International Research Journal in Global Engineering and Sciences. (IRJGES)
ISSN : 2456-172X | Vol. 3, No. 3, December 2018 - February, 2019 | Pages 125-130

An Outline of Machine Learning Techniques for Breast Cancer Prediction

Shiny R.M[1] , Jetlin C P[2]

Assistant Professor-Department of CSE,
Agni College of Technology, Chennai

ABSTRACT

Breast cancer is one of the major threat in middle aged women throughout the world. In today's world, this is the second most threatening cause of death in women. Early diagnosis can significantly reduce the chances of death. But it is not an easy due to several uncertainties in detection. Machine Learning techniques can be used to develop tools and that can be used as an effective mechanism for early detection and diagnosis of breast cancer for physicians which will greatly enhance the survival rate of cancer patients. This paper compares three of the most popular ML techniques which are used for breast cancer detection and diagnosis. The techniques are Support Vector Machine (SVM), Random Forest (RF) and Naive Bayes (NB). The probability will be calculated for each of these techniques and the algorithm with highest probability provides the more accurate results.

1.0 Introduction

In recent years, several studies have applied data mining algorithms on different medical datasets to classify Breast Cancer. These algorithms show good classification results and encourage many researchers to apply these kinds of algorithms to solve challenging tasks. Cancer is a heterogenous disease that may be divided into many types. According to World Health Organization, twenty five percent of the females are diagnosed with breasts cancer at some stage in their life. In UAE, forty third of feminine cancer patients are diagnosed with breast cancer. Accurately predicting a cancerous growth remains a difficult task for several physicians. The emergence of latest medical technologies and therefore the monumental quantity of patient information have impelled the trail for the event of latest methods within the prediction and detection of cancer. Recurrent breast cancer is cancer that comes back in the same or opposite breast or chest wall after a period of time when the cancer couldn't be detected. Though information assessment that is collected from the patient and a physician's intake greatly contributes to the diagnostic method, supportive tools could be superimposed to assist facilitate proper diagnoses. These tools aim to eliminate possible diagnostic errors and supply a quick method for analyzing the large chunks of data.

In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Machine Learning (ML), is a subfield of Artificial Intelligence (AI) that permits machines to learn without expilicit programming by exposing them to sets of information permitting them to learn a selected task through expertise .Over the previous few decades, machine

learning strategies have been widespread within the development of predictive models so as to support effective decision-making. In cancer analysis, these techniques could be used to determine completely different patterns during a information set and consequently predict whether or not a cancer is malignant or benign. The performance of such techniques will be evaluated supported the accuracy of the classification, recall, precision, and therefore the space underneath ROC. Recently, data processing has become a well-liked economical tool for data discovery and extracting hidden patterns from massive datasets. It involves the employment of refined knowledge manipulation tools to get antecedently unknown, valid patterns and relationships in massive dataset. We tend to apply three robust data processing classification algorithms i.e. SVM, Random forest and Naïve Bayes on a medium sized knowledge set that contained thirty five attributes and 198 cancer patient data.

The paper is arranged as follows: Section 2 delves with the learning methods of Machine Learning . Section 3 brings out the broad view of the experimental analysis. Section 4 renders about the results of the experiments. Section 5 concludes the report with the future work that can be carried out.

2.0 Learning Methods in Machine Learning

The learning method in ML techniques are often divided into 2 main categories, supervised and unsupervised learning. In supervised learning, a set of data instances are used to train the machine and are labelled to grant the right result. However, in unsupervised learning, there are no pre-determined knowledge sets and no notion of the expected outcome, which suggests that the goal is tougher to achieve.

Support Vector Machines:

A support vector machine (**SVM**) is a supervised machine learning model that uses classification algorithms for two-group classification problems. SVM is one of the supervised machine learning classification techniques that's widely applied in the field of cancer identification and prognosis. SVM functions by choosing critical samples from all categories called support vectors and separating the classes by generating a linear function that divides them as broadly as possible using of these support vectors.

