www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 6 Issue 2 Feb. 2017, Page No. 20388-20391
Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i2.40

Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree

Subrata Kumar Mandal
Information Technology Department,
Jalpaiguri Government Engineering College,
Jalpaiguri, West Bengal, India
mandal.skm@gmail.com

Abstract: **Breast cancer is one of the second leading causes of cancer death in women. Despite the fact that cancer is preventable and curable in primary stages, the vast number of patients are diagnosed with cancer very late. Established methods of detecting and diagnosing cancer mainly depend on skilled physicians, with the help of medical imaging, to detect certain symptoms that usually appear in the later stages of cancer. The objective of this paper is to find the smallest subset of features that can guarantee highly accurate classification of breast cancer as either benign or malignant. Then a relative study on different cancer classification approaches viz. Naïve Bayes(NB), Logistic Regression(LR), Decision Tree(DT) classifiers are conducted where the time complexity of each of the classifier is also measured. Here, Logistic Regression classifier is concluded as the best classifier with the highest accuracy as compared to the other two classifiers.**

Keywords: Breast Cancer, Classification Accuracy, Feature Extraction, Supervised machine learning, benign, cancer classification, malignant..

1. Introduction

1.1 Preliminaries

In breast cancer, cancer cells form in the tissues of the breast of the woman. The breast is formed up of lobes containing 15 to 20 sections and ducts. The most usual type of breast cancer begins in the cells of the ducts. Cancer that begins in the lobes or globules found in both breasts are other types of breast cancer. Warm, red, and swollen breast is an indicator for breast cancer. Age and health history can affect the risk of getting breast cancer. For discovering the different stages of the breast cancer, Chest X-ray, CT scan, Bone scan and PET scans are extensively used. The number of breast cancer disease is calculated to be 1.2 million among women every year according to projections by the World Health Organization. In the year 2014 an estimate of 2,32,714 new breast cancer incidences happened in women, whereas a total of 2,97,800 female patients died due to cancer in which 16.1% of the total death was in breast cancer occurred within the US. Since the early years of cancer research, biologists have used the traditional microscopic technique to assess tumor behavior for breast cancer patients. For the diagnosis and treatment of cancer, precise prediction of tumors is critically significant. The Latest machine learning techniques are progressively being used by life scientists to obtain appropriate tumor information from the databases. Among the existing techniques, supervised machine learning methods are the most popular in cancer diagnosis.

1.2 Basic Concepts used in Cancer Cell Detection

In this research paper we have used Lasso regression for feature extraction as well as feature selection and Gaussian Naïve Bayes, Logistic regression, Decision Tree, classifiers for cancer classification. These concepts are discussed as follows:

1.2.1 **Pearson Correlation Coefficient (PCC)**

It is a feature selection procedure which is employed to measure the strength of a linear connection between two variables, where the value of correlation coefficient r = 1 implies a perfect positive correlation and the value r = -1 implies a perfect negative correlation. Correlation between sets of data is a measure of how perfect they are associated to each other. The most common quantification of correlation in Statistics is the Pearson Correlation Coefficient. The coefficient value lies between -1 and 1.
The formula for Pearson Correlation Coefficient, ρ is:

$$\rho_{X,Y} = COV(X,Y) / \sigma_X \sigma_Y$$

Where:

- COV is the covariance.
- σ_X is the standard deviation of X.
- σ_Y is the standard deviation of Y

1.2.2 **Logistic Regression Classifier**

Logistic regression classifier is another technique borrowed by machine learning from the field of statistics.It is the go-to method for binary classification problems (problems with two class values).

1.2.3 **Naïve Bayes (NB) Classifier**

Naïve Bayes classifier is a statistical classifiers which can predict class membership probabilities such that the probability of a given tuple falls into a particular class. Naïve Bayes classifier is based on Bayes' theorem.

