BREAST CANCER PREDICTION USING MACHINE LEARNING APPROACHES

B. Kranthi kiran
JNTUHCEH, Kukatpally, Hyderabad.
Email: kranthikiran9@gmail.com

https://doi.org/10.26782/jmcms.2019.12.00012

Abstract

In recent days the fast-growing disease in most of the world's is breast cancer especially in women and, according to global statistics, represents a different level of cases that are hitting cancer and illnesses associated with related diseases, rendering it a major public health issue currently in the community. The diagnosis and treatment for this significantly contributed by the machine learning techniques that can be applied for patient data to detect the cancer stage at earlier stages can help patients receive appropriate medical treatment. In this paper, four classification methods have been used in the context of Bayes Net, Adaboost, Simple Logistic and Stochastic Gradient Descent, successfully. The primary goal is to test in terms of accuracy, uncertainty matrix, MAE and RMSE, consistency in the identification of information concerning efficiency and effectiveness of each algorithm.

Keywords: Classification, Machine learning, Stochastic Gradient Descent, Breast cancer

I. Introduction

Most people are effectively affected by breast cancer. The origin of this condition relies on many variables and cannot be established clearly. This takes an outstanding commitment by physicians and nurses to assess whether or not the tumor is harmless or malignant [I]. But that was many tests, including clump thickness, cell size homogeneity, cell form uniformity, etc, are dealt with in identifying breast cancer, the ultimate result can be troubling including for physicians [II]. In recent years, the use of machine learning and computing usually as detection methods has increased. These techniques require a dataset for processing the model for classification of data accordingly, in general, the dataset consists of instances that are essential for designing a system.

Every instance gathers some collection with significant attributes that define particular specifications of the domain. The output attribute value is expected to depend on the input attribute values. The attribute and the value it is assigned defines a function, that helps make a particular case a characteristic vector. The template developed by an algorithm can be viewed concerning the given input data for mapping into the corresponding output. Previous research revealed the significance of the same subject of research [III-VII], where they suggested the use of ML algorithms
to identify breast cancer. Machine learning is widely recognized as the method of forecasting breast cancer with its advances in the detection of critical features from complex datasets. Implementation of data mining methods in the medical field can assist to forecast results, reduce the cost of drugs, assist the health of people, and enhance the value of health care and save people's lives. The best way to classify benign and malignant tumors is through the use of machine learning classification techniques.

Machine learning methods will, therefore, make it easy for doctors to recognize breast cancer properly and recognize it as a healthy or malignant tumor. There is little doubt that patient analysis and doctor and specialist selections are the most important factors in identification, but know how to build systems with AI facility, for determination of the class of a given data purposes, together that makes it easy for medical specialists greatly. In this article, the main aim is to greatly contrast different classification research algorithms to predict a harmless disease in breast cancer. This paper discusses the numerous machine learning methods and utilizes many strategies and refers to the dataset of breast cancer.

For the rest of the paper, the discussion is organized in the following manner. In Section II we reviewed the basics of preprocessing phases of the system. Section III presents the proposed method used to design the breast cancer prediction system. Section IV includes Results and discussion. Finally section V ends with conclusion.

II. Background works

Feature Vectors:
To create a predictive model, we need training samples that are more relevance for getting the output of the prediction accurately, with the help of ML procedures, which is capable of detecting the presence of cancer. A sample collection is used to assess the model's efficiency in which ML algorithms involve that the dataset function vector is made up of sample input attributes as well as output attributes with the predicted variable existence of the output attribute. In the training set, ML algorithms use information from the function vector for getting the dimensionality of the problem. The ML procedure matrix can be in the mathematical form as

\[ a_1, ..., a_n, b \]

Where \( a_i \) treated to be input attribute and \( b \) treated to be the output attribute.

