COMPARATIVE ANALYSIS OF CLASSIFICATION APPROACHES FOR BREAST CANCER

Temesgen Abera Asfaw
Department of Computer Science & Engineering, College of Engineering, Jawaharlal Technological University, Telangana, India

ABSTRACT

Breast cancer is one of the greatest common diseases among women in Africa and worldwide. Accurate and early diagnosis is very significant phase in therapy and action. However, it is not an easy one due to some doubts in detection of breast cancer. Machine learning helps us to extract information and knowledge from this the basis of past experiences and detect hard-to-perceive pattern from large and noisy dataset. This paper compares and analysis the performance of machine learning algorithms, namely Decision Tree (DT), Logistic Regression (LR), Naïve Bayes (NB), and K-Nearest Neighbors (KNN) for detecting breast cancer. The data set used for comparison was from UCI Wisconsin original breast cancer data set. The result outcome shows that Logistic Regression performs better and classification accuracy is 96.93%.

Key words: Machine learning, Logistic Regression, K-Nearest Neighbors, Gaussian Naïve Bayes, Decision Tree, and Heart Disease.

Cite this Article: Temesgen Abera Asfaw, Comparative Analysis of Classification Approaches for Breast Cancer, International Journal of Computer Engineering and Technology 10(4), 2019, pp. 10-16. http://www.iaeme.com/IJCET/issues.asp?JType=IJCET&VType=10&IType=4

1. INTRODUCTION

Most of women are dying due to Breast Cancer around the world nowadays. The advent of new medical skills and the huge amount of patient data have motivated the path for the development of new plans in the prediction and detection of cancer. The volume of data generating is fast in the field of biomedical [1]. This data deliver a rich source of information for medical research. Machine learning helps us to extract information and knowledge from this the basis of past experiences and detect hard-to-perceive pattern from large and noisy dataset. This paper compares different machine learning classification algorithms for breast cancer namely Logistic Regression (LR), Naïve Bayes (NB), Decision Tree (DT) and K-Nearest Neighbors (KNN). Choosing the most accurate classifier between frequent classifiers is one of the focuses in model selection. Selecting the most accurate classifier is the main aim this research. As statistics indicated by World Health Organization, the bulk of people are
died by cancer and the number of death is increasing every year [2]. The major cause of death among female is breast cancer [3]. The majority population of breast cancer includes women in age 40 and above while it can occur in some younger age women with risky features [4]. Since, traditional techniques cannot success to processes huge data in medicine, so it is important to have computational tools for correct decisions. This experiment indicated Logistic Regression (LR) performs better than other classification approaches of machine learning.

2. RELATED WORK
This unit deals with the research work approved out in medical field related to prediction of breast cancer by using machine learning algorithms. Mangasarian et al. [5] use clinical data points to estimate the time of recurrence to forecast the cancer. Some of the studies [6] [7] [8] presented work allied to disease prediction and diagnosis using machine learning techniques, like decision tree for detection of cancer. According to Jin [9], KNN algorithm is one of the most often used classification algorithm in machine learning techniques due to its ease and resourcefulness in implementation. Belciug et al. [10] presented a comparative study of cluster network, Self Organizing Map and K-means in the detection of breast cancer, using the Wisconsin Prognostic Breast Cancer (WPBC) dataset [11] in which k-means performed batter. Chaurasia and Pal [12] examined the performance of artificial neural networks (ANNs), Logistic Regression (LR), and Dyadic decision trees (DDTs) in breast cancer recurrence prediction using the Breast Cancer Dataset. Angeline [13] compared the performance of Naïve Bayes, Decision tree (C4.5), K-Nearest Neighbor and Support Vector Machine to find the preeminent classifier in Wisconsin Breast Cancer (WBC) to predict the primary site of cancer. As per the analysis, SVM performs better than other. Abonyi [14] used fuzzy clustering algorithms in order to detect cancer on Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Lavanya et. al. [15] uses a hybrid and dynamic approach to improve the classification accuracies of WDBC dataset to a relatively higher degree with 10 fold cross validation. As per the analysis [16], cross validation is most effective methodologies to estimate the performance of a ML model on a validation set. Cruz [17] surveyed machine learning techniques and algorithms in cancer diagnosis and prognosis. Tan A.C. [18] performed bagged decision tree, C4.5 decision tree on micro array data of cancer and presented the performance analysis. Tsirogiannis [19] applied bagging techniques using decision trees, neural networks and SVM on medical databases. As per the analysis, bagging techniques show better accuracy.

