Diagnosis of Breast Cancer using Decision Tree Models and SVM

[1]Puneet Yadav, [2]Rajat Varshney, [3]Vishan Kumar Gupta

[1, 2]Bachelors in Computer Science and Engineering
[3]Professor, Dept. of Computer Science and Engineering, IMS Engineering college, Ghaziabad, UP, India

---

Abstract - *Breast cancer has become the second important cause of cancer deaths in women today world. It is the most common type of cancer today found in the women today .Decision tree and SVM is a most important technique in the medical field. Disease diagnosis is one of most important application of data mining to proving successful results. Breast Cancer Diagnosis are two medical applications which became a big challenge to the researchers. The use of machine learning and data mining techniques has changed the whole process of breast cancer Diagnosis. Breast Cancer Diagnosis using machine learning technique distinguishes benign from malignant breast lumps. These two problems are mainly in the scope of the classification problems. Most data mining methods which are commonly used in this domain are considered as classification category and applied prediction techniques assign patients to either a" benign" group that is non-cancerous or a" malignant" group that is cancerous. In this study, two powerful classification algorithms decision tree and Artificial Neural Network have been applied for breast cancer prediction. Experimental results show that algorithms has a effective results for this purpose with the overall prediction accuracy of decision tree is from 90% to 94%, and SVM has 94.5% to 97% respectively.*

Key Words: **malignant benign, SVM**

1. INTRODUCTION

An Breast cancer has become a common disease among women around the world and considered as the second largest prevalent type of cancer which causes deaths among women. However, it is also considered as the most curable cancer type as long as it can be diagnosed early. A group of rapidly dividing cells may form a lump or mass of extra tissue which is known as tumors. Tumors can be categorized either as cancerous (malignant) or non-cancerous (benign). Malignant tumors, which considered as a dangerous group, can penetrate and destroy healthy body tissues. The term, breast cancer, refers to a malignant tumor which has developed from the beast's cells. Malignant tumors, which considered as a dangerous group, can penetrate and destroy healthy body tissues. The term, breast cancer, refers to a malignant tumor which has developed from the beast's cells.

World Health Organization (WHO) statistics show that there are more than 1.2 billion women around the world which are diagnosed with breast cancer. In recent years, this graph has been reduced due to the effective Machine learning techniques. Recently, the advancement of data-driven techniques has introduced new and effective ways in the area of breast cancer diagnostics. Powerful expert and data-driven methods: Artificial Neural Network, fuzzy systems, decision tree, Support Vector Machine (SVM), Bayesian Network, etc. It goes without saying that data evaluation which has been attained from the patients can be considered as an important factor to develop an efficient and accurate diagnostic method. classification algorithms have been utilized to minimize the error of human. Which may happen during the treatment. Breast cancer prediction based on machine learning algorithms? The area under the ROC curve was used as a measurement of accuracy. Illustrated in Fig. 1

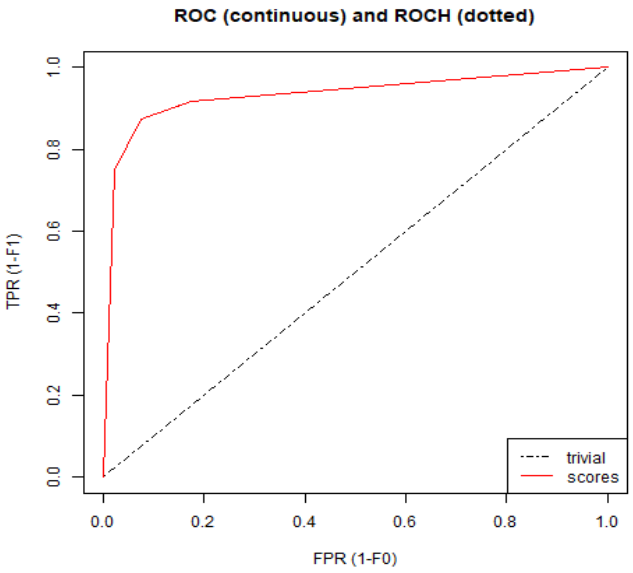

Fig. 1

The goal of this paper is using machine technique to predict benign cancer or the malignant one. Decision trees, Neural Networks, and SVM are powerful data mining techniques tools that can be used to achieve effective results. These algorithms construct their models using training data set then test the obtained models on the test data. Decision tree algorithms are based on constructing a tree that consists of nodes in which each node reflects a test on an attribute until you reach a leaf node. In neural networks, the dataset attributes are divided into three layers: Input, Hidden and output layer. Then, the first two layers are used to indicate the output layer. In this study,

machine learning algorithms will be tested using breast cancer Wisconsin data set, and then compared to result.