1.2.4 **Decision Tree**

An ensemble classifier blends a series of k learned models (or base classifiers), M1, M2,...,Mk, with the objective of designing an improved hybrid classification model, M*. D is the given data set which is used to create k training sets, D1,D2,...,Dk, where Di ($1 \leq i \leq k-1$) is likewise used to generate the classifier Mi.Given a new data tuple to classify, each of the base classifiers vote by returning a class prediction. Grounded on the votes of the base classifiers an ensemble returns a class prediction. An ensemble classifier can predict more accurate answer than its base classifiers.

2. Literature Survey

In this paper[1] authors demonstrate the comparison of different classification techniques like Bayes Network, Radial Basis Function, Pruned Tree and Nearest Neighbors algorithm using Waikato to Environment for Knowledge Analysis (WEKA) on large dataset. The data utilized in their research is the breast cancer data. It holds a total of 6291 data and a dimension of 699 rows and 9 columns. In this 75% of overall data is used training and the remainder is used for testing the accuracy of classification technique. Agreeing to the simulation result, highest accuracy is 89.71% which owe support to bayes network with the minimum time taken to build the model is 0.19 seconds and lowest average error is 0.2140 compared to others.

In the paper[2] prediction of the breast, cancer disease was done through Artificial Neural Network (ANN), Logistic regression, Naive Bayes techniques.The target of the research aims at giving the following results; firstly, it evaluates medical data set in terms of quality grammatically and secondly, it evaluates data mining methods with respect to their applicability to the data. Eventually, the knowledge drawn out from the data set is used for disease prediction by applying Artificial Neural Network (ANN), Logistic Regression, Naïve Bayes. It is found that these methods had highest lifting factor for most of the class values.

In the paper[3] the main focus of this work is in the analysis of breast cancer classification and prediction so that preventive measures can be made at an early stage before the onset of the breast cancer. Different data mining techniques such as Decision Tree, Clustering Algorithm are employed to achieve the objective. Observation mining technique common sensual on the data base fly in the ointment relationship and traditions takes are helpful in studying the progression of the disease.

Naive Bayes was used as a classifier in [4], and it yielded an accuracy of 96.6%

Some tools are giving impressive results as [5] when used RapidMiner to build an SVM classifier and achieved an accuracy of only 80%.

[6] built a hybrid classifier of Support Vector Machines and Decision Trees in WEKA resulting in 91% accuracy.

[7] had applied the supervised fuzzy clustering technique and reported an accuracy of 95.57%.
In the paper [8] by Leena Vig had presented an analysis using Random Forest classifiers, Artificial Neural Networks, Naïve Bayes a and Support Vector Machines. Results show that ANN's, Random Forests and SVMs are able to yield models with high accuracy, sensitivity and specificity whereas Naïve Bayes performs poorly.

In the paper [9] authors found the smallest subset of features from Wisconsin Diagnosis Breast Cancer (WDBC) dataset by applying confusion matrix accuracy and 10-fold cross validation method that can ensure highly accurate ensemble classification of breast cancer as either benign or malignant. For classification, the breast cancer data were first classified by Support Vector Machine (SVM) and Naïve Bayes classifiers and then finalize the classification process.

In this paper [21] by Diana Dumitru the Naive Bayes classifier was applied to the Wisconsin Prognostic Breast Cancer (WPBC) dataset, containing a number of 198 patients and a binary decision class: non-recurrent-events having 151 instances and recurrent-events having 47 instances. The testing diagnosing accuracy, that was the main performance measure of the classifier, was about 74.24%, in compliance with the performance of other well-known machine learning techniques.

In [22] A. Soltani Sarvestani, A. A. Safavi, N.M. Parandeh and M.Salehi provided a comparison among the capabilities of various neural networks such as Self Organizing Map(SOM), Multilayer Perceptron (MLP), Radial Basis Function (RBF) and Probabilistic Neural Network(PNN) which are used to classify WBC and NHBCD data. The performance of these neural network structures was investigated for breast cancer diagnosis problem. In the training set, PNN and RBF were proved as the best classifiers. But the PNN gave the best classification accuracy when the test set is considered. This Research work showed that statistical neural networks can be effectively used for breast cancer diagnosis as by applying several neural network structures a diagnostic system was constructed that performed quite well.