Cross-Validation:
It is a method used for evaluating the output of an algorithm-built design. A subset of the data presented is left aside during cross-validation, while the data that left used for purpose of training collection, is being applied to the system that can process the remaining procedure of classification algorithm. The portion for the unused data
collection, the test data utilized to assess corresponding system output by calculating its reliability based on the model's identification of the sample case. The data set is split into n subsets of equal size in n-fold cross-validation. It can be is carried out n times, for every subset used precisely once as a sample array. The template outputs of each subset used as the trial range are summed to calculate the total algorithm efficiency. The benefit of using n-fold cross-validation at the stage of every instance for a relevant dataset should be recognized for the output of the system can be tested within a single data set. The use of 10 and 20-fold cross-validation for good predictions is suggestible. ML algorithms for breast cancer are qualified to use a database of Wisconsin Breast Diagnostic Cancer (WDBC) as shown in Figure 1.

![Figure 1. Breast cancer Images (a) Benign, (b) Malignant [VIII]](image)

III. Proposed system

In this paper, we address the influence of the proposed model as depicted in the Figure 2 that predicts whether the person is affected by cancer or not and its stage also. The first goal was to obtain a database with statistical values of different instances. When our dataset is finished, there is a process of splitting the given data to training and testing in the ratio of 70:30 for train and test of four algorithms: Bayes Net, Adaboost, Simple Logistic, and Stochastic Gradient Descent.

![Figure 2. Breast cancer prediction system](image)

**Classification Algorithms:**

**Bayesian Belief Networks (Bayes Nets):**

It is the process of considering an entity that a certain number of inputs are not all subject to conditional autonomy. It is a representation of the graphic model for...
provides cooperative dissemination of probability for a collection of attributes. For example to explain its process, let every attribute is described in the network as a node. In the network, the link from node X to node Y is directed if X is parent Y, meaning that Y depends on X, or X node could affect the node Y. In other words, the node with some attribute depends on another node independently, by marking the parent node and gets state of it at that instance. Such effects seem to be expressed with the likelihood of quality at a node which is contingent upon the parent’s interest. Such likelihood values for a node are organized in a table format known as the CPT graph. To nodes without relatives, the CPT distributes the attribute at that node.

Adaboost Classifier:
It's another ensemble classification consisting of multiple classification algorithms of which the output is coupled with the output of those classification algorithms. This iteratively retrain the algorithm by choosing the training array depending on the accuracy of the previous training. The weight cost of each trained classifier relies on the exactness of each iteration.

Logistic regression:
Regression algorithms rely basically on the input data attributes that are sequentially supplied to the system for generating the required output data with the help of mathematical functions. Logistic regression is a mathematical formalism in which such connections are represented. In general, the data for simple logistic regression are n independent cases, each of which consists of the descriptive variable x, and a trial success or failure. For example, x can be the predictor for breast cancer, and success means this tumor is considered to be cancerous. Each assessment may have a value of x different. Nevertheless, it is practical, to begin with, a special case in which the explanatory variable x is a Yes-or-No variable, to incorporate logistic regression. The information then contains a number of outcomes for each of the two meanings of x (success or failure). There are only two p values, one for each x value.

Stochastic Gradient Descent (SGD):
In this technique, several samples are randomly selected instead of the complete data set for each iteration. In the gradient descent, the term "batch" refers to the total number of samples in a data set used to calculate the gradient for each iteration. In standard SGD optimization, the batch is considered the whole data set like the Batch Gradient Descent. Although it is useful to use the whole dataset to achieve minima in a less random or less random way, the problem arises if our datasets are very large. One thing to understand is that since SGD is generally noisier than a typical gradient decrease, due to the randomness of its descent, it usually took more iteration to reach the minima. Although it takes a greater number of iterations to reach the minimum point than the traditional descending gradient, it is still much simpler computationally than a standard descending gradient. Therefore, SGD is preferred over Batch Gradient Descent for optimizing the study algorithm in most scenarios.

IV. Simulation Results and Analysis
Numerous quality metrics were used in this study to determine the performance with the associated ML techniques. Similarly, the carried work in this
paper majorly focuses on the classification problem and its performance metrics are evaluated to check the strength of the process. Breast cancer prediction is positive, this means that the person affected with BC as a variable has the value of 1 (malignant). And if 0 (benign) is the target variable, it is a negative instance that says the patient would not suffer from cancer.

