www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 7 Issue 5 May 2018, Page No. 23891-23895
Index Copernicus Value (2015): 58.10, 76.25 (2016) DOI: 10.18535/ijecs/v7i5.03

Analysis of Machine Learning Techniques for Breast Cancer Prediction

Priyanka Gupta, Prof. Shalini L

B.Tech. Scope, VIT University,Vellore, Tamil Nadu, India
Assistant Professor (senior),Scope, VIT University,Vellore, Tamil Nadu, India

Abstract: Of all the deaths worldwide, the second most common cause of death is breast cancer. In 2011, the number of deaths was estimated to be 5.8 millions because of breast cancer. So, early diagnosis of cancer is very important. In this paper, machine learning techniques are explored in order to increase the accuracy of diagnosis. Methods such as CART, Random Forest, K-Nearest Neighbors are compared. The dataset used is obtained from UC Irvine Machine Learning Repository. The obtained accuracy prediction performances are proportionate to existing methods. However it is found that KNN algorithm has much better performance than the other techniques used in comparison.

Keywords: – Diagnosis, Algorithm, CART, Random Forest, Boosted Trees,K-Nearest Neighbor.

I. INTRODUCTION

The cancer that develops from breast tissue is breast cancer. It can be found in both men and women, but it is more common for women. Breast cancer is caused when malignant or cancerous tumor emerges in the breast. The tumor when matures, it spread to more parts of the body. Metastasize occurs through a primary route called lymphatic system which also produce and transport white blood cells and additional immune system cells which are cancer fighting. These cancer cells which are not killed by system's white blood cells pass through the lymphatic vessels to go in remote body sites, in order to form more tumors and preserve the disease mechanism. But some of these risks are preventable and some are not. To decrease the risk, people may take actions. Even a preventable risk is risky enough.

The traditional method for diagnosis of breast cancer was a invasive technique, known as breast biopsy, in which a piece of breast tissue is removed medically and then the sample is examined by a specialist. However, a less invasive method can be used now. In this technique, sample can be taken from minimally invasive needle aspirate method. The sample obtained through this method is easily digitized and used for computation based diagnostics. Awareness and research has helped in creating advances in diagnosis and treatment of breast cancer. Survival rates have grown and the number of deaths due to this disease is decreasing, due to early detection of cancer stages. A recent data has shown survival rate to be 88% after 5 years of diagnosis and to be 80% after 10 years of diagnosis.

Data Mining is the exploration of important information in massive amount of data. It is more used in healthcare systems. Data mining algorithm has provided great assistance for early stage prediction of breast cancer that has always been a challenging research problem. So, statistical data driven research has become a common method to many substantial areas like biotechnology and medicine. This trend is clearly visible in the studies of Cios et al.[1] and Houston et al.[2].

In this paper, we have compared data mining methods to conclude accuracy of each method. The paper is arranged as follows. The next section reports literature survey. Section 3 reviews the methodology used to conduct the prediction analysis. In Section 4, experimental results are presented. Conclusion and future work are given in the last section.

II. LITERATURE SURVEY

There have been several studies on the prediction problem by using various statistical approaches.

Through literature review, we only found a few studies which were related to medical diagnosis. But all the contributions that are made in this field are getting more advance day by day. The algorithms that are used are mainly data mining algorithms tested on dataset.

In this work [3], authors demonstrate different classification algorithms like Logistic Regression, Naïve Bayes, Artificial Neural Networks. The research aims at providing the result as it evaluates the data according to quality grammatically. It is established that these algorithms had maximum lifting factors for most of the class values.

In another paper [4], the prime focus is the study of breast cancer prediction and classification such that precautionary measures can be built at an early stage before the start of cancer. So to complete the prediction, data mining algorithms like Decision trees, Clustering are used.

In paper [5], Naïve Bayes was employed and it produced an accuracy of about 96%.

In another paper[6], impressive results were seen when RapidMiner was used to create SVM classifier and it showed accuracy of about 80%.

In paper[7], authors implemented fuzzy clustering method and showed an accuracy level of 95.57%.

In another paper[8], authors made a mixed classification model of SVM and Decision Tress which resulted in 91% accuracy.

Dataset Description:

The dataset is retrieved from University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. Sample appears periodically as Dr. Wolberg recorded his clinical cases. The database hence reflects the sequential arrangement of the data. This grouping data emerges immediately below.

The features in the dataset were calculated from a digitized image of a fine needle aspirate of a breast piece. They describe the properties of the cell nuclei found in the image. Various nuclei in each sample were examined. Attributes 2 to 10 have been used to produce instances. Every instance comprises one of the two available classes: benign or malignant. The dataset has 569 observations and 32 attributes (ID, diagnosis, 30 real-valued input features), with no missing values. The attributes (columns) information is as follows: (1) ID number (2) Diagnosis (M = malignant, B = benign) (2-32) Ten real-valued features are calculated for each and every cell nucleus: a) radius (mean of distances from center to points on the perimeter) b) texture (standard deviation of gray-scale values) c) perimeter d) area e) smoothness (local variation in radius lengths) f) compactness (perimeter^2 / area - 1.0) g) concavity (severity of concave portions of the contour) h) concave points (number of concave portions of the contour) i) symmetry j) fractal dimension (“coastline approximation” - 1). The mean, standard error, and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features.

