JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Selected Feature Based Improvements in Breast Cancer Detection

Ranjeet Kumar
Department of Electronics and Communication
Sarala Birla University
Ranchi, India

Vijay Kumar Singh
Department of Physics
Sarala Birla University
Ranchi, India

Arvind Kumar
Department of Electronics and Communication
BIT Sindri
Dhanbad, India

Abstract— The death due to breast cancer is increasing day by day. Survival rate can be increased by modern treatment methods, if cancer is detected at early stage with high accuracy. Several parameters of patients are stored in cancer dataset and all parameters are not equally important to predict cancer. In this article, different classifier indicators are compared with and without reduced attributes of Wisconsin Breast Cancer (WBC) dataset, by applying a feature selection algorithm. It was found that reduced attribute cancer dataset has better accuracy than original cancer dataset

Keywords—Breast cancer, feature selection, machine learning classifier

I. INTRODUCTION

It is reported that breast cancer is leading cause mortality rate in women globally. To identify the reasons to develop the cancerous breast cell is not an easy task. Cancerous tumor are of two main types, benign and malignant. Malignant tumor tendency to grow in uncontrolled fashion lead to death. Scientists are searching new technologies and drug to combat over cancer. By the help of more accurate early diagnosis, survival rate can be increased [1].

Researchers have been screening breast cancer by conventional methods like mammography, biopsy etc. In mammography method role of a radiologist has more importance. Different radiologist may have different interpretation. Biopsy is far better option than mammography, but it is insidious and costly [2, 3].

The huge past breast cancer data can be used to detect early stage with the help of data mining and machine learning. Different machine learning classifier algorithm are used to detect malignant and benign. Before applying machine learning algorithm data preprocessing is done .All the information attributes are not play equally important role in prediction of cancer. To enhance the classifier accuracy, feature selection method can be used [4].

In this article, some important parameters of classifier accuracy values are compared for feature selected cancer data and without feature selected cancer data. Naive Bayes classifier is used in this research. Best First search algorithm has been deployed for feature selection [5] .

.

II. LITERATURE REVIEW

In 2014 [6] the authors investigated the performance of different classification techniques. They analyzed the WBC dataset and found SMO (sequential minimal optimization) had a higher prediction accuracy (96.21%) than the tree and K nearest classifier.

In 2019 [7] the authors discussed two popular machine learning techniques for WBC dataset, artificial neural network and support vector machine(SVM) with data mining tool.SVM was found better than ANN(Artificial Neural Network) with performance accuracy of 96.9957%.

In 2019 the authors in [8] compared performance on three classifier J48,REP Tree and Naive Bayes. They foundJ48 had best performance in terms of accuracy.

In 2020 [9] the authors applied preprocessing techniques on breast cancer dataset. They applied 13 different classification algorithms and found result accuracy range between 72% to 98%.They suggested deep learning for high accuracy for future work.

In 2021 [1] Ajay Sharma purposed a method to improve early breast cancer detection accuracy upto 99.41%. They used three features selection methods correlation based, information gain based and sequential based methods. Different classifiers were applied on these feature subsets and best feature subset was selected.

In 2021 [10] authors found better result by applying MLP on more important attributes of WBC . Different feature selection algorithm was ranked by Extra Tree ensembles method.

in 2022 [11] this study evaluated the performance of, Logistic Regression, Naive Bayes, Multilayer Perceptrons, and Support Vector Machines using data from the Breast Cancer Surveillance Consortium (BCSC), which included 154,899 screening records, Out of the three, they discovered that the Multilayer Perceptron performed the best.

In 2022 [2] the authors applied wrapper based feature selection on WBC dataset. LR,LSVM and QSVM were applied for classification .QSVM with accuracy 97.1% was found best among other classifiers.

III. PROPOSED METHODOLOGY

The proposed method has following steps:

Step1: Apply a machine learning classifier on Wisconsin Breast Cancer (WBC) dataset.

Step2:Note down the parameters for classifier accuracy.

Step3:Apply feature selection algorithm on WBC dataset and find reduced attributes dataset.

Step4: Apply same machine learning algorithm to reduced attributes dataset that was applied on WBC dataset.

Step5: Compare the parameters for classification accuracy between selected feature and original WBC dataset.

