Intelligent Systems and Computer Technology
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC200204

Early Detection of Breast Cancer Using Ensemble Machine Learning Algorithm

Sainikhileaswar.P.S[a,1], Parthasarathy.G[b]
[a] *PG Scholar, School of Computing and IT, REVA University, Karnataka, India*
[b] *Associate Professor, School of Computing and IT, REVA University, Karnataka*

Abstract. As showed by World Health Organization (WHO), Breast cancer growth is the most incessant disease among ladies, 627,000 ladies died due to breast cancer in the year 2018, which implies about 15 present of all cancer deaths among ladies. In order to improve breast cancer results and endurance, early recognition is significant. For detecting breast cancer growth early for the most part AI techniques are used. Right now proposed flexible outfit utilizing ensemble procedure for examining breast cancer using Wisconsin Breast Cancer (WBC) dataset. The purpose of the introduced paper is to consider and explain how different classification algorithms works on our dataset and to identify best algorithm for outfit models, for example, Random Forest Voting Ensemble and XGBoost work and how ensemble systems make the performance of the predictive models better by improving their precision and accuracy.

Keywords. Breast-Cancer Dataset, Ensemble Learning, Bagging, Boosting, Voting, Classification.

1. Introduction

Cancer is the second most cause of death in the world. According to World Health Organization(WHO) 9.6 million people died in 2018 due to cancer Breast, lung, Cervix and colorectal are major type of cancers in women. 627,000 women died due to breast cancer in 2018, which means nearly 15 percent of all cancer deaths among women. Early detection is very essential in order to increase the breast cancer survival. Diagnosing breast cancer is done by tumour classification. Tumours are classified as malignant tumour and benign tumour. A normal breast on left and a breast with cancer on right with arrows One of the most challenging task for physicians is to accurately predict the cancer cells. Unfortunately this tumour classification will take 2 days and not all physicians are specialists in differentiating the tumour cells. The annual cost of medical errors is nearly 17 billion dollars. To overcome these medical errors machine learning can be used which will eliminate human errors in the diagnosis. The amount of data that is being generated in the medical field is tremendously fast. The data that is being generated can be utilized for medical research. We have utilized Wisconsin Breast Cancer (WBC) dataset which is taken from the UCI repository. The part of machine learning is that ,unlike humans brain ML can captures foreseen patterns from data in fraction of seconds. Categories in machine learning are supervised learning, unsupervised learning and reinforced learning. Multi layered preceptron, Decision Tree, k-nearest neighbour (KNN), Logistic Regression, Artificial Neural Network, Random Forest are some of the widely used ML techniques. The paper is arranged as follows: Section2 explains the

[1] Sainikhileaswar.P.S, REVA University, Bangalore-560064, Karnataka, India;
E-mail: nikhilpsv@gmail.com

flow of proposed work. Section3 describes the methodology. The feature selection process is explained in Section4 and section5 explains the model implementation. The paper ends in conclusion.

2. Proposed Work

In the first step of the process we have gathered the Wisconsin breast cancer dataset from UCI repository the collected data is explored using exploratory data analysis by plotting various types of graphs. The data collected is the pre-processed using standardization method. Features are selected using Univariate feature selection technique which uses ANOVA test then several classification algorithms are applied separately to register the accuracy. At the last,we have utilized ensemble techniques to get better accuracy.

3. Methodology

In the presented paper we utilized ensemble machine learning technique for breast cancer diagnosis.Pre-processing ,feature extraction and ensemble model are the three main fundamental sections in this model. In the presented paper we have utilized Wisconsin breast cancer (WBC) dataset. This WBC dataset is taken from the UCI repository .This data set consists of 569 Cases out of which 212 are malignant and 375 are benign and it has 32 attributes.

