Study and Analysis of Breast Cancer Data

Sonali Nandish Manoli,
Sri Jayachamarajendra College of Engineering,
Mysuru

Padma S.K
Sri Jayachamarajendra College of Engineering,
Mysuru

Abstract: **Breast cancer develops from breast tissue when cells in the region grow out of control. Signs of Breast Cancer may include a lump in the breast, a change in the breast shape, dimpling of skin, fluid coming from the nipple or a red scaly patch of the skin. The objective of this paper is to find the smallest set of features which can be used from the available Wisconsin Breast Cancer (WBC) Data set using supervised learning methods to detect breast cancer. For classification, the Breast cancer data is classified using Naive Bayes Classifier and Support Vector Machine (SVM) Classifier. Further the Principal Component Analysis (PCA), a Dimensionality Reduction technique (DRT) is used to obtain the smallest subset of features to get better performance measures to classify the data as either benign or malignant.**

Keywords **- Support Vector Machine, Naive Bayes, Principal Component Analysis, Wisconsin Breast cancer data set.**

I. INTRODUCTION

Various techniques are being used to detect cancer at an early stage. Among the use of various techniques, Supervised Machine learning is the most popular learning method used in cancer diagnosis [4].The data set used is Wisconsin Breast cancer (WBC) original data set which is publicly available in the UCI machine learning repository. The dataset involves recordings from a Fine Needle Aspirate

(FNA) Test [6].In this research paper, analysis is done by using the original data with all the features where the missing value of attribute has been obtained by taking the mean of the other values of the attribute. Two classifiers namely the Naive-Bayes and the SVM are used to analyse the data. Further feature extraction principle is used to eliminate the redundant features of the data. This is done by reducing the dimensionality of the data using Principal Component Analysis. The reduced feature data is again classified using Naives-Bayes and the SVM classifiers to improve the Accuracy of the data prediction

II. LITERATURE SURVEY

In the paper [1], by Sadhana et.al there is a proposal to compare the accuracies of two classifiers namely the SVM and Decision Tree (DT) for WBC by using accuracy indicator to evaluate classification efficiency of different classification algorithms. Overall, DT classification accuracy was found to be better than other classifier namely the SVM. They could obtain an accuracy of 94.54%.In the paper [2], by K. Sivakami; breast cancer prediction was done using the DT-SVM Hybrid Model. This study was performed on (WBC) dataset taken from the UCI machine learning repository. It had nine different attributes which varied significantly between benign and malignant samples. The accuracy obtained was 91% with an error rate of 2.58%.For IBL, the accuracy obtained was 85.36% with an error rate of 12.63%.For SMO the accuracy was 72.56% with an error rate of 5.96%.For Naive Bayes the accuracy obtained was 89.48% with an error rate of 9.89%.So this study revealed that DT-SVM hybrid model gave good accuracy. In the paper [3] by LeenaVig, experimental results were obtained for 3 performance measures namely accuracy, sensitivity, and specificity. Many classifiers like Artificial Neural Networks, SVM, Naive-Bayes and a Random Classifier with 100 decision trees were performed on the (WBC) data set and results were obtained.

The best accuracy obtained was 95.64% for Random Forest. The sensitivity obtained was 0.97 and specificity obtained was 0.94.In the paper [4] by Animesh Hazra et.al, the analysis of the data has been performed by using Naive Bayes, SVM classifier and the ensemble classifier. The result obtained was the best using Naive Bayes approach where the accuracy obtained by considering only 5 features out of the 32 features in the Wisconsin Breast Cancer Diagnosis (WBCD) Data set was 97.3978%.In the paper [5] by Kathija.A et.al ,the analysis has been made on WBC data set by using the Naive bayes and the SVM classifier and performance measures such as accuracy, sensitivity and specificity have been performed using 10-fold cross-validation technique. The best accuracy obtained was by using the Naive Bayes Classifier with an accuracy of 95.65%.

III. MOTIVATION

A lot of research is being done in the health-care sector to give better treatment to patients. Analysis of the data obtained from the patient plays a vital role in treatment and improvement of patient's health. Breast Cancer has become the leading cause of death in women, it is estimated that 13.4% of the women born today will be diagnosed with cancer at some stage in their lives[2].The breast is made up of lobes containing 15 to 20 sections and ducts. The most common type of breast cancer begins in the cells of these ducts. Cancer that starts in the lobes or lobules found in both the breasts are other types of breast cancer [4].In the domain of Breast Cancer data analysis a lot of research has been done in the domain of relatively high predictive classification. Hence there is a need to develop a system which helps in predicting the data better for early detection of the type of

tumour. It is classified into two categories namely Benign which is the non-cancerous tumour and Malignant which is cancerous tumour.