1.1 Data Set Description

The database therefore reflects this chronological grouping of the data. Table 1 summarized the attributes which are used for breast cancer diagnostics.

Table 1 Data set description

No.	Attribute	Description
1	Sample code number	Unique key
2	Clump thickness	Cancerous cells are grouped often in multilayers, while benign cells are grouped in monolayers.
3	Uniformity of cell shape	Cancer cells vary in size and shape.
4	Marginal adhesion	Normal cells tend to stick together, while cancer cells fail to do that
5	Single epithelial Cell Size	Epithelial cells that are enlarged may be a malignant cell. In benign tumors, nuclei are often not surrounded by the rest of the cell.
6	Bland chromatin	The texture of nucleus in benign cells
7	Normal nucleoli	Nucleus small structures that are barely visible in normal cells
8	Mitoses	The process of cell division
9	Class	Indication of a tumor category

1.2 DECISION TREES

The Decision trees algorithm consists of two parts: nodes and rules (tests). We construct the tree. In which each node reflect a test on an attribute the basic idea of this algorithm is to draw a flowchart diagram that contains a root node on top. All other (non-leaf) nodes represent a test until you reach a leaf node (final result). Decision tree algorithms have been widely used in data mining applications below are some important reasons that why decision trees are used in the area of data mining and classification:

Decision trees create user-friendly rules. They are considered one of easy to understand algorithms to the end user in Data mining.They show effective association among the dataset attributes and represent in an easy-to-understand form. Decision trees provide a clear indication of important attributes. Decision trees require less computation. They require less computation compared to other classification algorithms. When we implementing decision trees to detect breast cancer then leaf nodes are divided into two categories: Benign or Malignant. Rules will be established among the chosen data set attributes in order to determine if the tumor is benign or malignant.

1.1.2 Choose Inputs Variable

[1] "Clump.Thickness" "Cell.Shape" "Marginal. Adhesion"
[4] "Single.Epithelial.Cell.Size" "Bare.Nuclei" "Bland.C hromatin"
[7] "Normal.Nucleoli" "Mitoses" "Cancer"

1.1.2 Model Building -> decisionTree

1) root 559 250 1 (0.55 0.072 0.07 0.068 0.043 0.034 0.02 7 0.041 0.011 0.082)

2) Cancer>=0.5 374 67 1 (0.82 0.088 0.053 0.024 0 0.005 3 0.0027 0.0027 0.0027 0) 4) Cell.Shape< 1.5 285 18 1 (0. 94 0.039 0.025 0 0 0 0 0 0 0) *

5) Cell.Shape>=1.5 89 49 1 (0.45 0.25 0.15 0.1 0 0.022 0.0 11 0.011 0.011 0)
10) Single.Epithelial.Cell.Size< 2.5 66 28 1 (0.58 0.3 0.076 0.045 0 0 0 0 0 0) *

11) Single.Epithelial.Cell.Size>=2.5 23 15 3 (0.087 0.087 0. 35 0.26 0 0.087 0.043 0.043 0.043 0)

22) Cell.Shape< 3.5 10 4 3 (0.2 0.2 0.6 0 0 0 0 0 0 0) *

23) Cell.Shape>=3.5 13 7 4 (0 0 0.15 0.46 0 0.15 0.077 0.0 77 0.077 0) *

3) Cancer< 0.5 185 139 10 (0.011 0.038 0.1 0.16 0.13 0.09 2 0.076 0.12 0.027 0.25) 6) Cell.Shape< 6.5 97 71 4 (0.021 0.072 0.2 0.27 0.15 0.13 0.031 0.041 0.01 0.072) *
7) Cell.Shape>=6.5 88 49 10 (0 0 0 0.034 0.1 0.045 0.12 0. 2 0.045 0.44)

14) Cell.Shape< 9.5 46 29 8 (0 0 0 0.043 0.15 0.065 0.2 0.3 7 0.065 0.11) *

15) Cell.Shape>=9.5 42 8 10 (0 0 0 0.024 0.048 0.024 0.0 48 0.024 0.024 0.81) *

1.1.3 EvaluationsParameters

H Gini AUC AUCH KS MER MWL Spec.Sens95 Sens.Spe c95 ER Sens Spec
scores 0.759 0.83 0.915 0.93 0.844 0.077 0.074 0.66 0.898 0.077 0.918 0.926

