Classification and Prediction of Breast Cancer using Linear Regression, Decision Tree and Random Forest

S. Murugan
Professor, Department of Computer Science and Engineering
Sathyabama University
Chennai, India.
snmurugan@gmail.com

B. Muthu Kumar
Professor, Department of Computer Science and Engineering
Sathyabama University
Chennai, India.
anbmuthusba@gmail.com

S. Amudha
PG Student, Department of Computer Science and Engineering
Sathyabama University
Chennai, India.
amudha17s@gmail.com.

Abstract— **Breast Cancer is one of a major issue that some of the women are facing today. Earlier detection of cancer by performing detailed analysis based on the existing records which may assist the physicians in providing a better treatment to their patients. Data to analyze and predict the breast cancer are obtained from UCI Machine Learning Repository (Wisconsin Breast Cancer). The main objective is to classify whether the type of cancer is benign or malignant. Based on the available data set and the patient record, whether the disease is curable or non-curable is predicted. Thus the success rate of classification is 84.14% and the prediction percentage is 88.14%.**

Keywords—breast cancer, analysis, treatment, benign, malignant.

I. INTRODUCTION (*Breast cancer*)

Breast cancer is the most common cancer in women worldwide, with nearly 1.7 million new cases diagnosed in 2012 (second most common cancer overall). This represents about 12% of all new cancer cases and 25% of all cancers in women. Breast cancer is hormone related, and the factors that modify the risk of this cancer when diagnosed premenopausally and when diagnosed (much more commonly) postmenopausally are not the same. Except for skin cancers, breast cancer is the most common cancer among women in the U.S. In 2016, more than 246,660 cases of invasive breast cancer will be diagnosed in women and 2,600 in men in the U.S. Every 2 minutes, one case of breast cancer is diagnosed in a women in the U.S. Early detection and effective treatment contributed to a 37% decline in breast cancer mortality (deaths) between1990-2013. At the current rate, 13 million breast cancer deaths around world will occur in the next 25 years because of lack of medical awareness.

Classification methods are adopted for analyzing the type of cancer. There are many classification methods available in machine learning. In this work, four classification methods namely linear regression, decision tree and random forest are adopted. These method helps to determine the type of cancer at the earlier stage.

II. METHODS

A. *Linear Regression*

Linear regression is used to find the relationship between the attributes (variables). Relationship between independent variable and a dependent variable is determined. The independent variable is "class", other variables are dependent on the class variable and are called as dependent variable. Thus relation between class and other variables are determined using linear regression. This helps to determine which attribute is highly related to class variable and helps for identification of type of cancer.

B. *Decision Tree*

Decision tree is used for predicting the type of tumor present in patient. Relevant variables are chosen and processed by means of decision tree. It results in conditions that are responsible for the possibility of presence of tumor.

C. *Random Forest*

Random Forest, a large number of decision trees are created. Every observation is fed into every decision tree. The most common outcome for each observation is considered as final output. A new observation is fed into all trees and taking majority vote for each classification model. An error estimation is made for the classes which are not used for building the tree. This is called as an out of bag error estimate and is represented in percentage.