III. METHODOLOGY

In this study three classification methods are examined and compared. These four classification models are chosen to extract the most accurate model for predicting cancer survivability rate. These concepts are explained as follows:

A. Classification and Regression Trees (CART):

A CART model is represented as binary tree. Every root node shows a single input variable say x and a split point on such a variable (assume that variable to be numeric). Building a CART model includes selection of input variables and the precise split is selected using a greedy technique to diminish cost function. The tree structure ends using a preassigned stopping test. CART uses greedy method in which decision trees are built in a top-down recursive way. CART system can locate multivariate splits on the basis of linear combination of features. Multivariate splits are a type of feature construction in which new features are introduced on the basis of existing features. CART method is used in order to develop the trees.

For Regression predictive modeling, the cost function is decreased to select points is the sum squared error over all training samples.

For Classification, the Gini Index is applied which gives an expression of purity of leaf nodes.

$$G=\frac{2}{n}\frac{\sum_{i=1}^{n}(i\ y_i)}{\sum_{i=1}^{n} y_i}-\frac{n+1}{n}$$

B. Random Forest:

It is an ensemble learning algorithm for classification and regression. It is a meta evaluator that fits many decision tree classifiers on it and employ averaging to increase the predictive certainity and regulate over-fitting. The original input sample size is constantly similar to sub sample size but the samples are taken with replacement if bootstrap is true.

Accuracy of Random Forest is dependent on the fortitude of the respective classifiers and a measure of dependence between them. The main idea is to

preserve the strength of each classification model without changing their correlation.

Random Forests do not vary due to the number of features chosen at each split. This method provides internal evaluation of variable importance.

C. K-Nearest Neighbors

One of the most used algorithm is K-Nearest Neighbor in machine learning. It is an algorithm that is based on the occurances that is not necessary for learning stage. This method was first derived in the early 1950s.

The training sample combined with a choice and distance function of a class is based on the nearest neighbor classes. A comparison must be done with other elements before classifying a new element by using a similarity function. The K-Nearest Neighbors are then acknowledged. The class that comes highest among neighbors is appointed to the element to be classified. The neighbors are then granted weights by the distance which separates it to the recent elements.

The methodology is labor driven when provided with large amount of training sets. Nearest Neighbor classifiers works on learning through analogy that is, by matching a test tuple with training tuple that is more alike.

'closeness' is described through distance metric like Euclidean distance. The formula for Euclidean Distance between two points say , In a plane with point p_1 at (x_1, y_1) and point p_2 at (x_2, y_2), it is

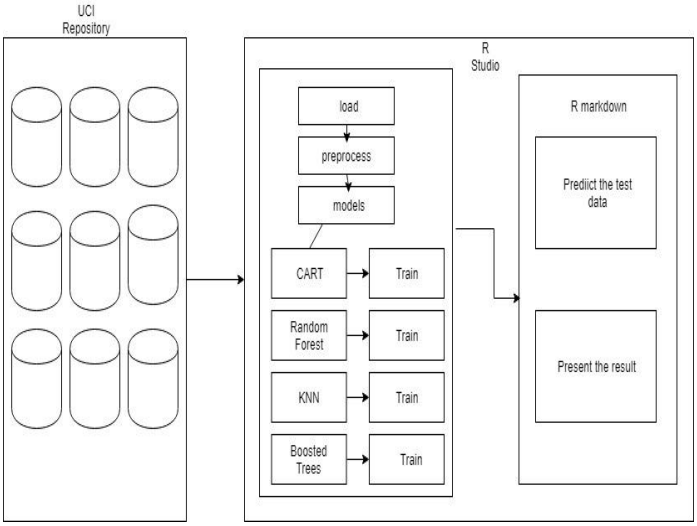
$E(D)=\sqrt{((x_1 - x_2)^2 + (y_1 - y_2)^2)}$.

D. Boosted Trees

Boosting involves N learners by implanting addition data in the training set. Random sampling is used with replacement to produce N new data sets from the original sets. In this method, weights are attached to each training tuple. A set of K-classifiers are iteratively studied. After a classifier is learned, the next classifier is allowed by updation of weights. The latest boosted classifier associates the votes of each single classifier, where the weight of each classifier's vote is defined as a function of its efficiency. In gradient boosting, it trains multiple models sequentially. Each new model gradually decreases the loss function of the whole system using Gradient Descent method. The learning methods consecutively train new models to give an increased accuracy estimate of the response variable.

The work flow diagram of breast cancer detection using these four algorithms are given in fig. 1.

