PROGNOSIS OF BREAST CANCER FROM MAMMOGRAMS

SANIYA AIMAN BAGALI[1], MUJAMIL DAKHANI[2]

[1]*Student Mtech (CNE) SIET Vijaypur Karnataka (INDIA)*
[2]*Assistant Prof (Mtech) SIET Vijaypur Karnataka (INDIA)*

---***---

Abstract - *Cancer has been characterized as a miscellaneous disease consisting of many different subtypes. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. Breast cancer is one among them. The importance of classifying breast cancer patients into high or low risk groups has led many research teams, from the biomedical field to study the application of machine learning (ML) methods. Therefore, these techniques have been utilized as an aim to model the progression and treatment of cancerous conditions. A variety of these techniques, including Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Decision Trees (DTs) have been widely applied in breast cancer research for the development of predictive models, resulting in effective and accurate decision making. The predictive models discussed here are based on various supervised ML techniques as well as on different input features and data samples. It determines whether the sample is cancerous (malignant) or non-cancerous (benign) and finally it compares the accuracy and precision of different algorithms and determines which among them is best. Predict whether the mammogram mass is benign or malignant.*

Key Words: ***Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs), Decision Trees (DTs), Machine Learning, Benign, Malignant.***

1. INTRODUCTION

The American Cancer Society defines cancer as a generic term for a large group of diseases that can affect any part of the body. Other terms are malignant tumors and benign tumors. Physicians need a reliable diagnosis procedure to distinguish between these tumors. But generally it is very difficult to distinguish tumors even by the experts. Hence automation of diagnostic system is needed for diagnosing tumors. Breast cancer is a type of cancer which affects the breast tissue which is most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk. Breast cancer risks can be reduced via early detection of the disease. In the last few decades, several data mining and machine learning techniques have been developed for breast cancer detection and classification, which can be divided into three main stages: preprocessing, feature extraction, and classification. To facilitate interpretation and analysis, the preprocessing of mammography films helps improve the visibility of peripheral areas and intensity distribution, these tests include Breast exam Mammogram Breast ultrasound Biopsy. As an alternative we can also use Machine Learning techniques for the classification of benign and malignant tumors. The prior diagnosis of Breast Cancer can enhance the prediction and survival rate notably, so that patients can be informed to take clinical treatment at the right time. Classification of benign tumors can help the patients avoid undertaking needless treatments. Thus the research is to be carried for the proper diagnosis of Breast Cancer and categorization of patients into malignant and benign groups. In this work, there were two challenges to automate the breast cancer diagnosis (i) determining which model best fits the data and (ii) how to automatically design and adjust the parameters of the machine learning model.

2. LITERATURE REVIEW

[1] David B. fogel et al. has discussed the evolving neural networks for detecting breast cancer and the related works used for breast cancer diagnosis using back propagation method with multilayer perceptron. In contrast to back propagation David B. fogel et al. found that evolution computational method and algorithms were used often, outperform more classic optimization techniques.

[2] Chih-Lin Chi et al., 2007 have presented an article on survival analysis of breast cancer on two breast cancer datasets. This article applies an Artificial Neural Networks (ANNs) to the survival analysis problem. Because ANNs can easily consider variable interactions and create a non-linear prediction model, they offer more flexible prediction of survival time than traditional methods. This study compares ANN results on two different breast cancer datasets, both of which use nuclear morphometric features. The results show that ANNs can successfully predict recurrence probability and separate patients with good and bad prognosis.

[3]Shajahan et al (2013) worked on the application of data mining techniques to model breast cancer data using decision trees to predict the presence of cancer. Data collected contained 699 instances (patient records) with 10 attributes and the output class as either benign or malignant. Input used contained sample code number, clump thickness, cell size and shape uniformity, cell growth and other results physical examination. The results of the supervised learning algorithm applied showed that the random tree algorithm had the highest accuracy of 100% and error rate of 0 while CART had the lowest accuracy with a value of 92.99% but naïve bayes' had the an accuracy of 97.42% with an error rate of 0.0258.

[4] Afzan Adam et al. have developed a computerized breast cancer diagnosis by combining genetic algorithm and Back propagation neural network which was developed as faster classifier model to reduce the diagnose time as well as increasing the accuracy in classifying mass in breast to either benign or malignant. In these two different cleaning processes was carried out on the dataset. In Set A, it only eliminated records with missing values, while set B was trained with normal statistical cleaning process to identify any noisy or missing values. At last Set A gave 100% of highest accuracy percentage and set B gave 83.36% of accuracy. Hence the author has concluded that medical data are best kept in its original value as it gives high accuracy percentage as compared to altered data.

[5] Naresh Khuriwal, Nidhi Mishra took data from Wisconsin Breast Cancer database and worked on breast cancer diagnosis. The results of their experiments proved that ANN and Logistic Algorithm worked better and provided a good solution. It achieved an accuracy of 98.50%.