3. Proposed Work

Today's real-world databases are highly vulnerable to noisy, missing and inconsistent data due to their typically massive size and their likely origin from multiple, miscellaneous sources. Hence data preprocessing is a necessary phase for classification purposes. Data preprocessing includes data cleaning, data dimensionality reduction, data transformation (data normalization, data binning) followed by classification.

Here, we have taken WDBC breast cancer dataset from UCI machine learning repository [10] as an input data. Our data cleaning technique includes removing the missing values if present, with the mean of the attributes. Data normalization brings the range of all attribute values between 0 and 1. In the following workflow diagram we represent breast cancer cell detection using Pearson Correlation Coefficient as a feature selection technique.By using PCC we have selected four dominat features i.e radius_se,area_se, concave points_worst, radius_worst attributes.

This WDBC dataset contains 569 instances and 32 attributes of which we have taken 70 percentage instances for training purpose and 30 percentage instances for testing purpose. These testing data are applied over three classification methods which detect whether the cell is malignant or benign.

DOI: 10.18535/ijecs/v6i2.40

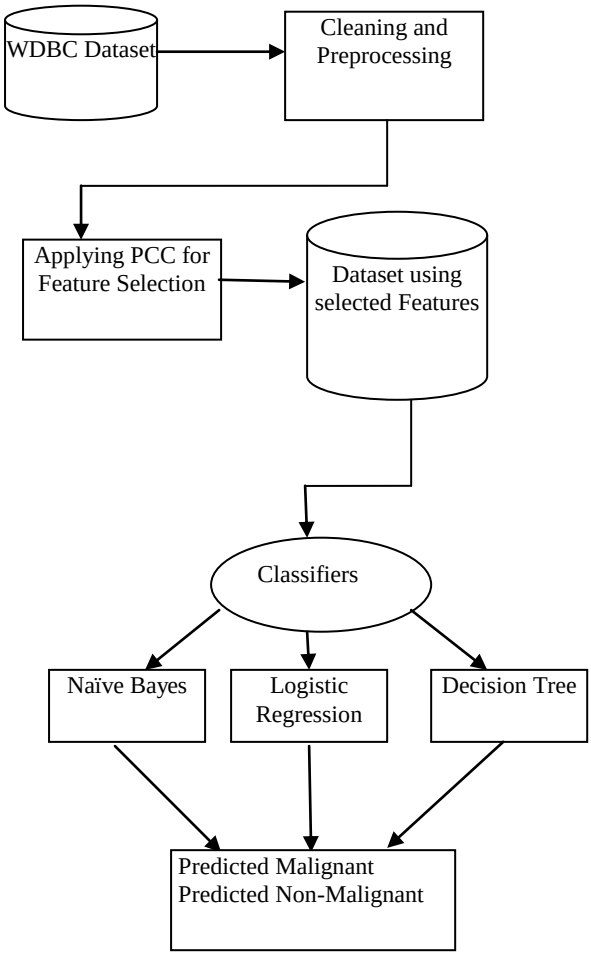

Fig 1: Workflow diagram for breast cancer cell detection using Pearson Correlation Coefficient.

4. Result and Discussion

In this paper we have conducted a comprehensive study on different classification techniques and provided a basis for comparison among them in terms of accuracy percentage and time complexity.

Table 1. Classification accuracies using the result of

Sl. No.	Name of Algorithm	Classification Accuracy (%)
1.	Naïve Bayes	94.40
2.	Logistic Regression	97.90
3.	Decision Tree	96.50

Here we use confusion matrix which is a table that is often used to describe the performance of a "classifier" or classification model on a collection of test data for which the true values are known. The level of effectiveness of the classification model is calculated with the number of incorrect and correct classification in each possible value of the variable being classified in the confusion matrix. In our testing dataset there are 102 malignant data out of which 100 have been predicted correctly and only 2 are predicted wrongly. Also, there are 68 benign data of which 2 are predicted wrongly and 66 are predicted correctly. This analysis is shown in Table 2.