3. MATERIALS
In this experimentation, we have used openly available breast cancer dataset from the University of Wisconsin Dr. William H. Wolberg [20] and taken from UCI machine learning repository (http://archive.ics.uci.edu/ml/). It includes 698 examples taken from fine needle articolates of human breast tissue. Since 16 instances of the dataset have missing information, 683 instances are used in our experiment, including 444 and 239 instances respectively belong to benign or not harmful and malignant or may be harmful classes. Every instance has nine attributes as shown in Table I [20].The dataset have two kinds of classification for breast cancer and its total number of Malignant (M) and Benign (B) Tumors are showed at fig. 1 below. It is always good to check the correlations between the features. From the figure below, the red around the diagonal advocates that features are correlated with each other. The yellow and green covers advocate some reasonable correlation and the blue boxes demonstration negative correlations as shown below fig. 2.
Breast cancer is one of observable disease in the area of medical diagnosis which is growing all year. A comparative analysis of most widely used five machine learning techniques is done on Wisconsin Breast Cancer Dataset to predict the breast cancer recurrence:

4.1. Logistic Regression (LR)

uses an equation related to Linear Regression but the result of logistic regression is a categorical variable whereas it is a value for other regression models. Binary results can be predicted from the independent variables. The result of dependent variable is distinct. Logistic Regression uses a artless equation which displays the linear relation between the independent variables. These independent variables sideways with their coefficients are combined linearly to form a linear equation that is castoff to predict the output [21]. Logistic Regression model equation

\[ \ln (\pi/1-\pi) = a_0 + a_1 x + a_2 x \] (1)
This is named the logistic function as the key method behind it is logistic function. The output can be expected from the independent variables, which form a linear equation. The output forecast has no limitations; it can be any value from negative infinity to positive infinity. But the output obligatory is a class variable (i.e., yes or no, 1 or 0). So, the outcome of the linear equation must be trampled into a minor range. Logistic function is used here to suppress the outcome value between 0 and 1.

4.2. Naïve Bayes (NB)
Bayesian classifier which is recognized as humble probabilistic classifiers. By using this technique, the probability of an occurrence which fits to a particular class is predicted. Entirely the features are supposed independent according to Bayes theorem which means there is no dependency amid the attribute value on a given class and the other attributes [22]. The modest method of Naïve Bayes classifier is presented in (3). According to this formula, C states the training instance set that is separated into K classes $C = \{C_1, C_2, ..., C_k\}$; the first probability of each classes stated as $p(C_j)$, where $j=1,2,...,k$. For a given instance illustrated as $d_i=(W_1, ..., W_j, ..., W_m)$ while feature words are stated as $W_j$, where $j=1,2,...,m$ belongs to a particular class $C_j$. In order to classify the document $d_i$, the probability of all the documents for a given $d_i$ should be estimated [23]. The posterior probability of class $C_j$ is estimated as it is shown

$$P(C_j|d_i) = P(D_i|C_j)P(C_j)/P(D_i) \quad \text{(2)}$$

4.3. Decision Tree (DT)
In DT each interior node represents a experiment on a feature, each branch means a result of the test and each leaf node represents a class label. The decision tree comprises two parts. First, in growing phase, based on recursive process and local optimal criteria, the training set is splitting until all or most of the outcomes of each partition have the same class label. The drawback of this method for building a tree is may over fitting happen. In order to handle the over fitting problem, pruning streamlines the tree with removing outliers to increase the accuracy. And the time for pruning the decision tree is very less than building decision tree using growth phase [24].

4.4. K-Nearest Neighbors (KNN)
classifies test data directly using training data. In order to classify any test data; it primarily calculates K value, which denotes the number of Nearest Neighbors. For all test data, it calculates the remoteness between all the training data and then sorts the remoteness. Then by using majority elective, class label will be chosen to the test data. The Euclidean distances are calculated as below:

$$\text{Distance}(p, q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)} \quad \text{(3)}$$

5. EXPERIMENTAL RESULT
A set of experiments were performed using pycharm 3.6 IDE in python programming language. To diagnosis breast cancer, four methods Decision Tree (DT), Logistic Regression (LR), Naïve Bayes (NB), and K-Nearest Neighbors (KNN) are performed. The comparative analysis of these machine learning techniques is performed in terms of Accuracy and we compared the dataset using standardization. With regular algorithm checking the initial run, it looks like DT performed the best given the dataset (all above 92% mean accuracy). Support Vector Machine has a surprisingly bad performance here. However, if we standardize the input dataset, its performance should improve. From the initial run, it looks like DT
performed the best given the dataset 92% mean accuracy. However, if we standardize the input dataset, its performance should advance.

**Figure 2** Performance comparison before standardization

**Table 1** Performance measure of machine learning techniques without using Standardization

| Classification Algorithms | Accuracy |
|---------------------------|----------|
| Decision Tree (DT)        | 92%      |
| Logistic Regression (LR)  | 62%      |
| K-Nearest Neighbors (KNN) | 74%      |
| Naïve Bayes (NB)          | 61%      |

Evaluation result of algorithm using Standardized Dataset shows that Logistic Regression performs better comparison result than other algorithms. We have used pipelines to standardize the data and build the model for each fold in the cross-validation test yoke. Finally we have got impartial estimation of how each model with standardized data might perform on unseen data.

