International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Ensemble Classification Algorithms for Breast Cancer Prognosis

Nitasha, Rajeev Kumar Bedi, SK Gupta

Abstract: ***Breast Cancer is the second highest reason for the death rate among women as well as men too in world. In this paper, we used Data mining classification algorithms to find the presence of breast cancer whether it is benign or malignant and analysis is done on the basics of accuracy and time taken in build model. The data is collected from WISCONSIN of UCI machine learning Repository, which includes patient's samples. The dataset undergoes different algorithm with and without feature selection.***

Keywords: Ensemble Classification algorithms, Feature Selection techniques, Vote Naive Bayes & J48, Random Forest

I. INTRODUCTION

Data mining techniques are very useful in medical science, where it is difficult for the doctor to predict the disease at the specific time. Data mining is speedy growing in the medical science field. Data mining techniques are not only useful in predicting the disease whereas it is very difficult for the doctor to predict the breast Cancer disease by measuring the patient symptoms [1]. To predict the Breast cancer at earlier stage is very important as to predict later. The large amount of data is generated in the hospitals and laboratory's including the tests and the mammography's reports [2]. To predict the disease these data can be analyzed and processed by the doctors for earlier prediction of the breast cancer disease. Different kind of disease such as cardiovascular, diabetes, Cancer, Asthma etc. can be predicted using data mining [3].

Cancer is the most deadly disease among the other. Human body is the composition of the cells, when the irregular growth of cell starts increasing it leads to the tumor or cancer. According the WHO report, 1.5 million cases are registered in 2010[4].

According to the report of 2012-2014 38% men and women are suffering from the cancer [5]. There are different types of data mining techniques such as classification, clustering, association, regression etc. [6]. To predict the breast cancer disease at earlier stage classification techniques is the best among other to predict them. There are different classifiers such as Adaboost, SVM, Artificial neural network, KNN, J48, etc. which are used by the other researchers but none of them have compared standard classifier with hybrid classifier using vote ensemble classifier. Therefore this paper will explain to fill the gap.

In this paper, Section II contains Literature survey. Section III explains Data Mining Algorithms Section IV explains Experimental setup and Result explained in Section V .Section VI contains the Conclusion and future scope.

II. LITERATURE SURVEY

Akshya Yadav, Imlikumla Jamir, Raj Rajeshwari Jain, Mayank Sohani [7], survey the Comparsion of Machine learning Algorithms for the breast cancer prediction. The Comparsion is made between the SVM, ANN, and KNN and noticed that SVM have high accuracy of 97.2%.

Authors Ahmed Iqbal Pritom, Md. Ahadur Rahman Munshi et al [8] explain the survey of predicting the breast cancer recurrence using effective classification and feature selection used attribute selection method for improving the result of algorithms and shows that navies Bayes and Decision tree have better result based on ROC curve.

Kashish Goyal, Prakiti Sodhi, Preeti Aggrawal and Mukesh Kumar [9] aim to find the cancer status whether it is benign or malignant using predication system Ad boost, SVM and Random forest algorithms and compared on the bases of accuracy which show that reduced dataset Adaboost and Logistic Regression show 97.92% each whereas initial dataset the Random forest show high accuracy then the others.

Author Dono Sara Jacob, Rakhi Viswan, V manju, L PadmaSuresh, Sine Raj [10] gives a survey that compares the classification algorithms and clustering algorithms and stated that the classification algorithms are better than of the clustering algorithms in the breast cancer prediction. It also shows that SVM and C5.0 shows same accuracy.

S.Padmapriya et al [11] present the survey of classification algorithms using weka tool. ID3 and C4.5 classifiers are compared in the Comparsion analyses of the breast cancer prediction; as a result the C4.5 gives more accuracy then that of ID3 algorithm. They used SEER dataset and explained various breast cancer applications.

Author Chintan Shah and Anjali G. Jivani's [12] research shows that Naïve Bayes is more superior to the Decision tree and KNN using different parameters. The result is based on the accuracy and lowest time.