[6] Kaewchinporn et al.9 presented a new classification algorithm tree bagging and weighted clustering (TBWC) combination of decision tree with bagging and clustering. This algorithm is experimented on two medical datasets: cardiocography1, cardiocography2 and other datasets not related to medical domain.

[7] Delen et al.10 had taken 202,932 breast cancer patients records, which then pre-classified into two groups of "survived" (93,273) and "not survived" (109,659). The results of predicting the survivability were in the range of 93% accuracy.

[8] S.Kharya worked on breast cancer prediction and stated that artificial neural networks are widely used. The paper featured about the advantages and short comings of using machine learning methods like SVM, Naive Bayes, Neural network and Decision trees.

3. METHODOLOGY:

In this work, many applied techniques were tested for the subsequent stages of processing and analysis of the breast cancer dataset.

Stage 1: Preprocessing. As a part of this research, processing was performed on the raw breast cancer data to scale the features using the Standard Scaler module. Standardization of datasets is a common requirement for many machine learning estimators.

Stage2: Feature Selection. Usually, feature selection is applied as a preprocessing step before the actual learning. However, no algorithm can make good predictions without informative and discriminative features; therefore, to keep the most significant features and reduce the size of the dataset, we implemented PCA using randomized SVD. The module used for feature selection was implemented in using the Python scikit-learn library. All selection strategies were based to many criteria to extract the best features.

Stage 3: Machine Learning Algorithm. Usually, ensemble machine learning algorithms allow better predictive performance compared with a single model. This can be considered machine learning competition, where the winning solution was used as a model for breast cancer diagnosis. In this paper, the following heterogeneous ensembles machine learning algorithms were used to classify the given data set: support vector machine (SVM), K-nearest neighbor (KNN) decision tree (DT), logistic regression (LR).

Stage 4: Parameter Optimization. Genetic Programming (GP) is a type of evolutionary algorithm (EA) that generalizes the genetic algorithm. GP is a model for testing and selecting the best choice among a set of results. Based on biological evolution and its fundamental mechanism (mutation, crossover, and selection), GP generates a solution. The use of GP is the reason for its flexibility; it can model systems where the structure of the desired models and the key features are not known.

Classification algorithms in Machine Learning:

1. Naïve Bayes' classifier

Naive Bayes Classifier is a probabilistic model based on Baye's theorem. It is defined as a statistical classifier. It is one of the frequently used methods for supervised learning. It provides an efficient way of handling any number of attributes or classes which is purely based on probabilistic theory. Bayesian classification provides practical learning algorithms and prior knowledge on observed data. Let X is a data sample containing instances, Xi where each instances are the breast cancer risk factors (modifiable and non-modifiable). Let H be a hypothesis that X belongs to class C which contains (unlikely, likely and benign cases). Classification is to determine P(Hj|X), (i.e., posteriori probability): the probability that the hypothesis, Hj (unlikely, benign or likely) holds given the observed data sample X. ▯ P(Hj) (prior probability): the initial probability of the hypothesis in the class; P(Xi): probability that sample data is observed for each attribute, i; P(Xi|H) (likelihood): the probability of observing the sample's attribute, Xi given that the hypothesis holds in the training data X; and posteriori probability of a hypothesis Hj (unlikely, likely or benign), P(Hj|Xi), follows the Baye's theorem as follows:

For example, for a variable X with i attributes (breast cancer risk factors) expressed as:

X = {X1, X2, X3, X4… X1} and Hj= {unlikely, likely, benign}.

2. Decision Trees

J48 decision trees classifier is a simple decision learning algorithm, it accepts only categorical data for building a model. The basic idea of ID3 is to construct a decision tree by employing a top down greedy search through the given sets of training data to test each attribute at every node. It uses statistical property known as information gain to select which attribute to test at each node in the tree. Information gain measures how well a given attribute separates the training samples according to their classification.

It is suitable for handling both categorical as well as continuous data. A threshold value is fixed such that all the values above the threshold are not taken into consideration. The initial step is to calculate information gain for each attribute. The attribute with the maximum gain will be preferred as the root node for the decision tree.

Given a set S of breast cancer cases, J48 first grows an initial tree using the divide-and-conquer algorithm as follows: If all the cases in S belong to the same class or S is small, the tree is a leaf labeled with the most frequent class in S; Otherwise, choose a test based on a single attribute with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition S into corresponding subsets S1, S2,……, Sn for a dataset containing n cases according to the outcome for each case, and apply the same procedure recursively to each subset.

