International Journal of Research in Signal Processing, Computing & Communication System Design
4 (2), 2018, 33-35
http://www.publishingindia.com/ijrsp

Multi-Class Classification of Breast Cancer Using Machine Learning

Parneet Kaur Vohra[1], Boda Bhavani[2*] and Nagamani Gonthina[3]
[1]Assistant Professor, BVRIT Hyderabad, Hyderabad, Telangana, India.
Email: Praneeth.vohra@gmail.com
[2]M.Tech (CFIS), JNTUH College of Engineering Hyderabad, Hyderabad, Telangana, India.
Email: bodabhavani04@gmail.com
[3]Assistant Professor, BVRIT Hyderabad, Hyderabad, Telangana, India.
Email: gnvsk1986@gmail.com
*Corresponding Author

Abstract: Cancer is a big issue in the whole world. It has many subtypes, which includes Blood cancer, Skin cancer, Lung cancer, Breast cancer, etc. Breast cancer is one of the most leading causes of death among women. The factors that cause this disease cannot be easily determined. The early detection of abnormalities in breast enables the doctor to treat the breast cancer easily. The diagnosis process which determines whether the cancer is benign or malignant also requires a great deal of effort from the doctors and physicians.

A variety of Machine learning algorithms have now been applied to detect breast cancer, which includes Artificial Neural Networks (ANN), Bayesian Belief Networks (BBN), Support Vector Machines (SVM) and Decision Tree (DT) [1]. Many research papers about classification of breast cancer have only considered two classifiers such as a high and low-risk group. But, the binary classification detects cancer at the later stages, which is difficult to cure and the other drawback is it is error-prone i.e., the results of binary classification are not accurate. The error rate can be still decreased by multi-classifying the cancer data. The various Multi-class classification algorithms are Neural Networks, K-Nearest Neighbors, Boosting, Decision Trees etc. In this work, the three algorithms SVM, KNN, Gaussian Naïve Bayes algorithms are used for classification and K-means algorithm is used for clustering. The performance of these algorithms is analyzed.

Keywords: Artificial Neural Networks (ANN), Benign, Malignant, Support Vector Machines (SVM).

I. Introduction

Breast cancer occurs as a result of accumulation of dead cells commonly referred to as "Tumor". The reason there are many deaths of breast cancer is that, the tumor is not recognized in a patient in the early stages. A tumor does not mean cancer. Tumors are two types- Benign (harmless) and Malignant (harmful). Therefore, initially the classification of the dataset into two classes (Benign and Malignant) is done. Later, the malignant tumor has been clustered into two clusters (Severe and Moderate).

The algorithms used for classification are Support Vector Machine, Naïve Bayes, K-Nearest Neighbors. Support Vector Machine is a supervised learning algorithms used for classification. It differentiates the classes by using hyper plane. The algorithm outputs an optimal hyper plane that perfectly classifies the classes. Naïve Bayes algorithm is based on applying Bayes theorem with strong independent assumptions. K-Nearest Neighbors is an algorithm that makes predictions for an instance by searching through the entire training set for the k most similar instances. To determine the nearest neighbor, the distance measure called Euclidean Distance is used.

The algorithm used for clustering is K-means algorithm. It aims to partition n observations into k clusters in which each observation belongs to one cluster based on the mean value.

II. Related Work

In a previous research [2], a predictive model is developed to see how many women diagnosed with breast cancer have been survived. For this work, the comparison between three classification models- SVM, ANN and SSL (Semi Supervised Learning) has been made with the SEER cancer database. In this dataset, a variable called survivability was also considered which had the value "yes" for the patient who survived and "no" for the patient who didn't survive. The other features considered for this work include tumor size, the number of nodes and the age at the time of diagnosis.

By comparing the performance of these three algorithms, the work concluded that the accuracy of SVM, ANN and SSL was 65%, 51% and 71% respectively. For evaluating the performance the 5-fold cross validation has been used. Based upon the result, the work proposed that SSL model was the good candidate for survival analysis of breast cancer diagnosed patients.

Here, the point to be noted is that no preprocessing steps were mentioned by the authors of the work. The work is just proceeded with the SEER dataset and the box-plot graphs were used to visualize the performance. If the dataset is preprocessed by various methods like data cleaning, data transformation, data integration, reducing redundant data, etc., the results would have been more accurate.