III. DATA MINING ALGORITHMS

A. Hybrid Naïve Bayes and J48

Voting (Voting for the classification and Averaging for the regression techniques) is the concept of prediction of data mining which combines the classification from different multiple models or from the same type of model have different data.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author

Nitasha*, Pursuing M tech CSE from Beant College of Engineering and technology, PTU, Punjab, nitashadalmotra25@gmail.com

Rajeev Kumar Bedi Associate Professor of CSE Dept. in BCET, Gurdaspur Punjab, rajeevbedi12@gmail.com

Dr. SK Gupta Professor of CSE Dept. in BCET Gurdaspur, skgbcetgsp@gmail.com

Naïve Bayes is a classification algorithms based on Bayes Theorem which gives the probability of the event occurring given the probability of another event that has be occurred already[13]. $P(\frac{A}{B}) = \frac{P(\frac{B}{A})P(A)}{P(B)}$, where P (A) is the probability of event before the evidence is seen

Whereas P (A/B) is the probability of event after the evidence is seen.

J48 algorithm is used to generate a decision tree developed by Ross Quinlan. It is used for the classification techniques. The roots and notes are generated according to the data sets and no backtracking is done [14]. Using voting the ensemble of Naïve Bayes and J48 algorithm is done to find the presence of breast cancer.

To apply the ensemble voting with Naive Bayes and J48. Select Classify>Choose>Meta>Vote.

Then under the vote the classifier are added as Naive Bayes and J48 classifiers.

Combination Rule> Average of Probabilities> Ok

Cross -Validation Fold >10

Percentage split> 66% > Start

B. Random Forest

Random Forest is a supervised algorithm which makes the forest with combination of trees. It is an extension of bagging for decision trees which can be used as classifier or regression [15]. As the number of trees increases the accuracy increases or vice versa. The decision trees are constructed using greedy algorithm which selects the best splits points at each step in the algorithm. Fig 1 shows the flowchart of Random forest classifier.

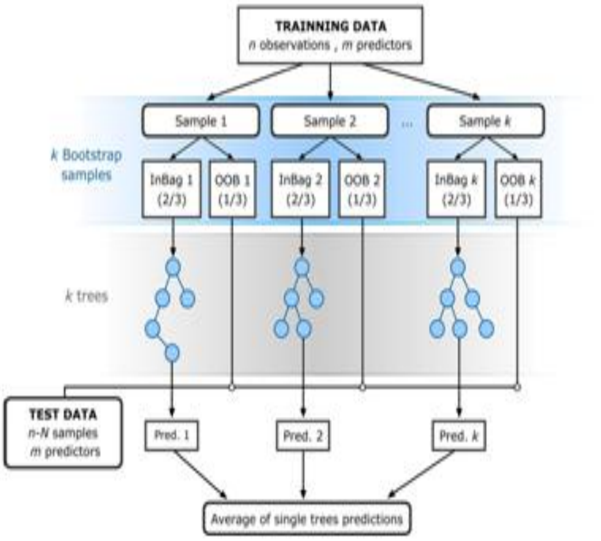

Fig 1 Flowchart of Random Forest Classification Algorithm

IV. EXPERIMENTAL SETUP

The Breast Cancer directory available in UCI Machine Learning Repository contains Four different databases containing datasets related to breast cancer institutions namely University Medical center Ljubljana, University of Wisconsin USA (Original), University of Wisconsin Madison (Prognostic), and Diagnostics [16]. For this paper University of Wisconsin (Original) database is processed and analyzed as it contains 11 attributes and the values are in range of1-10 and no missing values. Attribute Class have value 2 for Benign and 4 for malignant. The first methodology is to preprocess the dataset and standardized the data. The 11 attributes are Sample Code Number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, and Class.

A. Methodology

This research paper explains the applicability of the selected standard methods with the ensemble classifier, and compares the algorithms with the hybrid ensemble classifier (Meta) vote which is a hybrid method of navies Bayes and J48 algorithms. The Comparsion is based on the standard dataset and the reduced datasets by feature selection method of Wrapper Subset Evaluator using Best First search method in WEKA [17].

- ***Data preprocessing***

In this experiment the dataset have no missing values. The data has been normalized and preprocessed by the feature selection. The attributes have value scale from 1-10. In this the normal attributes are tested as well as after the feature selection explained in section 4.1.2, the reduced or effective data attributes are selected and the algorithms are performed under it.