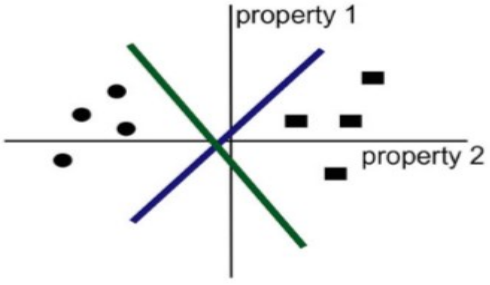

Fig.1 SVM generated by hyperplanes

Random Forest:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees. The RF methodology relies on an algorithmic approach within which each iteration involves choosing one random sample of size N from the data set with replacement, and another random sample from the predictors without replacement. Then the data obtained is partitioned. The out-of-bag data is then dropped and the above steps repeated many times depending on how many trees are needed. Finally, a count is made over the trees that classify the observation in one category and within the other. Cases are then classified based on a majority vote over the decision tree.

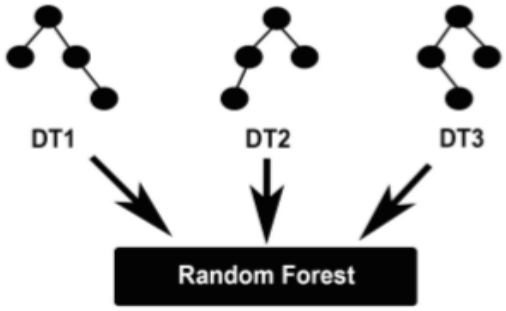

Fig.2 A Visual of random forest working

Naive Bayes:

It is a classification technique based on *Bayes* Theorem with an assumption of independence among predictors. In simple terms, a *Naive Bayes* classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.BN is a subfield of probabilistic graphical models that are used for prediction and knowledge representation of uncertain domains. BN correspond to a widely used structure in machine learning known as the directed acyclic graph (DAG). This graph consists of many nodes, each equivalent to a variable and the node edges represents direct dependence among the corresponding nodes in the graph.

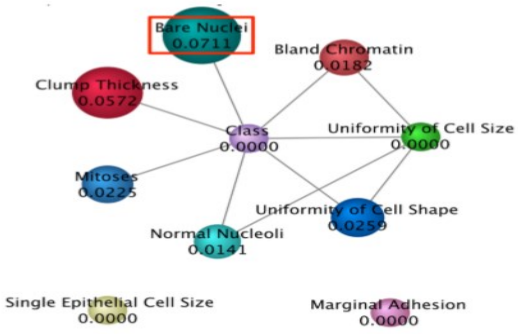

Fig.3 DAG Model With Breast Cancer Attributes

3.0 Experimental Analysis:

Our investigation is based on the initial Wisconsin breast cancer data set that's obtained from the UCI Machine Learning Repository, an online open source repository. This data set was collected periodically over 3 years by Dr. William H. Wolberg from the University of Wisconsin Hospitals and consists of 669 instances, where the cases are classified as either malignant or benign. 458 of the cases are benign and 241 are malignant. The ten attributes are: • Clump Thickness • Cell Size Uniformity • Cell shape Uniformity • Marginal Adhesion • Single epithelial cell Size • Bare Nuclei • Bland chromatin • normal Nuclei • Mitoses • class.

Training set:

The classifier are tested using the k – fold cross validation methodology. This validation technique can randomly separate the training set into k subsets where one of the k-1 subsets are used for testing and the rest for training. 10-fold cross-validation is the preferred k value utilized in most validation in ML and will be used in this paper. This implies nine subsets will be used for training of the classifier and the remaining one for the testing. This system is used to avoid over fitting of the training set, which is likely to occur in small data sets and large number of attributes.

Fig.4 10-k Cross Validation Method

Simulation software:

In this paper, the Waikato environment for knowledge Analysis (WEKA) software was used as a ML tool. WEKA - an open-source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems. WEKA is a Java based open supply tool that was initially released to the public in 2006 under the GNU General Public License. This tool provides many ml techniques and algorithms as well as the classification techniques that are being investigated in this paper. Alternative features include data preprocessing, clustering, feature selection evaluation and rule discovery algorithms. Data sets are accepted in many formats like CSV and ARFF. Besides being an open-source tool, Weka is additionally attractive due to its portability and ease of use GUI.

