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

Diagnosis of Breast Cancer using Decision Tree and Artificial Neural Network Algorithms

Autsuo Higa
Toyohashi University of Technology
Toyohashi, Japan

Abstract: Breast Cancer Diagnosis and Prognosis are two medical applications which have posed a challenge to the researchers. The use of machine learning and data mining techniques has revolutionized the whole process of breast cancer Diagnosis and Prognosis. Breast Cancer Diagnosis distinguishes benign from malignant breast lumps and Breast Cancer Prognosis predicts when Breast Cancer is likely to recur in patients that have had their cancers existed. Thus, 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. Hence, the breast cancer diagnostic problems are basically in the scope of the widely discussed classification problems. In this study, two powerful classification algorithms namely decision tree and Artificial Neural Network have been applied for breast cancer prediction. Experimental results show that the aforementioned algorithms has a promising results for this purpose with the overall prediction accuracy of 94% and 95.4%, respectively.

Keywords: Breast Cancer Diagnostics, Machine learning Techniques, Artificial Neural Networks, Decision Tree, Data Mining

1. INTRODUCTION

Breast cancer has become a common disease among women around the world and considered as the second largest prevalent type of cancer which cause deaths among women [1]. 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 are known as tumors [2]. 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. Based on the World Health Organization statistics, there are more than 1.2 billion women around the world which are diagnosed with breast cancer. However, in recent years, this trend has been reduced due to the effective diagnostic techniques which can cure the cancer if it is diagnosed in an appropriate time.

Recently, the advancement of data-driven techniques have introduced new and effective ways in the area of breast cancer diagnostics. Data mining and expert systems have not only actively utilized in the medical problems, but also they have widely used in other industrial applications [3][4][5]. To name some of the powerful expert and data-driven methods: Artificial Neural Network, fuzzy systems, decision tree, Support Vector Machine (SVM), Bayesian Network, etc. [6][7][8]. It goes without saying that data evaluation which have been attained from patients can be considered as an important factor to develop an efficient and accurate diagnostic method. To this end, classification algorithms have been utilized to minimize the error of human errors which may happen during the treatment.

Breast cancer prediction based on machine learning algorithms has attracted the attention of many researchers recently. For example, Lunin et al. [9] evaluated the accuracy of Neural Network in 5, 10, and 15-year breast cancer specific survival. They use a data set with 951 patients. The area under the ROC curve was used as a measurement of accuracy and the AUC values for neural networks are 0.909, 0.886, and 0.833 for 5, 10, and 15-year breast cancer specific survival, respectively. They also use logistic regression in their paper and the AUC values for logistic regression are: 0.897, 0.862, and 0.858, respectively. In [10], authors present an analysis in rate of survivability with three data mining techniques: Naïve Bayes, the back-propagated neural network, and C4.5 decision tree algorithm. SEER dataset has been used for their research. The accuracy of prediction for these techniques Naïve Bayes, the back-propagated neural network, and C4.5 decision tree are 84.5%, 86.5%, and 86.7% respectively. They also show that the C 4.5 has the best performance in this case.

One of the approaches toward the breast cancer prediction is diagnosis via mammography images which is considered as image processing and classification. In [11], authors proposed a method for automatic segmentation of the mammogram images and then classified them as a malignant, benign or normal based on the decision tree J48 algorithm. The accuracy of their method for breast cancer diagnosis via mammography images for positive prediction and negative prediction are 94% and 98.5%, respectively.

Authors in [12] propose a method which use Support Vector Machines (SVMs) and decision tree for classifying 100 breast cancer patients into two classes: Benign and Malignant. They concluded that on the basis of the accuracy the SVM (with the accuracy of 98%) is better than the decision tree (96% of accuracy). In the second stage of their method k-mean clustering technique has been used to partition the above two classes of patients into three categories: Poor, Intermediate, and Good to determine whether the patient is in urgent need of chemotherapy with respect to the survival time of the patient.

The goal of this paper is using machine to predict whether a has a benign cancer or malignant one. Decision trees and Neural Networks are powerful data mining techniques tools that can be used to achieve that. Both 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 reflect 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, the two algorithms will be tested using breast cancer Wisconsin data set [13], and then compared to each other based on their ability to predict cancerous tumors.