3.1 Data Pre-Processing

Data Pre-Processing is a data mining method which is basically used to transform raw data into efficient data. Usually real world data can have irrelevant and missing parts(not having all necessary or appropriate parts). Data Pre-processing is the best method for setting such issues. In the presented paper we pre-processed utilizing standardization method. Here in the presented paper we have made various visualizations for better understanding of the data. Exploratory Data Analysis (EDA) is the most important step in pre-processing. EDA gives more insight into the data which is very essential. In EDA we will look into duplicates, count of rows and columns, mean, median, missing values, quantiles, and correlation between variables, data types and distribution of data. For exploratory data analysis several graphs were plotted using Matplotlib which is a visualization library in python used to plot graphs for the better understanding of the data. the plotted for number of malignant and benign cases from dataset. the non-overlapping scatter plot to get the good understanding of values distributed. the relationship between attributes and Violin graph to show the distributive and quantitative information over a level of categorical factors to compare the distribution.

4. Feature Selection

Feature selection is a method which is used to identify the related features from the dataset and remove irrelevant and less important features. Feature selection will have huge impact on the whole performance of the model. Feature selection will reduce over-fitting, reduce training time and improves accuracy. Univariate selections, feature importance, correlation matrix with heat map are some of the feature selection techniques that are easy to use and also provide good results. In the presented paper we have utilized univariate selection for feature selection process.

4.1. Univariate Selection method

The Biggest challenge in ML is to select the best features to train the model. Univariate feature selection can be used to select those features that have the strong relationship with the output variable. Here using univariate selection technique we selected top 20 attributes for the determination of breast cancer. univariate feature selection calculation utilized ANOVA(Analysis of Variance) test for finding detail relation between input variables and the output variable.

5. Implementing The Model

In the execution we utilized ensemble learning strategy which consolidates different models to make the performance of the predictive models better by improving their exactness, ie. Accuracy.

5.1. Ensemble learning

An ensemble is the specialty of joining the decisions from various models to improve the general performance. Ensemble learning which is a powerful system to improve the performance of your AI model. The three most well-known strategies for consolidating the forecasts from various models are bagging, boosting and voting.

5.1.1. Random Forest

Random Forest is an ensemble machine learning algorithm that uses bagging technique. It is an extension of the bagging estimator algorithm. The decision trees in random forest are base estimators.

5.1.2. XGBoost

XGBoost [15] is an ensemble machine learning algorithm, XGBoost means extreme Gradient Boosting and XGBoost is used for both regression and classification problems. XGBoost uses the boosting technique; it builds a strong classifier by combining number of weak classifiers. XGBoost is a regularized boosting technique thus it reduces over-fitting.

Table 1. Classification Report

	Precision	**Recall**	**F1-score**	**support**
Benign	0.98	0.98	0.98	59
Malignant	0.98	0.98	0.98	41
Avg/total	0.98	0.98	0.98	100

5.1.3. Voting

Voting is one of the easiest ways of combining the predictions from several machine learning algorithms. In this method multiple models are used to make predictions then the voting classifier is used to wrap the models by considering each models prediction as a vote.

6. Results And Discussion

In the presented paper we utilized ensemble learning for breast cancer diagnosis using bagging,boosting and voting.Pre-processing,feature selection and ensemble learning are the three main fundamental sections in this model.In the presented paper we have utilized Wisconsin breast cancer (WBC) dataset.This WBC dataset is taken from the UCI repository.This data set consists of 569 Cases out of which 212 are malignant and 375 are benign and it has 32 attributes.The data collected is further pre-processed using standardization method.Since large features will impact on the model implementation we used limited input variables that are more related.Features are selected using Univariate feature selection technique.In this model we selected 20 features Since large features will impact on the model implementation.Then several classification algorithms are applied separately to register the exactness.At the last,we have utilized ensemble methods to increase the performance of the model.The accuracy achieved using ensemble learning is better than the individually achieved accuracy.The accuracy achieved using various classification algorithms and ensemble algorithms are shown in Table1 .