978-1-5386-3243-7/17/$31.00 ©2017 IEEE

III. LITERATURE SURVEY

Janos Abonyi and Ferenc Szeifert (2003) proposed a supervised fuzzy clustering algorithm in which each object can be represented by more than one class label. The input variables are selected based on the analysis of clusters by Fisher's interclass separability criteria. To examine the performance evaluation two models are taken that is, wine dataset and Wisconsin breast cancer dataset. In Wisconsin breast cancer dataset the classification distribution is 65.5% benign and 34.5% malignant respectively. The implementation of supervised fuzzy clustering algorithm results is 95.57% average classification accuracy, with 90.00% as worst and 95.57% as best performance. These results indicates that the proposed clustering algorithm is effectively utilizes the class label. F.Paulin and A. Santhakumaran (2011) presents a study on breast cancer by using Feed Forward Artificial Neural Network. The performance of the network are evaluated by various training algorithms such as Batch Training, Batch Gradient Algorithms, Quasi-Newton Algorithms, Levenberg-Marquardt, Resilient Back propagation methods. The highest classification accuracy of 99.28% is achieved when using Levenberg-Marquardt method. Bekaddour Fatima and Chikh Mohammed Amine (2012) provides an approach for recognizing breast cancer diagnosis using Adaptive Neuro Fuzzy Inference System. In this approach CAD (Computer Aided Diagnosis) method is adopted for pattern recognition, aiming to doctors in making diagnosis decision. This results in knowledge extraction and classification of breast cancer disease using neuro-fuzzy approach for explaining human decision. Nahato et. al. (2014) built classifiers which will predict the presence or absence of a disease by learning from the attributes that are extracted from dataset. In this work rough set relation with back propagation neural network is used. It consists of two stages: first stage is handling of missing values and selecting appropriate attribute from the dataset and the second stage is performing classification using back propagation neural network. The accuracy obtained for breast cancer dataset is 98.60%. AhmetMertet.al. (2014) explains the features of independent component analysis on breast cancer decision support system. They perform reduction to one feature from 30 features and are used to evaluate the diagnostic accuracy of classifiers such as k-nearest neighbor (k-NN), artificial neural network (ANN), radial basis function neural network (RBFNN), and support vector machine (SVM).The results are categorized on tumors as benign and malignant in terms of specificity, sensitivity, accuracy, F-score which improves in diagnostic decision support with reduced computational complexity.

IV. EXISTING METHOD

The existing methods helps to identify and analyze the type of cancer present in patient. The results indicates only the presence of tumor and it doesn't provide any information whether the cancer present in the patient is curable (benign) or non-curable (malignant). The proposed methods indicates the results whether the tumor is curable or non-curable.

V. PROPOSED METHOD

i. LINEAR REGRESSION

Regression analysis is used to establish relationship between two variables. One variable is called predictor variable and the other is called response variable.

In Wisconsin breast cancer dataset, the relationship between the attributes are calculated and the attributes which are strongly related with one another are identified and are considered for analysis. The important attribute is class and the relation between class with other attributes are calculated. The dataset contains ten attributes. Each attribute contains their unique feature. The attribute description follows.

- **Clump thickness:** Benign cells tend to be grouped in monolayers, while cancerous cells are often grouped in multilayer.
- **Uniformity of cell size/shape:** Cancer cells tend to vary in size and shape. That is why these parameters are valuable in determining whether the cells are cancerous or not.
- **Marginal adhesion:** Normal cells tend to stick together. Cancer cells tend to loose this ability. So loss of adhesion is a sign of malignancy.
- **Single epithelial cell size:** Is related to the uniformity mentioned above. Epithelial cells that are significantly enlarged may be a malignant cell.
- **Bare nuclei:** This is a term used for nuclei that is not surrounded by cytoplasm (the rest of the cell). Those are typically seen in benign tumors.
- **Bland Chromatin:** Describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells the chromatin tends to be coarser.
- **Normal nucleoli:** Nucleoli are small structures seen in the nucleus. In normal cells the nucleolus is usually very small if visible at all. In cancer cells the nucleoli become more prominent, and sometimes there are more of them.

The attribute values ranges from

Clump Thickness	1 - 10
Uniformity of Cell Size	1 - 10
Uniformity of Cell Shape	1 - 10

Marginal Adhesion	1 - 10
Single Epithelial Cell Size	1 - 10
Bare Nuclei	1 - 10
Bland Chromatin	1 - 10
Normal Nucleoli	1 - 10
Mitoses	1 - 10
Class	(2- benign, 4 - malignant)