Fig.1.: Architecture of the system



IV. Experimental Results:

For better efficiency, the data repository is first preprocessed. The data is collected from UCI Machine Learning Repository. In next step, the raw dataset is preprocessed to transform numeric values to nominal. This step is carried out using discretization method. After this step, all the valid attributes are selected. The next step is to use various classification algorithms to organize dataset into benign and malignant. Finally, performances of all models are discussed.

In this experiment, we have used confusion matrix. It is a table which is used to measure the performance of a classification model on a group of data for which true values are already known. By measuring the number of correct and incorrect classification, the level of accuracy is measured.

Sensitivity is also measured which calculates the fraction of positives which are measured correctly while specificity calculates the fraction of negatives which are measured accurately.

Kappa is the measure of interrater reliability. This occurs when data collectors provide the similar result to the same data.

In our testing data for boosted tree classifier, there are 107 benign cases and 63 malignant cases. For CART, there are 100 benign cells are predicted correctly and 57 malignant cells are predicted correctly and as a result, CART shows an accuracy of 92%. Another classifier Random Forest shows an accuracy of 96.5% by predicting 104 benign and 60 malignant correctlyOut of which, 103 benign are predicted correctly and 61 malignant are predicted correctly. The analysis is shown in the table 1.

TABLE I: CONFUSION MATRIX OF THE DATASET OBTAINED USING BOOSTED TREE CLASSIFIER

		Predicted	
		Malignant	Benign
Actual	Malignant	103	2
	Benign	4	61

All the classification models when trained with the breast cancer data provides the following results shown in table 2.

Table 2: Performance of classification models

Classification Models	Accuracy	kappa	Sensitivity	Specificity
CART	0.9235	0.8366	0.9346	0.8906
Random Forest	0.9647	0.9243	0.9524	0.9720
K-Nearest Neighbor	0.9700	0.9360	0.9474	0.9839
Boosted Trees	0.9647	0.9248	0.9683	0.9626

It can be seen here that CART achieved an accuracy of about 92%, which is not the best. Random Forest had 96.5% accuracy same as that of boosted trees. The best accuracy of 97% is shown by K-Nearest Neighbor algorithm.

V. CONCLUSION

Breast cancer is one of the leading causes of death in women. So, early detection is very important. Breast cancer detection done by less invasive technique gives digitized result which makes prediction easier. There are many algorithms available that can be used in order to train the dataset and get the accurate results. In this paper, four classification algorithms are used in order to find their accuracy. The different classification models trained on the dataset showed different accuracy. The most accurate model was K-Nearest Neighbor. The classification model such as Random Forest and Boosted Trees showed the similar accuracy. Therefore, the most accurate classifier can be used to detect the tumor so that the cure can be found in early stage.

REFERENCES

[1] Cios KJ, Moore GW. Uniqueness of medical data mining. Artificial Intelligence in Medicine 2002; 26:1-24.

[2] Houston, Andrea L. and Chen, et. al.. Medical Data Mining on the Internet: Research on a Cancer Information System. Artificial Intelligence Review 1999; 13:437-466.

[3] K. Shiny,"Implementation of Data Mining Algorithm to Analysis Breast Cancer", International Journal for Innovative Research in Science and Technology, vol.1, no.9, (2015), pp.207-212.

[4] M. Kumar, S. S. Tomar and B.Gaur, "Mining based Optimization for Breast Cancer Analysis: A Review",International Journal of Computer Applications, vol. 19, no. 13,(2015).

[5] Gayathri, B. M., & Sumathi, C. P. (2016). An Automated Technique using Gaussian Naïve Bayes Classifier to Classify Breast Cancer. International Journal of Computer Applications, 148(6).

[6] Priyanka Jain & Santosh Kr. Vishwakarma (2016). Collaborative Analysis of Cancer Patient Data using Rapid Miner. International Journal of Computer Applications, 145, 8-13.

[7] Nauck, D., & Kruse, R. (1999). Obtaining interpretable fuzzy classification rules from medical data. Artificial intelligence in medicine, 16(2), 149-169.

[8] K.Sivakami,"Mining Big Data:Breast Cancer Prediction using DTSVM Hybrid Model." International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-5,August 2015.

[9] S.B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Informatica 31(2007) 249-268, 2007.

[10] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition. San Fransisco:Morgan Kaufmann; 2005.

[11] M. Kumar, S. S. Tomar and B.Gaur, "Mining based Optimization for Breast Cancer Analysis: A Review",International Journal of Computer Applications, vol. 19, no. 13,(2015).

[12] K. Arutchelvanand R. Periasamy, "Analysis of Cancer Detection System Using Datamining Approach",International Journal of Innovative Research in Advanced Engineering, vol. 2, no. 11, (2015)

[13] K. Balachandran and R. Anitha, "Ensemble based optimal classification model for pre-diagnosis of lung cancer",2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), IEEE,(2013).

[14] Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkanen L, Joensuu H. Artificial neural networks

applied to survival prediction in breast cancer. Oncology 1999; 57:281- 6.
[15] William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.