III. Proposed Methodology

Previously, much research has undergone to classify the breast cancer into two classes- Benign and Malignant. The main disadvantage of binary classification is that, it leads to some error i.e., there will be some error in some cases if we use binary classification. For instance, assume we have a set of balls of many different colors. We have two baskets of pink and blue color. The work is to segregate the balls into either pink basket or blue basket based on the intensity of the picked ball color. If the ball is pink color (or blue color), put the ball into pink basket (or blue basket).

But, what if the picked ball is green colored? Assuming that green is nearer to blue (Intensity), the ball is placed into blue basket. This guarantee's that there is an error in classification process. The below Fig. 1 depicts the disadvantage of binary classification and Fig. 2 depicts how the multi-class classification overcomes this disadvantage.

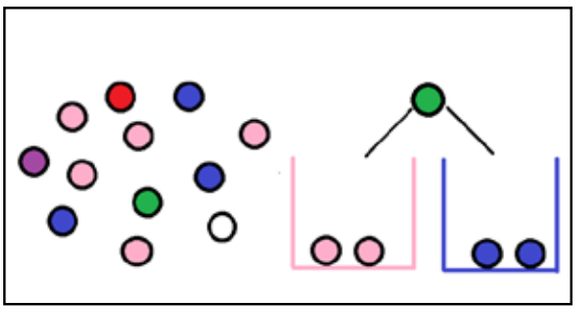

Fig. 1: Disadvantage of Binary Classification

This error can be overcome if we use one more basket of green color and place the green color ball in green color basket. A reasonable amount of error has been decreased due to introducing one more class.

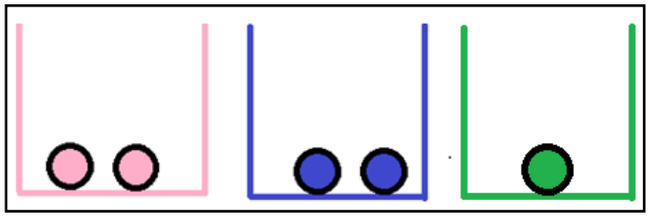

Fig. 2: Overcoming the Advantage of Binary Classification

As shown in the Fig. 2, the green ball is placed in newly introduced green basket to reduce the error. In this way, as the number of classes' increases the error rate will decrease.

Therefore, this work mainly focuses on classifying the breast cancer into three classes. So, this is multi-class classification problem. Since dataset is unavailable that has three class labels, initially the dataset is classified into two classes (Benign and Malignant) and cluster the Malignant into two types- severe risk and moderate risk.

Initially, cancer is binary classified into Benign or Malignant stage, using Support Vector Machine (SVM), Gaussian Naïve Bayes (NB) and K-Nearest Neighbors algorithms as shown in Fig. 3.

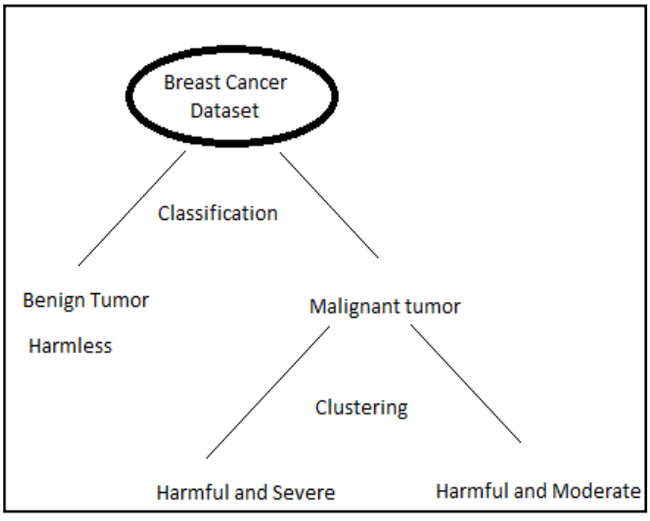

Fig. 3: Flowchart of the Process

The three algorithms are applied on the training set to classify into two classes and tested on the testing data. The performance of these algorithms is showed in the Fig. 4 which shows there is very poor performance in case of Support Vector Machine as compared to the other two.