Linear regression for Wisconsin Breast Cancer dataset is shown in Figure 1. The attributes are analyzed with each other attribute. The independent attribute class is performed regression with dependent attribute. "Class" attribute is performed with other attributes such as "clump_thickness, shape_uniformity, size_uniformity, bland_chromatin, normal_nucleoli, bare_nucleoli, mitoses, marginal_adhesion, epithelial_size". The accuracy varies for each attributes. The highest accuracy obtained is 84.15%.Thus the accuracy obtained is by excluding the attribute mitoses.

```
Call: lm(formula = class ~ clump_thickness +
shape_uniformity + size_uniformity +
marginal_adhesion + epithelial_size +
bare_nucleoli + bland_chromatin +
normal_nucleoli, data = wbcd)
Residuals:
     Min       1Q   Median       3Q      Max
-1.67976 -0.16600 -0.02453  0.11442  1.52764
Coefficients:  Estimate Std. Error t value
Pr(>|t|)
(Intercept)       1.505412   0.032613  46.160
< 2e-16 ***
clump_thickness   0.063518   0.007108   8.936
< 2e-16 ***
shape_uniformity  0.031286   0.012464   2.510
0.012300 *
size_uniformity   0.043806   0.012723   3.443
0.000611 ***
marginal_adhesion 0.016693   0.007910   2.110
0.035194 *
epithelial_size   0.020559   0.010261   2.004
0.045509 *
bare_nucleoli     0.090711   0.006429  14.109
< 2e-16 ***
bland_chromatin   0.038179   0.010043   3.801
0.000157 ***
normal_nucleoli   0.037237   0.007379   5.046
5.8e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1
Residual standard error: 0.3801 on 674 degrees
of freedom
Multiple R-squared:  0.8433, Adjusted R-
squared:  0.8415
F-statistic: 453.5 on 8 and 674 DF,  p-value:
< 2.2e-16
```

Fig. 1. Linear Regression for Wisconsin Breast Cancer Data Set

ii. .DECISION TREE

Decision tree is a graph used to represent results in the form of tree. Based on the results, the decisions are obtained and the methodology about treatment process for cancer. The results are shown in Figure 2

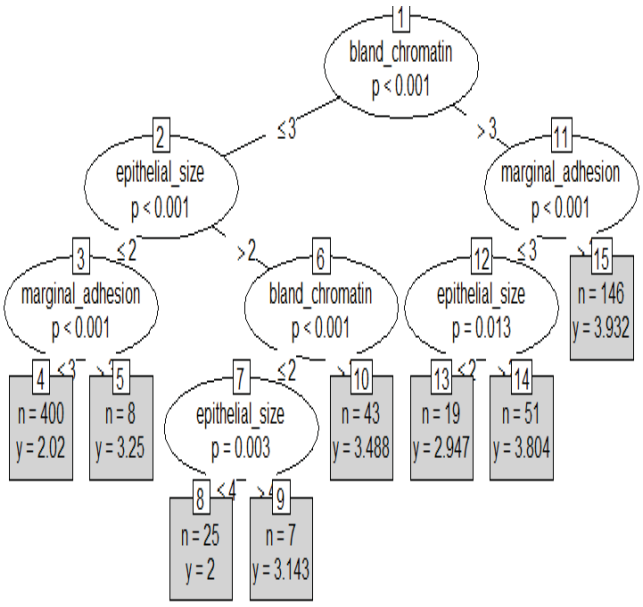

Fig. 2. Decision tree for the Treatment Process

- the order to perform decision tree analysis.

< or = (2,3) - condition to check for the presence of tumor in each attribute.

p - probability of getting malignant tumor.

n (no) - probability of getting benign tumor.

y (yes) - probability of getting malignant tumor.