IV. PROPOSED METHODOLOGY

In the study, the Wisconsin Original Breast Cancer Dataset with 699 samples has been considered. The data set has 16 missing values in the bare nuclei attribute. To eliminate the missing values imputation has been performed on the 16 missing values by considering the mean of the 1st nearest neighbour above and below the missing value in the same attribute. The resultant data set is the whole data set considered for the study. It is divided into training and testing data. In order to measure the performance, 80% or 559 samples are randomly chosen as training data and the remaining 20% or 140 samples are chosen as testing data. This method is applied 10 times by randomly taking 20% data in each iteration which is tested each time. The average accuracy is observed for 10 iterations for each classifier used.

The Naive-Bayes classifier is a simple probabilistic classifier based on the Bayes' Theorem with strong independent assumptions between features. The use of Naive Bayes is that it is easy to train, fast to classify and it can handle real and discrete data very well during classification. It is suited for the WBC data set considered since it has discrete values. The classifier predicts the class membership probability such that the probability of a given tuple falls into either the Benign or Malignant class.

The Support Vector Machine (SVM) classifier uses hyper planes to separate instances of various classes. The data is mapped to a higher dimensional space where it can be linearly separated using kernel trick. The optimal separator is a line which can efficiently separate the given input data into two classes namely Benign and Malignant by giving maximum margin between the two classes. Further, steps have been taken to obtain the accuracy equal to or greater than the results obtained by using the above mentioned classifiers. It is done by using the feature extraction method .The feature extraction methodology used here is the Principal Component Analysis (PCA) which gives the principle components of a data by reducing the Dimensionality of the original Data set. Several Iterations have been conducted by considering the principal components in the decreasing order of the result of the PCA obtained for the WBC dataset because decreasing order indicates maximum variance of obtained feature. Such features cannot be eliminated or ignored to get good accuracy for prediction of results. The study has been conducted by considering values of data obtained from the 1st principal component which results in a 699x1 Matrix for the PCA reduced dataset, the 1st principal component has the highest variance. It is followed by considering the 1st two principal components which have the highest and the second highest variance which results in a 699x2 matrix and so on till the 9th principal component which results in data having 699x9 matrix of PCA reduced data set. The PCA reduced data has been divided into training data and testing data. The reduced data is again classified using the Naive Bayes and the SVM approach. The workflow of the Breast Cancer Detection for the PCA reduced data is as shown in Figure 1. The classification is evaluated by using performance measures such as Accuracy, Confusion Matrix, Precision, Recall, and Specificity.

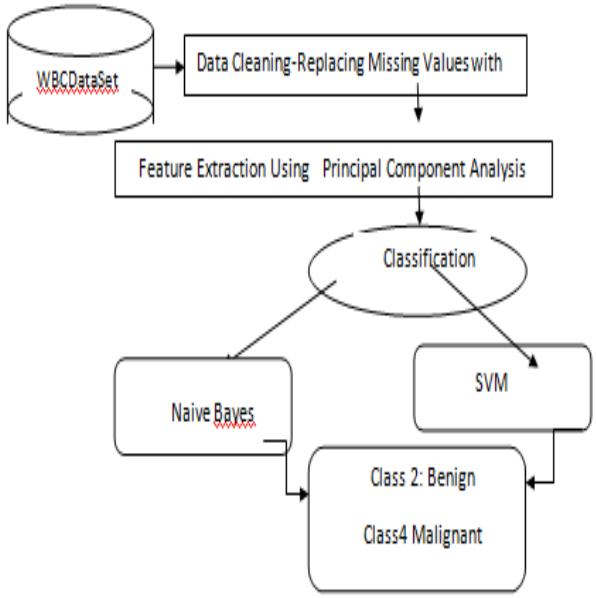

Figure 2 : Workflow Diagram for breast cancer cell detection using PCA

V. PERFORMANCE EVALUATION

The measures are calculated using TP and TN which are true positive and negative tuples classified by classifiers. FP and FN are positive and negative tuples which are incorrectly classified.

1. Confusion Matrix

For Naive Bayes Classifier considering whole data result is as shown in Table I

Actual Class	Predicted class	
	Is Benign	Is Malignant
Is Benign	78	3
Is malignant	3	56
Total =140	81	59

Table I

For Naive Bayes Classifier considering PCA reduced data result is as shown in Table II

Actual Class	Predicted class	
	Is Benign	Is Malignant
Is Benign	92	1
Is malignant	3	44
Total =140	95	45

Table II

For SVM Classifier considering whole data result is as shown in Table III

Actual Class	Predicted class	
	Is Benign	Is Malignant
Is Benign	79	3
Is malignant	2	56
Total =140	81	59