- ***Feature selection***

The number of attributes in dataset are 11 but for better resultant and it is reasonable to remove the irrelevant attribute, which may affect the training time. In this paper the WrapperSubSetEval Attribute Evaluator is used with the Best first Search Method is used. This feature selection method is used both in the ensemble classifier and hybrid classifier vote hybrid of Navies Bayes and J48 classifier which will reduce the attribute according to the algorithms.

V. RESULTS

As the goal of this research paper is to predict the Benign or Malignant condition of the breast cancer in the human being. The classifiers are evaluated in terms of Correctly Classified Instances, Incorrectly Classified Instance and Time taken to build the model shown in Table 1. Table 2 explain the classifier overall performance.

Table 1: Result of the Experiment

Evaluation Criteria	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy (%)	Time Taken (sec)
Random Forest	275	13	95.4861	0.14
Voting (Navies Bayes & J48)	269	19	93.4028	0.03
Feature Selection Random Forest	277	11	96.1806	0.11
Feature Selection Voting	271	17	94.0972	0.00

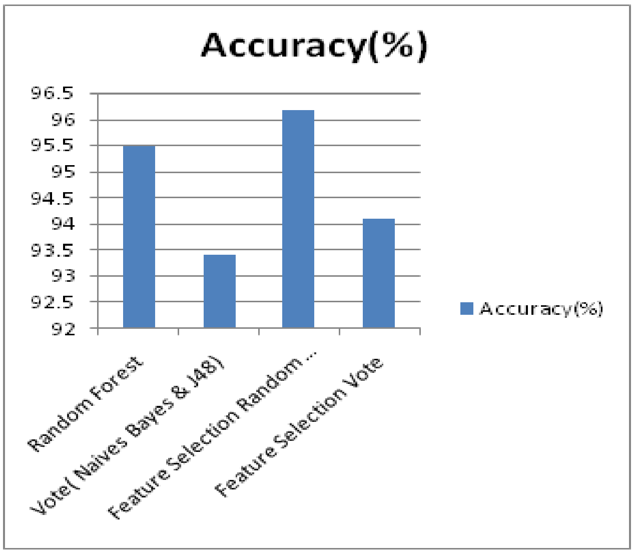

Fig2: Graph for the Accuracy of the Algorithms

From table1 and Fig 2 it is observed that with the feature selection method of Wrapper Sub Set evaluator the Feature selection Random Forest and Vote Hybrid of Navies Bayes and J48 shows better result than the initial dataset with 96.1806% and 94.0972% respectively. As the time taken to build model of random forest is 0.11 and 0 sec for vote whereas the initial dataset have 0.14sec and 0.03 sec respectively. Table 2 represents the overall performance of classifiers.

Table2: Classifier overall performance

Algorithms Performance	TP Rate	FP Rate	Precision	F-measure	ROC Area
Random Forest	0.955	0.044	0.955	0.955	0.980
Vote (Navies Bayes & J48)	0.934	0.066	0.934	0.934	0.957
Feature Selection Random Forest	0.962	0.037	0.962	0.962	0.979
Feature Selection Vote	0.941	0.058	0.941	0.941	0.947

VI. CONCULSION AND FUTURE SCOPE

To detect the benign or malignant condition of breast cancer so that it can prevent at the earlier stage is an important from getting worse. In this paper, different classification algorithms are designed to predict the breast cancer earlier.

WrapperSubSetEvaluator feature selection method and Best First Search method is used to reduce the attributes which are best among the other to predict the cancer. The classification algorithms Random Forest and Vote hybrid of Navies Bayes and J48 are compared with initial dataset means without any feature selection techniques. Random Forest with feature Selection shows better result than the others classification algorithms. The performance is evaluated in terms of precision, recall, F-measure.