Initially the attribute bland_chromatin is considered. The probability is considered to be 0.001. The value for bland_chromatin is 3. If the value is less than or equal to 3. Then epithelial_size is considered. The value for epithelial_size is 2. If the value of epithelial_size is less than or equal to 2,marginal_adhesion is checked. It contains yes or no conditions, which indicates yes and no for malignant tumor. If it is less than 3, the probability is 400, which indicates benign tumor and 2.02 indicates malignant tumor. If the value is greater than 5, the probability is 8 for benign tumor and 3.25 for malignant. After analysis of epithelial_size less than 2, greater than two is to be analyzed. Similarly the attribute are analyzed and the probability is determined. The outcome is, higher chance for getting malignant tumor because the value of getting malignant is 3.932 which is higher when compared with other attributes mentioned in decision tree.

iii. RANDOM FOREST

Random forest is ensemble learning method for classification. In the random forest approach, a large number of decision trees are created. Every observation is fed into every decision tree. The most common outcome for each observation is used as the final output. A new observation is fed into all the trees and taking a majority vote for each classification model.

An error estimate is made for the cases which were not used while building the tree. That is called an **OOB (Out-of-bag)** error estimate which is mentioned as a percentage.

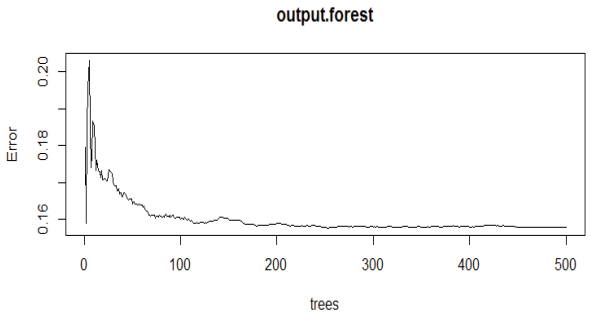

Fig. 3. Random Forest for Error Estimation
X-axis – number of trees.
Y-axis - error rate.

In the considered case study, regression tree is chosen for analyses of random forest. The number of trees considered are 500 trees. The number of variables tries to split is 2. The result is obtained as 88.14% which indicates the percentage of variance of the attributes considered in random forest. The residual sum of squares (RSS), also known as the sum of squared residuals (SSR) or the sum of squared errors of prediction (SSE) is 0.110228 and it is a measure of the discrepancy between the data and an estimation model (regression).

CONCLUSION

The methods linear regression, decision tree and random forest are analyzed for predicting the type of treatment that can be offered for the patient with breast cancer. The success rate of classification is 84.14% obtained by linear regression. The prediction percentage is 88.14% obtained by random forest. As a future enhancement, how the preprocessing of data set can be improvised for increasing the success rate in classification and prediction process in the determination of breast cancer.

ACKNOWLEDGMENT

The authors would like to acknowledge that this work has been carried out at DST-FIST sponsored Cloud Computing Lab (order Saction No. : **SR/FST/ETI-364/2014** Dated: **21 November, 2014**), School of Computing, Sathyabama University.

REFERENCES

[1] Abonyi, J., & Szeifert, F. (2003). Supervised fuzzy clustering for the identification of fuzzy classifiers. Pattern Recognition Letters, 24, 2195-2207

[2] Kemal Polat ∗, Salih Güne¸s , Breast cancer diagnosis using least square support vector machine, Digital Signal Processing 17 (2007),Elsevier.

[3] Mehmet Fatih Akay, Support vector machines combined with feature selection for breast cancer diagnosis, Expert Systems with Applications 36 (2009),Elsevier.

[4] Janos Abonyi *, Ferenc Szeifert, Supervised fuzzy clustering for the identification of fuzzy classifiers, Pattern Recognition Letters 24 (2003), Elsevier.

[5] Nahato, K. B., Nehemiah, H. K., & Kannan, A. (2015). Knowledge Mining from Clinical Datasets Using Rough Sets and Backpropagation Neural Network. Comp. Math. Methods in Medicine, 2015.

[6] AhmetMert., Niyazi KJ.,Erdem Bilgili., & Aydin Akan., (2014) . Breast Cancer Detection with Reduced Feature Set Hindawi Publishing Corporation.Computational and Mathematical Methods in Medicine.

[7] From https://cran.r-project.org/