Table 2., Accuracy of algorithms

Algorithm	Accuracy
Logistic	95.0%
KNN Classifier	96.0%
Linear SVC	94.0%
Gaussian Kernel	59.0%
Decision Trees	95.0%
BernoulliNB	59.0%
MultinomialNB	90.0%
Artificial Neural	96.0%
Voting Ensemble	97.0%
Random Forest	97.0%
XGBoost	98.0%

7. Conclusion

Breast cancer is one of the leading cause of cancer deaths among women so early detection of breast cancer is critical. In the presented paper we have used ensemble machine learning algorithm for diagnosis and early detection of breast cancer using

WBC dataset collected from UCI repository. It is seen from the table that the ensemble algorithm achieved 98 percent accuracy.

References

[1] http://www.who.int/cancer/detection/breastcancer/en/ (Accessed January 2020)

[2] E.A.Bayrak,P.Kırcı and T.Ensari, Comparison of Machine LearningMethods for BreastCancer Diagnosis .2019Scientific Meeting onElectrical-Electronics Biomedical Engineering and Computer Science(EBBT), Istanbul, Turkey, 2019, pp.1.

[3] N. Jafarpisheh, N. Nafisi and M. Teshnehlab, Breast cancer relapseprognosis by classic and modern structures of machine learning algorithms.2018 6th Iranian Joint Congress on Fuzzy and IntelligentSystems (CFIS), Kerman, 2018, pp. 120-122.

[4] Z. Wang et al.,Breast Cancer Detection Using Extreme LearningMachine Based on Feature Fusion With CNN Deep Features.inIEEE Access, vol. 7, pp. 105146-105158, 2019.

[5] M. Gupta and B. Gupta .A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques .2018 SecondInternational Conference on Computing Methodologies and Communication (ICCMC), Erode, 2018, pp. 997-10020

[6] M. NEMISSI, H. SALAH and H. SERIDI, ”Breast cancer diagnosisusing an enhanced Extreme Learning Machine based-Neural Network,”2018 International Conference on Signal, Image, Vision and theirApplications (SIVA), Guelma, Algeria, 2018, pp. 1-40.

[7] N. M. J. Kumari and K. K. V. Krishna,Prognosis of Diseases UsingMachine Learning Algorithms: A Survey. 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT),Coimbatore, 2018, pp. 1-9.

[8] N. Khuriwal and N. Mishra, ”Breast cancer diagnosis using adaptivevoting ensemble machine learning algorithm,” 2018 IEEMA Engineer Infinite Conference (eTechNxT), New Delhi, 2018, pp. 1-5.

[9] S. Sathya, S. Joshi and S. Padmavathi,Classification of breast cancerdataset by different classification algorithms. 2017 4th InternationalConference on Advanced Computing and Communication Systems(ICACCS), Coimbatore, 2017, pp. 1-4.

[10] L. Liu, Research on Logistic Regression Algorithm of Breast CancerDiagnose Data by Machine Learning .2018 International Conferenceon Robots Intelligent System (ICRIS), Changsha, 2018, pp. 157-160.

[11] M. Amrane, S. Oukid, I. Gagaoua and T. Ensar˙I, Breast cancer classification using machine learning. 2018 Electric Electronics, Computer Science, Biomedical Engineerings’ Meeting (EBBT), Istanbul, 2018, pp. 1-4.

[12] A. P. Pawlovsky and M. Nagahashi,A method to select a good setting for the kNN algorithm when using it for breast cancer prognosis. IEEEEMBS International Conference on Biomedical and Health Informatics (BHI), Valencia, 2014, pp. 189-192.

[13] UCI Machine Learning Repository. Retrieve from http://archive.ics.uci.edu/ml (Accessed January, 2020) .

[14] B.R.A.Cirkovic, A. M. Cvetkovic, S. M. Ninkovic and N. D. Filipovic, ”Prediction models for estimation of survival rate and relapse for breast cancer patients,” 2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE), Belgrade, 2015, pp. 1-6.

[15] A. Vignesh, T. Yokesh Selvan, Ganesh Krishnan, Arjun N. Sasikumar, V. D. Ambeth Kumar, “Efficient Student Profession Prediction Using XGBoost Algorithm”, Lecture Notes on Data Engineering and Communications Technologies, Volume 35, pp 140-148, 2020.