1.1.4 Accuracy

90.29

1.1.5 Decision Tree Result

Actual	Predicted
1	1
0	0

1	1
1	1
0	0
0	0
1	1
1	1
1	1
1	0
1	1

Decision trees algorithm on a single attribute. Our data set contains multiple attributes that need to be included. Therefore, a complicated chart that describes multiple relationships (rules) among these attributes will be delivered using Weka application. A major step in classification is to have a test set and training set. Otherwise, the evaluation results will not be reliable. In this study we commonly used ratio to split a dataset into 80% training set and 20% test set. Next step is to decide which decision tree algorithm should be used for a given problem. The tree is applied to each row in the database after it is constructed. Performing initial testing on all decision tree algorithms using our dataset. In addition, simplicity is one of its unique features, the output of algorithm can be easily understood by the end user and it satisfies the performance measure. Illustrated in Fig. 2

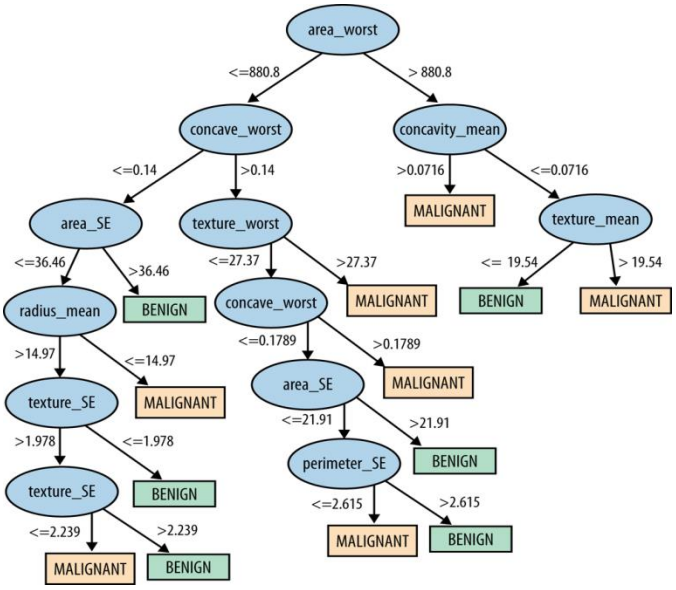

Fig. 2

2. ARTIFICIAL NEURAL NETWORKS

Neural networks (NNs) have been widely used in different fields as a discerning tool in recent years. Using neural network in classification of breast cancer dataset has become a popular intelligent technique. NNs are transmission function of mapping from input to output. If each different input is regarded as a form of input mode, the mapping to the output is considered as output response model, the mapping from input to output is undoubtedly the issue of pattern classification. Any neural network model must be trained before it can be considered intelligent and ready to use. Neural networks are trained using training sets, and then they can predict the result in the test set. Below are two major factors which make Artificial Neural Network (ANN) as a robust classification algorithm:

Neural networks are adaptive in nature. A neural network is composed of "living" units or neurons. This algorithm learns information from data. Learning is the most interesting feature of neural networks.

Neural networks are massively parallel in nature.The use of neural network to classify breast cancer data is illustrated in Fig. 3. In this study, the input attribute are: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses. An Intermediate cell is called the hidden layer units. The output of the hidden layer is considered as the input of two output units, corresponding to a result of the diagnosis of breast cancer, benign or malignant tumor

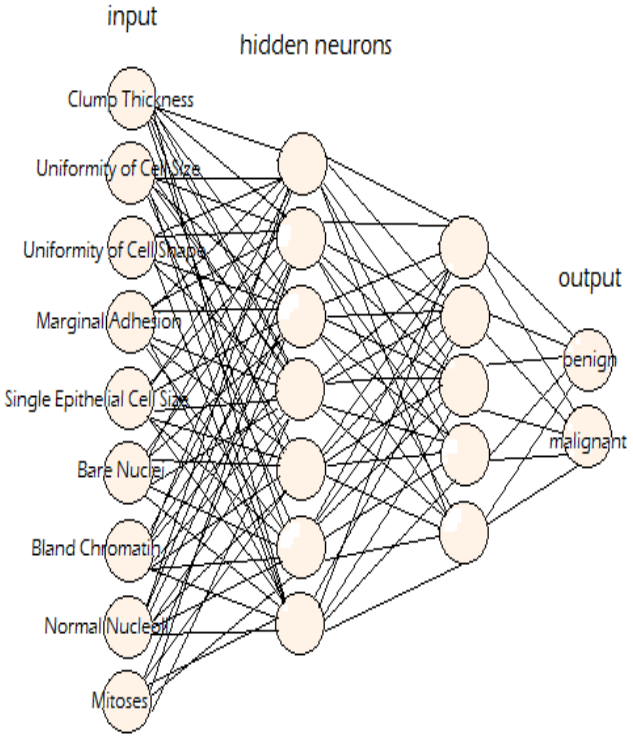

Figure 3. Artificial Neural Network for breast cancer prediction

3. SVM (Support Vector Machine)

Support Vector Machine (SVM) is a supervised machine learning algorithm which is used for both classification and regression problems. In this algorithm, we plot each

data item as a point in n-dimensional space where n is number of features you have with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well Illustrated in Fig. 4.