REFERENCES

1. Hamza Turabieh, "A Hybrid ANN-GWO algorithm for prediction of cancer disease", American journal of operation research, vol.6, pp.136-146, (2016).
2. Manjula Sanjay Koti and B.H. Alamma, "Predictive Analytics Techniques Using Big data for Healthcare Databases", IEEE, Smart Intelligent Computing application Springer, Singapore vol-105, pp.679-686, (2019).
3. MD .Robel Mia, "A comprehensive study of data mining techniques in healthcare, medical and bioinformatics", International conference on computer, communication, chemical, material and electronic engineering, IEEE pp.1-4, (2018).
4. M Baldonodo, C.C.K. Chang, L.Gravano, A. Paepcke, "The Stanford digital library metadata architecture", International journal Digital library, vol-1, pp.108- 121, (1997).
5. Tawseef Ayoub Shaikh and Rashid Ali, "Applying Machine Learning Algorithms for Early Diagnosis and Prediction of Breast Cancer Risk", 2nd International conference on Communication, Computing and Networking, vol.46, pp.589-598, (2018).
6. K.Kourou, T.P Exarchos, K.P. Exarchos, M.V. karamouzis, D.I. Fotiadis, "Machine Learning Applications in Cancer prognosis and prediction", Computer Structure Biotechnology, vol 13, pp.8-17 (2015).
7. Akshya Yadav, Imlikumla Jamir, Raj Rajeshwari Jain, Mayank Sohani, " Comparative Study of machine Learning Algorithms for Breast Cancer prediction- Review" International journal of Scientific Research in computer Science, Engineering and Information Technology, vol.5, issue2, pp.727-734, (2019).
8. Ahmed Iqbal Pritom, Md. Ahadur Rahman Munshi, Shaded Anzarus Sabab, Shihabuzzaman Shibab, "Survey on Breast cancer prediction using WEKA Tool", Imperial journal of interdisciplinary research, vol-2, issue-4, 2016.
9. Kashish Goyal, Prakriti Sodhi, Preeti Aggarwal and Mukesh Kumar, "Comparative Analysis of Machine Learning Algorithm for Breast cancer prediction", Springer Nature Singapore pte Ltd, International conference on communication, computing and networking, vol-46, pp.727-734, (2019).
10. Dono Sara Jacob, Rakhi Viswan, V Manju, L PadmaSuresh , Shine Raj, "A Survey on Breast Cancer Prediction using Data mining Techniques", IEEE conference on Emerging Devices and Smart system,2-3 March, pp.256-258, (2018).
11. S. Padmapriya, M Devika, V Meena, S.B Dheebika and R. Vinodhini: "A survey on breast cancer analysis using data mining techniques", IEEE, vol-2, issue-4, pp.970-974, (2014).
12. Chitan Shah and Anjali G.Jivani," Comparsion of data mining algorithms classification for the breast cancer prediction", pp.1-4, (2013).
13. M. Navya Sri, J.S.V.S. Hari Priyanka, D. Sailaja and M. Ramakrishna Murthy, "A Comparsion analysis of Breast Cancer Data Set Using Different classification methods", Springer pte Ltd. vol-104, pp.175-181, (2019).
14. Sivangi, P. "Supervised learning approach for breast cancer classification", Int. J. Emerge. Trends Technology computer science, vol-1, Issue-4, pp.125-129 (2012).
15. L. Bremain," Random Forest", Machine Learning, vol- 45, pp-5-32, (2001).
16. Center of Machine Learning and Intelligent System, UCI Repository, University of Wisconsin hospital USA.
17. Motaz M. H Khorshid, Tarek H. M. Abou-El-Enien, Ghada M.A Soliman , "Hybrid classification Algorithms for terrorism prediction in middle and east Africa" , International Journal of Emerging Trends and Technology in computer science, vol-4, issue-3, pp.23-29, (2015).

AUTHORS PROFILE

Nitasha completes BTech Information and Technology from Beant college of Engineering and Technology under Punjab Technical University Punjab. Now pursuing Mtech in CSE Department from Beant College of Engineering and Technology Gurdaspur under PTU.

Paper published on Review on Breast Cancer Prediction using data mining Algorithms in International journal of Computer Science Trends and Technology. Research work in Data Mining.

Rajeev Kumar Bedi Associate Professor in CSE department in Beant College of Engineering and Technology Gurdaspur Punjab. Pursuing PhD in Cloud computing. Research work in cloud computing, Mobile cloud computing, OOPS and Web Development.

Dr. Sunil Kumar Gupta professor in CSE department of Beant college of Engineering and Technology Gurdaspur Punjab. Research work in Checkpoint and cloud computing.