press* 2003:188-193.

[2] American Cancer Society, 2007. *Breast Cancer Facts & Figures*. American Cancer Society.

[3] Daniel F. Roses (2005). Clinical Assessment of Breast Cancer and Benign Breast Disease, In: *Breast Cancer*: Vol. 2, Ch. 14, M. N. Harris [editor], Churchill Livingstone, Philadelphia.

[4] Salama, Gouda I., M. Abdelhalim, and Magdy Abd-elghany Zeid. "Breast cancer diagnosis on three different datasets using multi-classifiers." *Breast Cancer (WDBC)* 32, no. 569 (2012): 2.

[5] Borges, Luís C., Viriato M. Marques, and Jorge Bernardino. "Comparison of data mining techniques and tools for data classification." In *Proceedings of the International C* Conference on Computer Science and Software Engineering*, pp. 113-116. ACM, 2013.

[6] Aruna, S., S. P. Rajagopalan, and L. V. Nandakishore. "Knowledge based analysis of various statistical tools in detecting breast cancer." *Computer Science & Information Technology* 2 (2011): 37-45.

[7] Senturk, Zehra Karapinar, and Resul Kara. "Breast cancer diagnosis via data mining: performance analysis of seven different algorithms." *Computer Science & Engineering* 4, no. 1 (2014): 35.

[8] Mandal, Subrata Kumar. "Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree." *International Journal Of Engineering And Computer Science* 6, no. 2 (2017).

[9] UCI Machine Learning Repository and breast cancer site:archive.ics.uci.edu/ml site:ics.uci.edu.

[10] Witten, Ian H., Eibe Frank, Mark A. Hall, and Christopher J. Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[11] L. Breiman, J. Friedman., R. Olshen, C. Stone, (1984), *Classification and Regression Trees*Wadsworth, Belmont, CA.

[12] Kalmegh, Sushilkumar. "Analysis of WEKA data mining algorithm REPTree, Simple CART and RandomTree for classification of Indian news." *Int. J. Innov. Sci. Eng. Technol*2, no. 2 (2015): 438-446.