1.1 Data Set Description

Samples arrive periodically as Dr. Wolberg reports his clinical cases. 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	Value
1	Sample code number	Unique key	ID Number
2	Clump thickness	Cancerous cells are grouped often in multilayers, while benign cells are grouped in monolayers.	(1-10)
3	Uniformity of cell Size	Cancer cells vary in size and shape.	(1-10)
4	Uniformity of cell shape		(1-10)
5	Marginal adhesion	Normal cells tend to stick together, while cancer cells fail to do that	(1-10)
6	Single epithelial Cell Size	Epithelial cells that are enlarged may be a malignant cell.	(1-10)
7	Bare nuclei	In benign tumors, nuclei is often not surrounded by the rest of the cell.	(1-10)
8	Bland chromatin	The texture of nucleus in benign cells	(1-10)
9	Normal nucleoli	Nucleus small structures that are barely visible in normal cells	(1-10)
10	Mitoses	The process of cell division	(1-10)
11	Class	Indication of a tumor category	2 - Benign 4 - Malignant

The initial data preprocessing resulted in using only 10 attributes from the chosen dataset, which are (clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses, class). Taking away the (Sample code number) attribute since it is not going to be useful for our purpose. The chosen algorithms will be implemented using Weka3 [14] which is a Data mining software written in Java.

2. DECISION TREES

Decision trees algorithm consists of two parts: nodes and rules (tests). 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 to a single or multiple attributes until you reach a leaf node (final result). Decision tree algorithms have been widely used in data mining applications due to the fact that they are powerful classification tools [15]. Below are some important reasons that why decision trees are used in the area of data mining and classification:

- *Decision trees create understandable rules*: They are considered one of the friendliest algorithms to the end user in data mining. They initiate relationships among the dataset attributes in an easy-to-understand form.
- *Decision trees provide a clear indication to important attributes:* a major part of establishing rules between attributes is indicating the importance level of each one.
- *Decision trees require less computation:* They require less computation compared to other classification algorithms such as mathematical formulae.

When implementing decision trees algorithm to detect breast cancer, 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. Figure 1. shows an example of using decision tree approach for breast cancer detection.

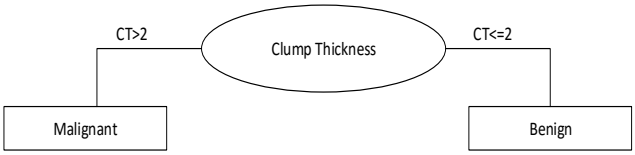

Figure 1. Decision tree example for Breast Cancer Detection

This figure illustrates a 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. Decision trees algorithm will be judged and evaluated based on its ability to predict cancerous cells. A major step in classification is to have a test set that is different from the used training set. Otherwise, the evaluation results will not be reliable. In this study, Pareto principle is employed [16] as 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. Weka offers multiple decision tree algorithms, such as J48, Random forest and Decision stump. J48 is the implementation of decision tree algorithm ID3 that creates a binary tree [17]. The tree is applied to each row in the database after it is constructed. After performing initial testing on all decision tree algorithms using our dataset, we found out that J48 algorithm is relatively faster than other decision tree algorithms. In addition, simplicity is one of its unique features, the output of this algorithm can be easily understood by the end user and it satisfies the performance measure. Therefore, J48 decision tree algorithm has been used

in this study and below is the classifier output after running in Weka.