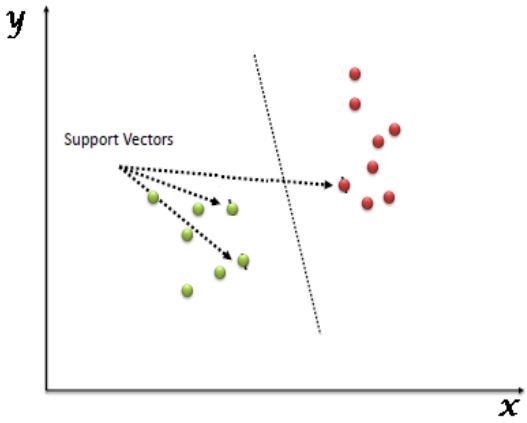

Fig. 4

Support Vectors are simply the co-ordinates of individual examination. Support Vector Machine is a borderline which best segregates the two classes.Working of SVM as we got routine to the process of segregating the two classes (malignant and benign) with a hyper-plane. Identify the right hyper-plane here we have three hyper-planes A, B and C. Now, identify the right hyper-plane to classify star (malignant) and circle (benign). You need to remember a rule to identify the right hyper-plane Select the hyper-plane which segregates the two classes better. In this scenario, hyper-plane "B" has excellently performed this job.Identify the right hyper-plane here, we have three hyper-planes A, B and C and all are segregating the classes well. Here, maximizing the distances between nearest data points (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as Margin.Let's look in fig.5

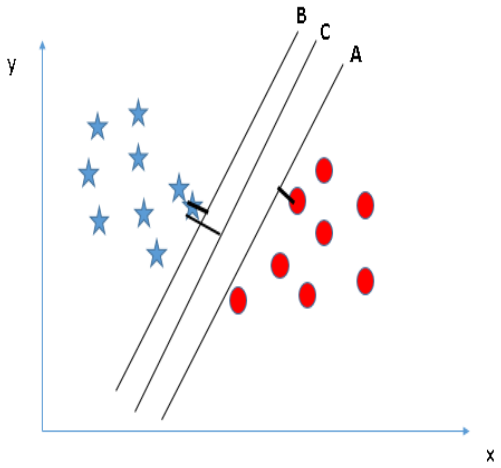

Fig.5

4. CONCLUSIONS

Decision tree and Neural Networks are powerful data mining techniques that can be used to classify cancerous tumors. Decision tree algorithm creates user-friendly rules that indicates important attributes and requires less computation compared to other algorithms such as Neural Networks. On the other hand.Various data mining techniques are available in medical diagnosis, where the objective of these techniques is to assign patients to either a 'healthy' group that does not have a certain disease.Data mining have proved the ability to reduce the number of error rate in decisions. Decision tree and SVM are the most popular and effective data mining methods. DT provides a pathway to find "rules" that could be evaluated for separating the input samples into one of several groups without having to express the functional relationship directly.

REFERENCES

1. International Journal of Computer Applications Technology and Research Volume 7–Issue 01, 23-27, 2018, ISSN:-2319–8656.

2. http://www.imaginis.com/general-information-on-breast-cancer/what-is-breast-cancer-

3. Übeyli, E. D. (2007). Implementing automated diagnostic systems for breast cancer detection. Expert Systems with Applications, 33(4), 1054-1062.

4. https://www.safaribooksonline.com/library/view/data-science-for/9781449374273/ch04.html

5. https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-

6. f0812effc72Shrivastava, Shiv, Anjali Sant, and Ramesh Aharwa. "An Overview on Data Mining Approach on Breast Cancer Data." International Journal of Advanced Computer Research (2013): n. pag. Web.

7. Tike Thein1, Htet Thazin, and Khin Mo Mo Tun. "An Approach for Breast Cancer Diagnosis Classification Using Neural Network." Advanced Computing: An International Journal (ACIJ) 6 (2015)

8. Wolberg, WIlliam. "UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set." UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set. University of Wisconsin Hospitals Madison, Wisconsin, USA, n.d. Web. Oct. 2015.'