A. Dataset description

Wisconsin breast cancer (WBC) dataset has 30 features and 569 instances. Features are computed from a digitized image of fine niddle aspiration (FNA) of a suspicious or abnormal breast mass. In the borderline region of a set of cell nuclei, active contour shapes, also called snakes are initialized. By deforming the customized snakes to original shape of nuclei, helps to examine the nuclei shape, size and texture. 569 different possible combinations of digitized images of FNA were examined to best classify between benign and malignant types of tumor [12] .

B. Bayesian classifier

Naive Bayes is a supervised machine learning probabilistic based classifier ,which is conditionally independent to any other features in the model dataset. It is assumed that all features equally contribute to produce outcome. Mathematically ,it can be presented as

$$PE(F_1 \ldots\ldots F_n) = \frac{PE(C)PE(F_1 \ \ldots . F_n/C)}{PE(F_1 \ldots\ldots F_n)}$$

Where, PE is the probability, C is the class variable and $F_1 \ \ldots . F_n$ are features variables.

C. Performance Parameters

In this article, several confusion matrix parameters are used to compare results of cancer dataset with and without attributes selection. Eight parameters are used in this research.

True Positive(TP)

The observation is predicted positive and is actually positive.

False positive(FP)

The observation is predicted positive and is actually negative.

Precision

It is ratio of true positive to total number of positive prediction by model.

Recall

It is the ratio of true positive to sum of true positive and false positive.

F-Measure

Traditional F-measure can be calculated with combination of precision and recall with formula

$$F = \frac{2 \times precision \times recall}{precision + reacll} \quad [12]$$

Matthew's Correlation Coefficient(MCC)

It is a correlation coefficient between actual and predicted series based on TP ,TN, FP and FN. It returns value from $-1 \ to + 1$. [13]

Receiver Operating Characteristics(ROC)

It is the plot of false positive rate on X-axis and true positive rate on Y-axis [12].

Kappa Statistics

It compares between observed accuracy with an expected accuracy [10].

$$Kappa\ Statistics = \frac{Observed\ Accuracy - Expected\ Accuracy}{1 - Expected\ Accuracy}$$

IV. EXPERIMENTAL RESULTS

A. Naive Bayes classifier apply on WBC dataset

When Naive Bayes classifier is applied on Wisconsin Breast Cancer(WBC) dataset, following results has been noted down .

Table I. Values of confusion matrix parameters, when Naive Bayes classifier applied on WBC dataset.

Parameters	Value
Kappa Statistics	0.8418
TP Rate	0.926
FP Rate	0.086
Precision	0.926
Recall	0.926
F-Measure	0.926
MCC	0.842
ROC	0.976

B. Naive Bayes Classifier apply on reduced features of WBC dataset

The Best First search algorithm was applied for features selection. Table II shows the values obtained by different confusion matrix parameters by applying Naïve Bayes classifier on selected or reduced features of WBC dataset.

Table II. Values of confusion matrix parameters, when Naïve Bayes classifier applied on reduced WBC dataset.

Parameters	Value
Kappa Statistics	0.8715
TP Rate	0.940
FP Rate	0.074
Precision	0.940
Recall	0.940
F-Measure	0.940
MCC	0.872
ROC	0.980

C. Comparison of different classifier indicators

Table III shows the comparison of different confusion matrix parameters between WBC dataset with (or purposed method) and without selected features, when applied Naive Bayes Classifier.

Table III. Comparison of Confusion matrix parameters

Parameters	**Without Feature Selection**	**With Feature Selection**
Kappa Statistics	0.8418	0.8715
TP Rate	0.926	0.940
FP Rate	0.086	0.074
Precision	0.926	0.940
Recall	0.926	0.940
F-Measure	0.926	0.940
MCC	0.842	0.872
ROC	0.976	0.980

D. **Comparison chart**

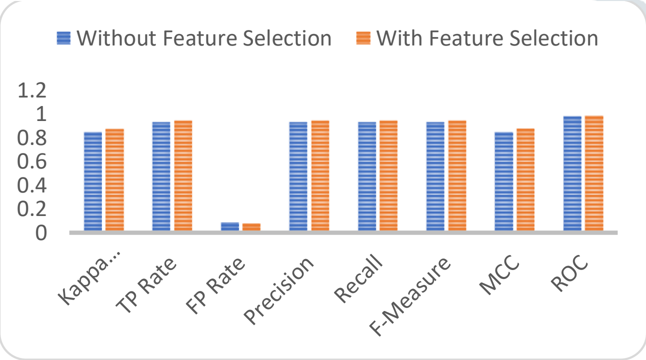

Fig 2. Comparison of confusion matrix parametrs

Fig 2 shows comparison of 8 confusion matrix parameter indicators taken for experiment. For high classification accuracy, except FP(False Positive) rate, all other 7 parametrs values should be high. It is observed that the value of FP decreases from 8.6% to 7.4%, hence Precision increased in selected features dataset. Except FP rate, other 7 confusion matrix parametrs vales is higher for selected features dataset eg. TP rate is increased from 9.26% to 9.40%.