**Figure 3** Performance comparison after standardization
Comparative Analysis of Classification Approaches for Breast Cancer

Table 2 Performance measure of machine learning techniques using Standardization

| Classification Algorithms | Accuracy |
|---------------------------|----------|
| Scaled Decision Tree (ScaledDT) | 90.5% |
| Scaled Logistic Regression (ScaledLR) | 96.9% |
| Scaled K-Nearest Neighbors (ScaledKNN) | 95.8% |
| Scaled Naïve Bayes (ScaledNB) | 92.9% |

6. CONCLUSION
Core motivation of the research work accompanied is to advance the prediction of breast cancer in demand to increase the accuracy of diagnosis. Most of the studies are presented which have been planned in several years and emphasis on the advance of predictive models for breast cancer prognosis using machine learning approaches and classification. The comparative analysis of four widely used machine learning methods Logistic Regression (LR), Naïve Bayes (NB), Decision Tree (DT) and K-Nearest Neighbors (KNN) are performed. The performance in terms of accuracy, Logistic Regression is better as compared to other methods.

REFERENCES
[1] Hanahan, Douglas, and Robert A. Weinberg. "Hallmarks of cancer: the next generation." cell 144.5 (2011): 646-674.
[2] WHO. WHO | Cancer. World Health Organization. Retrieved July 05, 2014, from http://www.who.int/topics/cancer/en/
[3] Cancer Facts & Figures 2010 | American Cancer Society. (n.d.). Retrieved November 19,2014 from http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2010/index
[4] Tawam Hospital | Medical News. (n.d.). Retrieved November 19, 2014, from http://www.tawamhospital.ae/english/news/print.aspx?NewsID= 367
[5] Mangasarian, Olvi L., W. Nick Street, and William H. Wolberg. "Breast cancer diagnosis and prognosis via linear programming." Operations Research 43.4 (1995): 570-577.
[6] Zhou, Zhi-Hua, and Yuan Jiang. "Medical diagnosis with C4. 5 rule preceded by artificial neural network ensemble." IEEE Transactions on information Technology in Biomedicine 7.1 (2003): 37-42.
[7] Lundin, Mikael, et al. "Artificial neural networks applied to survival prediction in breast cancer." Oncology 57.4 (1999): 281-286.
[8] Delen, Dursun, Glenn Walker, and Amit Kadam. "Predicting breast cancer survivability: a comparison of three data mining methods." Artificial intelligence in medicine 34.2 (2005): 113-127.
[9] Marsilin, Jini R., and G. Wiselin Jiji. "An efficient cbir approach for diagnosing the stages of breast cancer using knn classifier." Bonfring International Journal of Advances in Image Processing 2.1 (2012): 1.

[10] S. Belciug, F. Gorunescu, A. B. Salem, and M. Gorunescu., “Clustering-based approach for detecting breast cancer recurrence” In Proceedings of the International Conference on Intelligent Systems Design and Applications (ISDA), (2010):533–538.

[11] M. Lichman. 2015. UCI Machine Learning Repository. Retrieve from http://archive.ics.uci.edu/ml (Accessed November, 2016)

[12] Chaurasia, Vikas, and Saurabh Pal. "Data mining techniques: To predict and resolve breast cancer survivability." International Journal of Computer Science and Mobile Computing 3.1 (2014): 10-22.

[13] Christobel, Angeline, and Y. Sivaprakasam. "An empirical comparison of data mining classification methods." International Journal of Computer Information Systems 3.2 (2011): 24-28.

[14] Abonyi, Janos, and Ferenc Szeifert. "Supervised fuzzy clustering for the identification of fuzzy classifiers." Pattern Recognition Letters 24.14 (2003): 2195-2207.

[15] Lavanya, D., and K. Usha Rani. "Ensemble decision tree classifier for breast cancer data." International Journal of Information Technology Convergence and Services 2.1 (2012): 17.

[16] http://www.openml.org/a/estimation-procedures/1(Accessed October, 2016)

[17] Cruz, Joseph A., and David S. Wishart. "Applications of machine learning in cancer prediction and prognosis." Cancer informatics 2 (2006).

[18] Tan, Aik Choon, and David Gilbert. "Ensemble machine learning on gene expression data for cancer classification." (2003).

[19] Tsirigannis, G. L., et al. "Classification of medical data with a robust multi-level combination scheme." Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on. Vol. 3. IEEE, (2004).

[20] W. 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, vol. 87, pp. 9193-9196, 1990.

[21] Chao-Ying Joanne, PengKukLida Lee, Gary M. Ingersoll –“An Introduction to Logistic Regression Analysis and Reporting “, September/October 2002 [Vol. 96(No. 1)]

[22] I. Kononenko, “Machine learning for medical diagnosis: history , state of the art and perspective,” vol. 23, 2001.

[23] L. Yuan, "An Improved Naive Bayes Text Classification Algorithm In Chinese Information Processing." Science, pp.267–269, 2010.

[24] D.Lavanya, & D.K.U. Rani, "Performance evaluation of decision tree classifiers on medical datasets," International Journal of Computer Applications, vol.26(4). 2011.

http://www.iaeme.com/IJCET/index.asp

editor@iaeme.com