Confusion Matrix (CM): It is the process of investigating for a classification model to be governed by equating the correct/incorrect classification of the number of positive instances and the correct/incorrect classification of the many negative instances shown in Table 1.

| Actual Negative | Predicted Negative | Predicted Positive |
|-----------------|--------------------|--------------------|
| Actual Negative | TN                 | FP                 |
| Actual Positive | FN                 | FP                 |

Check for the effectiveness of ML techniques by applying the required dataset with attributes that are run on the suitable test dataset. The test data set included 30 percent of a prediction for breast cancer. It was tested for cross-validation of 10-fold to predict the BC. Several metrics shown in Table 2 have been used for evaluating the parameters of the classifier of four techniques.

| Classifier name          | MAE     | RMSE    | Accuracy  |
|--------------------------|---------|---------|-----------|
| Bayes Net                | 0.3297  | 0.4566  | 72.02%    |
| Adaboost                 | 0.3526  | 0.4329  | 70.27%    |
| Simple Logistic          | 0.3589  | 0.4291  | 75.17%    |
| Stochastic Gradient Descent | 0.3007  | 0.5484  | 69.93%    |

Figure 3. Comparison of MAE, RMSE for different ML classifiers
V. Conclusion

In this paper, there is an explanation of various ML methods and solicitations used in BC diagnostics and prognosis to analyze the attributes for getting the class of BC by applying the dataset. The strategies adopted by ML techniques demonstrated the noteworthy potential for the benefits of the algorithm for better accuracy for prediction. The classification accuracy is an important criterion for evaluation, but it is not the only criterion. Also, the schemes produce different outcomes because of different approaches followed by respective algorithms. We can conclude that Simple Logistics performs better in the prediction of breast cancer for this dataset.

References

I. B.Kranthi kiran, Padmaja.Pulicherla, Classification and Enrichment of Unlabeled Feedback Data using Machine Learning. International Journal of Engineering and Advanced Technology, ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019.

II. Chen, W.; Zheng, R.; Baade, P.D.; Zhang, S.; Zeng, H.; Bray, F.; Jemal, A.; Yu, X.Q.; He, J. Cancer statistics in China, 2015. CA Cancer J. Clin. 2016, 66, 115–132.

III. Dr. Padmaja.Pulicherla, Job Shifting Prediction and Analysis Using Machine Learning(2019), et al 2019 J. Phys.: Conf. Ser. 1228 012056.
IV. Elias Zafiropoulos, Ilias Maglogiannis, and Ioannis Anagnostopoulos. 2006. A support vector machine approach to breast cancer diagnosis and prognosis. *Artificial Intelligence Applications and Innovations* (2006), 500–507.

V. Gouda I Salama, M Abdelhalim, and Magdy Abd-elghany Zeid. 2012. Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC) 32, 569 (2012), 2.

VI. Padmaja.Pulicherla, Image Map: Alternative for Password Based Authentication, International Journal of Recent Technology and Engineering, ISSN: 2277-3878, Volume-8 Issue-3, September 2019.

VII. Padmaja.Pulicherla, Retrieving Songs By Lyrics Query Using Information Retrieval, International Journal of Engineering and Advanced Technology, ISSN: 2249 – 8958, Volume-8 Issue-6S, August 2019.

VIII. Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. 2011. The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering* 13, 2 (2011), 22–30.

IX. William H Wolberg, W Nick Street, and Olvi L Mangasarian. 1992. Breast cancer Wisconsin (diagnostic) data set. *UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/]* (1992).

X. Wenbin Yue, Zidong Wang, Hongwei Chen, Annette Payne Xiaohui Liu, “Machine Learning with Applications in BreastCancer Diagnosis and Prognosis”, mdpi design, May 2018.

XI. Yichuan Tang. 2013. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239* (2013).