3. k-Nearest Neighbor (k-NN): K-Nearest Neighbor is a supervised machine learning algorithm as the data given to it is labeled. It is a nonparametric method as the classification of test data point relies upon the nearest training data points rather than considering the dimensions (parameters) of the dataset. It is employed in solving both classification and regression tasks. In Classification technique, it classifies the objects based on the k closest training examples in the feature space. The working principle behind KNN is it presumes that alike data points lie in same surroundings. It reduces the burden of building a model, adapting a number of parameters, or building furthermore assumptions. It catches the idea of proximity based on mathematical formula called as Euclidean distance, calculation of distance between two points in a plane. Suppose the two points in a plane are A(x0, y0) and B(x1, y1) then the Euclidean distance between them is calculated as follow. An object to be classified is allotted to the respective class which represents the greater number of its nearest neighbors. If k takes the value as 1, then the data point is classified into the category that contains only one nearest neighbor. Given a new input data point, the distances between that points to all the data points in the training dataset are computed. Based on the distances, the training set data points with shorter distances from the test data point are considered as the nearest neighbors of our test data. Finally, the test data point is classified to one of the classes of its nearest neighbor. Thus the classification of the test data point hinges on the classification of its nearest neighbors. Choosing the value of K is the crucial step in the implementation of KNN algorithm. The value of K is not fixed and it varies for every dataset, depending on the type of the dataset.

4. Support Vector machine

Support Vector Machine is a supervised machine learning algorithm which is doing well in pattern recognition problems and it is used as a training algorithm for studying classification and regression rules from data. SVM is most precisely used when the number of features and number of instances are high. A binary classifier is built by the SVM algorithm. This binary classifier is constructed using a hyper plane where it is a line in more than 3-dimensions.The hyper plane does the work of separating the members into one of the two classes. Hyper plane of SVM is built on mathematical equations. The equation of hyper plane is WTX=0 which is similar to the line equation y= ax + b. Here W and X represent vectors where the vector W is always normal to the hyper plane. WTX represents the dot product of vectors. As SVM deals with the dataset when the number of features are more so, we need to use the equation WTX=0 in this case instead of using the line equation y= ax + b. If a set of training data is given to the machine, each data item will be assigned to one or the other categorical variables. A SVM training algorithm builds a model that plots new data item to one or the other category.

5. Performance Evaluation:

The performance evaluation criteria allow the measurement of the accuracy of the models developed using the training dataset. The results of the classification are recorded on a confusion matrix. A confusion matrix is a square which shows the actual classification along the vertical and the predicted along the vertical. All correct classifications lie along the diagonal from the north-west corner to the south-east corner also called True Positives (TP) and True Negatives (TN) while other cells are called the False Positives (FP) and False Negatives (FN). If the unlikely case is considered positive then likely and benign are called negatives, if likely is considered as positive then unlikely and benign are considered negatives and the same also applies if benign is called the positive.

3. RESULTS

Sl. No	Algorithms	Accuracy	Precision
1.	Logistic Regression	92.10	95.31
2.	K-Nearest-Neighbor	92.23	96.55
3.	Support Vector Machine	92.78	95.78
4.	Decision Tree	92.62	96.33s

4. CONCLUSION

In this study four different classification techniques were used for the prediction of breast cancer risk and their performance was compared in order to evaluate the best classifier. Our work is mainly focused on advancement of predictive models to achieve good accuracy in predicting valid disease outcomes using supervised machine learning methods. The analysis of the results signify that the integration of multidimensional data along with different classification, feature selection and dimensionality reduction techniques can provide auspicious tools for inference in this domain.

REFERENCES

[1] Nidhi Mishra, Naresh Khuriwal- “Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm”, 2018 IEEMA Engineer Infinite Conference (eTechNxT), 2018.
[2] Yi-Sheng Sun, Zhao Zhao, Han-Ping-Zhu,”Risk factors and Preventions of Breast Cancer” International Journal of Biological Sciences.
[3] Vikas Chaurasia, BB Tiwari and Saurabh Pal – “Prediction of benign and malignant breast cancer using data mining’s techniques”, Journal of Algorithms and Computational Technology.
[4] S. Wang, H. Li, J. Li, Y. Zhang, and B. Zou, “Automatic analysis of lateral cephalograms based on multi resolution decision tree regression voting,” Journal of Healthcare Engineering, vol. 2018, Article ID 1797502, 15 pages, 2018.
[5] D.Jha, J. Kim, and G.Kwon,“Diagnosis of Alzheimer’s disease using dual-tree complex wavelet transform, PCA, and feed forward neural network,” Journal of Healthcare Engineering, vol. 2017, Article ID 9060124, 13 pages, 2017.
[6] Haifeng Wang and Sang Won Yoon – Breast Cancer Prediction using Data Mining Method, IEEE Conference paper.
[7] Tuba kiyan, Tulay Yildirim “Breast cancer diagnosis using statistical neural networks”, Journal of Electrical and Electronic Engineering.
[8] P.K.Chenniappan & N.Anusheela “Early detection of cancer cells to lung cancer survival data by neural network testing” International journal of neural network and application.