4.0 Results and Discussion:

This section describes the parameters and presents the results that assists the 3 classifiers that are being investigated in this paper.

i.Accuracy:

The classifier accuracy is a measure of how well the classifier will properly predict cases into their correct category. It is the number of correct predictions divided by the total number of instances within the data set. it's worth noting that the accuracy is highly dependent on the threshold chosen by the classifier and may therefore change for various testing sets. Hence, accuracy may be calculated using the subsequent equation:

$$Accuracy = \left(\frac{K_{TP} + K_{TN}}{K_P + K_N}\right) \times 100\%$$

ii.Recall:

Recall, also commonly referred to as sensitivity, is the rate of the positive observations that are correctly predicted as positive. This measure is desirable, particularly in the medical field because how many of the observations are properly diagnosed. in this study, it is more important to properly identify a malignant tumor than it is to incorrectly identify a benign one.

$$Recall = \left(\frac{K_{TP}}{K_P}\right) \times 100\%$$

RECALL VALUES.

	SVM	RF	BN
Benign	97.4%	96.9%	96.5%
Malignant	96.3%	95.9%	98.3%
Average	97.0%	96.6%	97.1%

iii.Precision:

Precision, also commonly called confidence, is the rate of both true positives and true negatives that are identified as true positives. This shows how well the classifier handles the positive observations but doesn't say much regarding the negative ones. The precision values for all 3 techniques are shown in Table:

PRECISION VALUES.

	SVM	RF	BN
Benign	98.0%	97.8%	99.1%
Malignant	95.1%	94.3%	93.7%
Average	97.0%	96.6%	97.2%

iv.ROC area:

A receiver operating characteristics (ROC) graph is a way to visualize a classifiers performance by showing the trade-off between the price and advantage of that classifier. roc is one

amongst the foremost common and useful performance measure for data processing techniques. Percentage values for the roc area of all 3 techniques are shown in Table:

AREA UNDER ROC VALUES

	SVM	RF	BN
Benign	96.4%	99.8%	99.0%
Malignant	96.8%	99.9%	99.2%
Average	96.6%	99.9%	99.1%

5.0Conclusion:

ML techniques are widely used in the medical field and have served as a useful diagnostic tool that helps physicians in analyzing the available data as well as designing medical expert systems. This paper conferred three of the most common ML techniques normally used for carcinoma detection and diagnosis, namely Support Vector Machine (SVM), Random Forest (RF) and Bayesian Networks (BN). the main features and methodology of each of the three machine learning techniques was represented. Performance comparison of the investigated techniques has been applied using the original Wisconsin breast cancer data set. Simulation results obtained has proved that classification performance varies based on the strategy that is selected. Results have showed that SVMs have the highest performance in terms of accuracy, specificity and precision. However, RFs have the highest probability of properly classifying tumor.

References:

a) *WHO — Breast Cancer: Prevention and Control (2015) Retrieved 20 Jan 2015, from WHO — http://www.who.int/cancer/detection/breastcancer/en/index1.html*
b) *Y. Elobaid, T.-C. Aw, J. N. W. Lim, S. Hamid, and M. Grivna, "Breast cancer presentation delays among Arab and national women in the UAE, a qualitative study," SSM - Popul. Heal., Mar. 2016.*
c) *S. Conrady and L. Jouffe, Bayesian Node Analysis. 2013.*
d) *Weka 3.5.6, An open source data mining software tool developed at university of Waikato, New Zealand,http://www.cs.waikato.ac.nz/ml/weka/ 2009.*
e) *http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-key-statistics*
f) *http://www.wcrf.org/int/cancer-facts-figures/data-specificcancers/breast-cancer-statistics*
g) *https://archive.ics.uci.edu/ml/datasets*
h) *Wolberg, William H., and Olvi L. Mangasarian. "Multisurface method of pattern separation for medical diagnosis applied to breast cytology".*
i) *Abdelaal, Medhat Mohamed Ahmed, et al. "Using data mining for assessing diagnosis of breast cancer." Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on. IEEE, 2010.*
j) *Liu, Huan, and Lei Yu. "Toward integrating feature selection algorithms for classification and clustering.".*
k) *Vanaja, S., and K. Ramesh Kumar. "Analysis of feature selection algorithms on classification: a survey." International Journal of Computer Applications (2014).*