Table III

For SVM Classifier for PCA reduced data result is as shown in Table IV

Actual Class	Predicted class	
	Is Benign	Is Malignant
Is Benign	93	1
Is malignant	2	44
Total =140	95	45

Table IV

VI. RESULT AND DISCUSSIONS

The accuracy of classification obtained is 95.71 % for Naive Bayes and 97.14% for SVM for the whole data. Further accuracy of classification obtained is 97.14% for Naive Bayes and 97.86% for SVM for PCA reduced data by considering only 2 features of the PCA reduced data set. So this paper reveals that the SVM classifier is a better classifier which provides an accuracy of 97.86% which is 1.96 % more than the accuracy of 95.90% reported as the highest in the UCI Machine Learning Repository

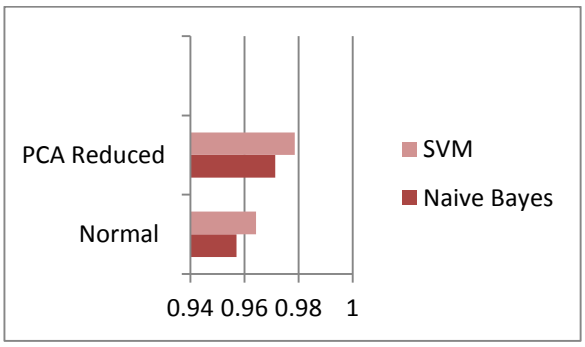

Figure 2. ACCURACY FOR TESTING SAMPLES

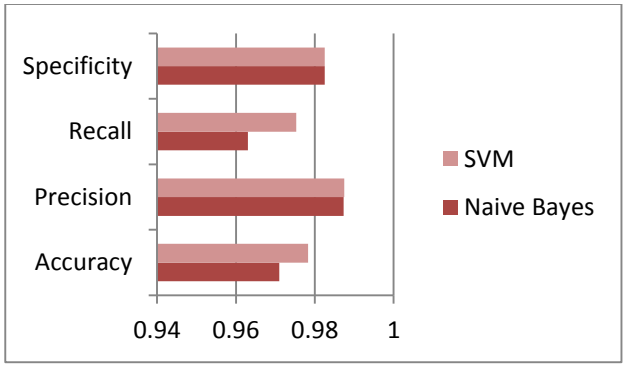

Figure 3. RESULTS OF ALL PERFORMANCE MEASURES FOR WHOLE DATA

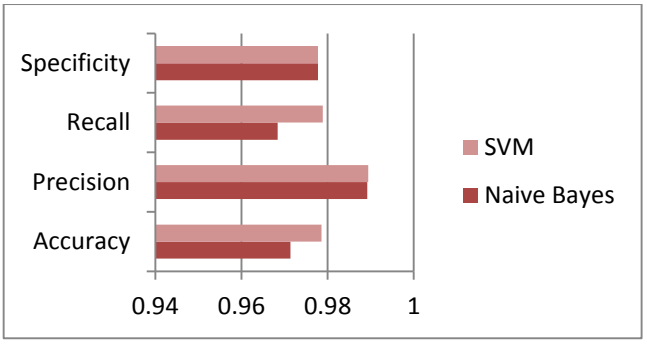

Figure 4 .RESULTS OF ALL PERFORMANCE MEASURES FOR PCA REDUCED DATA

VII. CONCLUSION

From the analysis we can conclude that the model is useful in predicting breast cancer from tumour data, there is also scope for analysis using other classifiers and dimensionality reduction techniques which may help in better understanding of larger data sets with many more features in near future. Further work is in progress to develop classifiers using WBC Diagnostic and WBC Prognostic data which will help in the early detection of breast cancer in patients so that early treatment will help in improving the lifespan of patients.

REFERENCES

[1] Sadhana and Sankareswari, "A Proportional Learning Of Classifiers Using Breast Cancer Datasets", International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 3, Issue. 11, pg.223 – 232, November 2014.

[2] LeenaVig, "Comparative Analysis of Different Classifiers for the Wisconsin Breast Cancer Dataset", Open Access Library Journal, Volume 1e660, 2014.

[3] Sivakami, "Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model." International Journal of Scientific Engineering and Applied Science (IJSEAS) -Volume-1, Issue-5, ISSN: 2395-3470, August 2015.

[4] Animesh Hazra et.al, "Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms", International Journal of Computer Applications (0975 – 8887) Volume 145 – No.2, July 2016.

[5] Kathija.A and Shajun Nisha, "Breast Cancer Data Classification Using SVM and Naive Bayes Techniques" International Journal of Innovative Research in Computer and Communication Engineering.Vol.4,Issue12,December 2016.

[6] Wisconsin Original Breast Cancer Dataset. http://archive.ics.uci.edu/ml/datasets/breast + cancer + Wisconsin+original.