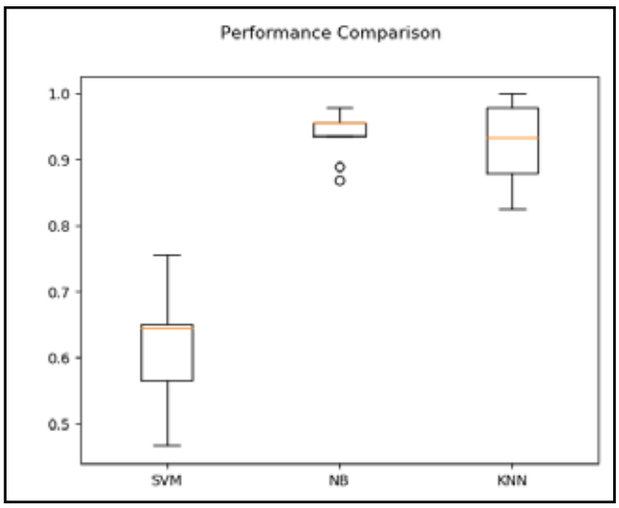

Fig. 4: Performance Comparison of Various Algorithms

This surprisingly bad performance of SVM is because there are many types of features in the dataset and the feature with the largest value dominates the feature with the smallest value i.e., variables measured at different scales do not contribute equally to the analysis. One solution for this is to standardize the dataset. Data standardizing procedures equalizes the range and data variability.

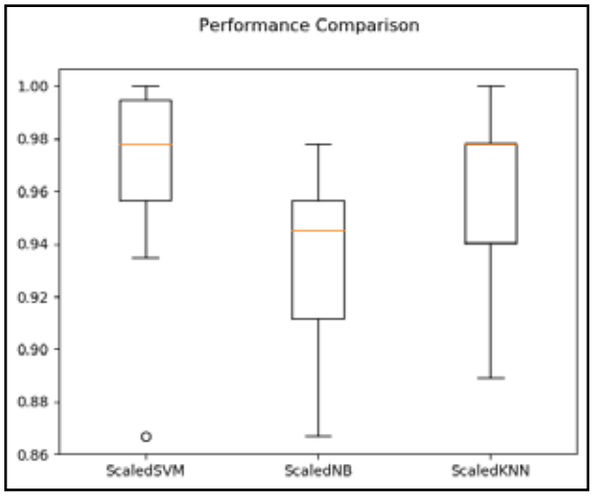

Fig. 5: Performance after Standardizing Data

The pipelines are used to standardize the data and build the model for each fold in cross-validation. The Fig. 5 shows the change in the performance of the algorithms after standardizing the dataset. The performance of the SVM algorithm has increased, whereas the Naïve Bayes performance is decreased. Therefore, SVM is the best algorithm and is used after standardizing the data.

Now, the malignant tumor has to be clustered into two clusters and depending upon to which cluster the tuple belongs, we classify the tuple into either severe risk or moderate risk. For this purpose the K-means clustering algorithm is used as there are no labels in the dataset.

K-means clustering is a type of unsupervised learning algorithm, which is used when we don't have labels in the data. Here k represents the number of clusters to be formed. The algorithm works iteratively and assigns each data point to one of k groups based on the features.

IV. Conclusion

The binary classification of breast cancer dataset to two classes- Benign and Malignant leads to error and it detects the cancer in later stages. These two disadvantages of binary classification can be overcome by multi-class classification. Initially binary class algorithms- SVM, Naïve Bayes, and K-Nearest Neighbors are used to classify dataset into two classes. Later, Malign tuples are clustered into two clusters using K-Means clustering algorithm. Therefore, the dataset is classified into three classes which will result in reduction of error. Among the three algorithms used for classification, SVM initially showed poor performance due to the presence of unscaled data. Later after standardizing the data, the performance has increased in the case of SVM. Later, the malignant patients are clustered into two classes called severe risk and moderate risk by using K-means clustering algorithm.

References

[1] P. Dhivyapriya, and S. Sivakumar, "Classification of cancer dataset in data mining algorithms using R tool," *International Journal of Computer Science Trends and Technology (IJCST)*, vol. 5, no. 1, pp. 79-83, January-February 2017.

[2] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal,* vol. 13, pp. 8-17, 2015.

[3] K. Menaka, and S. Karpagavalli, "Breast cancer classification using support vector machine and genetic programming," *International Journal for Innovative Research in Computer and Communication Engineering*, vol. 1, no. 7, pp. 1410-1412, September 2013.

[4] A. Mohamed, "Survey on multiclass classification methods," November 2005.