Table 2. Confusion matrix of the dataset obtained using Logistic Regression Classifier

		Predicted	
		Malignant	**Benign**
Actual	**Malignant**	100	2
	Benign	2	66

Table 3. Execution time of the classification algorithms

Sl. No.	Name of Algorithm	Execution Time (in µs)
1.	Naïve Bayes	0.0137357
2.	Logistic Regression	0.0141579
3.	Decision Tree	0.0150908

In Table 3 we have obtained the execution times of each of the three classification algorithms. The execution times for Naïve Bayes, Logistic Regression and Decision Tree algorithms are 0.0137357, 0.0141579 and 0.0150908 microseconds respectively.

5. Conclusion

Comparing to all other cancers, breast cancer is one of the major causes of death in women. So, the early detection of breast cancer is needed in reducing life losses. In this paper we have applied techniques namely data cleaning, feature selection, feature extraction, data discretization and classification for predicting breast cancer as accurately as possible. Our study reveals that Logistic Regression Classifier gives the maximum accuracy with reduced subset of features (four) and time complexity of this algorithm is least compared to other two classifiers. This work can further be enhanced by identification of particular stage of breast cancer, can be done in near future.

References

[1] Mohd,F.,Thomas,M.,2007.Comparison of different classification techniques using WEKA for Breast cancer.IFMBE proceedings 15:520-523.

[2] 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.

[3] 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).

[4] 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).

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

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

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

[8] LeenaVig, "Comparative Analysis of Different Classifiers for the Wisconsin Breast Cancer Dataset", Open Access Library Journal,Volume 1 | e660,2014.

[9] Kathija, Shajun Nisha ,"Breast Cancer Data Classification Using SVM and Naïve Bayes Techniques", International Journal of Innovative Research in Computer and Communication Engineering- Vol. 4, Issue 12, December 2016.

[10] https://archive.ics.uci.edu/ml/machine-learning-databases /breast cancer-wisconsin/wdbc.data.

[11] http://www.nationalbreastcancer.org/breast-cancer-facts

[12] Daniele Soria, Jonathan M. Garibaldi, Elia Biganzoli, Ian O. Ellis,"A comparison of three different methods for classification of breast cancer data."Machine Learning and Applications, 2008. ICMLA'08.Seventh International Conference on. IEEE, 2008.

[13] Jiawei Han, Jian Pei, Micheline Kamber "Data Mining Concepts and Techniques", Third Edition, Elsevier Inc, 2012, ISBN:978-0-12-381479-1.

[14] Kim W, Kim KS, Lee JE, Noh DY, Kim SW, Jung YS, Park MY, Park RW, "Development of novel breast cancer recurence prediction model using support vector machine." Journal of breast cancer 15.2 (2012): 230-238.

[15] C. D. Katsis, I. Gkogkou, C.A. Papadopulos, Y.Goletsis, P. V. Boufounou, G. Stylios "Using artificial immune recognition systems in order to detect early breast cancer." International Journal of Intelligent Systems and Applications 5.2 (2013): 34.

[16] H. Li, G. Hong and Z.Guo, "Reversal DNA methylationpatterns for cancer diagnosis",2014 8th International onference on Systems Biology (ISB), IEEE, (2014)

[17] S. S. Shrivastava, V. K. Choubey and A.Sant, "Classification Based Pattern Analysis on the Medical Data in Health Care Environment",International Journal of Scientific Research in Science, Engineering and Technology, vol. 2,no.1, (2016)

[18] 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)

[19] Mehmet Fatih Akay,"Support vector machines combined with feature selection for breast cancer diagnosis." Expert Systems with Applications 36 (2009) 3240–3247.

[20] 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)

[21] Diana Dumitru,"Prediction of recurrent events in breast cancer using the Naive Bayesian classification."*Annals of the University of Craiova-Mathematics and Computer Science Series* 36.2 (2009): 92-96.

[22] Sarvestan Soltani A, Safavi A. A., Parandeh M. N. and Salehi M., "Predicting Breast Cancer Survivability using data mining techniques,"Software Technology and Engineering (ICSTE), 2nd International Conference, 2010, vol.2, pp.227-231.