V. CONCLUSION

Early breast cancer detection with high precision is highly important to increase the survival rate of cancer patients. As all the features of WBC dataset are not equally important for detection of cancer, so feature selection algorithms can removed less important features from original dataset. In this paper, Best First search algorithm was applied on WBC dataset for feature selection. Naive bayes classifier was use to calculate 8 different confusion matrix parameters for selected feature WBC dataset and without feature selected WBC dataset. Experimental results show that breast cancer detection is more précised with feature selection method.

Further, different classifiers with different feature selection algorithms can be applied on WBC dataset and can be compared for the highest accuracy.

REFERENCES

[1] A. Sharma and P. K. Mishra, "Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis," *Int. J. Inf. Technol.*, vol. 14, no. 4, pp. 1949–1960, 2022, doi: 10.1007/s41870-021-00671-5.

[2] R. Hasan and A. S. M. Shafi, "Feature Selection based Breast Cancer Prediction," *Int. J. Image, Graph. Signal Process.*, vol. 15, no. 2, pp. 13–23, 2023, doi: 10.5815/ijigsp.2023.02.02.

[3] S. Tounsi, I. F. Kallel, and M. Kallel, "Breast cancer diagnosis using feature selection techniques," *2022 2nd Int. Conf. Innov. Res. Appl. Sci. Eng. Technol.*, pp. 1–5, 2022, doi: 10.1109/IRASET52964.2022.9738334.

[4] D. Lavanya and D. K. U. Rani, "Analysis of Feature Selection with Classification : Breast Cancer Datasets," *Indian J. Comput. Sci. Eng.*, vol. 2, no. 5, pp. 756–763, 2011, [Online]. Available: http://demo.pohonkeputusan.com/files/ANALYSIS OF FEATURE SELECTION WITH CLASSFICATION BREAST CANCER DATASETS.pdf.

[5] M. R. Wijaya, R. Saptono, and A. Doewes, "The Effect of Best First and Spreadsubsample on Selection of a Feature Wrapper With Naive Bayes Classifier for The Classification of the Ratio of Inpatients," *Sci. J. Informatics*, vol. 3, no. 2, pp. 139–148, 2016, doi: 10.15294/sji.v3i2.7910.

[6] V. Chaurasia and S. Pal, "A Novel Approach for Breast Cancer Detection using Data Mining Techniques," pp. 2456–2465, 2014.

[7] E. A. Bayrak and P. Kırcı, "Comparison of Machine Learning Methods for Breast Cancer Diagnosis," pp. 4–6, 2019.

[8] T. Padhi, "Breast Cancer Analysis Using WEKA," *2019 9th Int. Conf. Cloud Comput. Data Sci. Eng.*, pp. 229–232, 2019.

[9] M. Alshammari and M. Mezher, "A Comparative Analysis of Data Mining Techniques on Breast Cancer Diagnosis Data using WEKA Toolbox," vol. 11, no. 8, pp. 224–229, 2020.

[10] M. Bahrami, "Wise Feature Selection for Breast Cancer Detection from a Clinical Dataset," *2021 28th Natl. 6th Int. Iran. Conf. Biomed. Eng.*, pp. 160–164, 2021, doi: 10.1109/ICBME54433.2021.9750287.

[11] S. Yin, R. Sundararajan, and G. Nanda, "Assessing the Impact of Unbalance in Data on Predicting Breast Cancer Occurrence Using Machine Learning Models," *Acta Sci. Med. Sci.*, vol. 6, no. 2, pp. 159–170, 2022, doi: 10.31080/asms.2022.06.1178.

[12] S. Ray, "Selecting Features for Breast Cancer Analysis and Prediction," 2020.

[13] Y. S. Deshmukh, M. Professional, P. Kumar, R. Karan, and S. K. Singh, "Breast Cancer Detection-Based Feature Optimization Using Firefly Algorithm and Ensemble Classifier," pp. 1048–1054, 2021, doi: 10.1109/ICAIS50930.2021.9395788.

.