Classifier output:

```
=== Run information ===
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     breast-cancer-
weka.filters.unsupervised.attribute.Remove-R1
Instances:    699
Attributes:   10
              Clump_Thickness
              Uniformity_of_Cell_Size
              Uniformity_of_Cell_Shape
              Marginal_Adhesion
              Single_Epithelial_Cell_Size
              Bare_Nuclei
              Bland_Chromatin
              Normal_Nucleoli
              Mitoses
              Class
Test mode:    split 80.0% train, remainder test
=== Classifier model (full training set) ===
J48 pruned tree
------------------
Uniformity_of_Cell_Size <= 2
|  Bare_Nuclei <= 3: 2 (405.39/2.0)
|  Bare_Nuclei > 3
|  |  Clump_Thickness <= 3: 2 (11.55)
|  |  Clump_Thickness > 3
|  |  |  Bland_Chromatin <= 2
|  |  |  |  Marginal_Adhesion <= 3: 4 (2.0)
|  |  |  |  Marginal_Adhesion > 3: 2 (2.0)
|  |  |  Bland_Chromatin > 2: 4 (8.06/0.06)
Uniformity_of_Cell_Size > 2
|  Uniformity_of_Cell_Shape <= 2
|  |  Clump_Thickness <= 5: 2 (19.0/1.0)
|  |  Clump_Thickness > 5: 4 (4.0)
|  Uniformity_of_Cell_Shape > 2
|  |  Uniformity_of_Cell_Size <= 4
|  |  |  Bare_Nuclei <= 2
|  |  |  |  Marginal_Adhesion <= 3: 2 (11.41/1.21)
|  |  |  |  Marginal_Adhesion > 3: 4 (3.0)
|  |  |  Bare_Nuclei > 2
|  |  |  |  Clump_Thickness <= 6
|  |  |  |  |  Uniformity_of_Cell_Size <= 3: 4 (13.0/2.0)
|  |  |  |  |  Uniformity_of_Cell_Size > 3
|  |  |  |  |  |  Marginal_Adhesion <= 5: 2 (5.79/1.0)
|  |  |  |  |  |  Marginal_Adhesion > 5: 4 (5.0)
|  |  |  |  Clump_Thickness > 6: 4 (31.79/1.0)
|  |  Uniformity_of_Cell_Size > 4: 4 (177.0/5.0)

Number of Leaves  :          14
Size of the tree :     27
Time taken to build model: 0.01 seconds
=== Evaluation on test split ===
Time taken to test model on training split: 0 seconds
=== Summary ===
Correctly Classified Instances         130               92.8571 %
Incorrectly Classified Instances        10                7.1429 %
Kappa statistic                          0.8485
Mean absolute error                      0.092
Root mean squared error                  0.2429
Relative absolute error                 20.2164 %
Root relative squared error             50.6609 %
Coverage of cases (0.95 level)          98.5714 %
Mean rel. region size (0.95 level)      70      %
Total Number of Instances               140
=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure
MCC      ROC Area  PRC Area  Class
               0.911     0.040     0.976     0.911     0.943      0.852
0.955     0.962     2
               0.960     0.089     0.857     0.960     0.906      0.852
0.955     0.893     4
Weighted Avg.  0.929     0.057     0.934     0.929     0.929
0.852     0.955     0.937
=== Confusion Matrix ===
  a  b   <-- classified as
 82  8 |  a = 2
  2 48 |  b = 4
```

Looking at the confusion matrix, we can see that the algorithm successfully predicted 82 benign and 48 malignant cases with a predictive accuracy rate equal to 92.8571 %. To optimize the results, we ran 10 tests using different training and test sets every time, the algorithm successfully predicted 94 % cases on average. More details are available in Figure 3.

3. ARTIFICIAL NEURAL NETWORKS

Neural networks (NNs) have been widely used in different fields as an intelligent tool in recent years. Recently, using neural network in classification of breast cancer dataset has become a popular intelligent tool [18]. Generally speaking, NNs is 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 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 solution in the test set. Below are two major factors which make Artificial Neural Network (ANN) as a powerful classification algorithm:

- *Neural networks are adaptive:* A neural network is composed of ‘‘living’’ units or neurons. It can learn or memorize information from data. Learning is the most fascinating feature of neural networks.
- *Neural networks are naturally massively parallel:* This is the structural similarity of ANNs to biological ones. Though in some cases neural network models are implemented in software on ordinary digital computers, they are naturally suitable for parallel implementations.

The use of neural network to classify breast cancer data is illustrated in Fig. 2. In this study, the input nodes 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. Intermediate cell is called the hidden layer units, whose output

are only in the internal network, not a part of all the network output. 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.

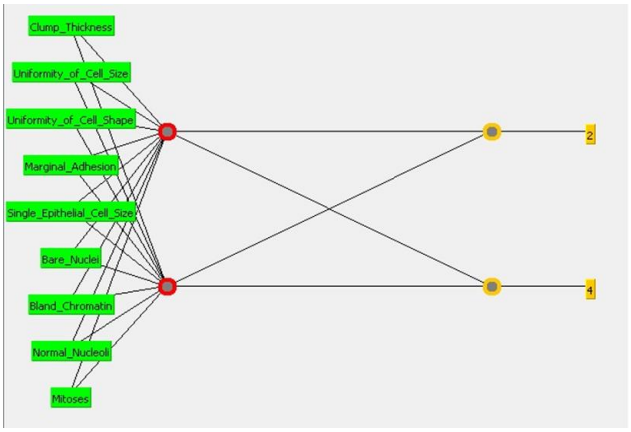

Figure 2. Artificial Neural Network for breast cancer prediction

Instead of using Conventional validation to divide the dataset into training set and test set, we use 10-fold cross validation. One of the main reason for using cross validation is to assess how the result of network will generalize to independent data and how accurately a predictive model will perform in practice. Therefore, cross validation is a fair way to generalize the performance of the neural network. In 10-fold cross-validation, the original sample is randomly partitioned into 10 equal sized subsamples. Of the 10 subsamples, a single sub-sample is retained as the validation data for testing the model, and the remaining 10-1 subsamples are used as training data. The cross-validation process is then repeated 10 times (the *folds*), with each of the 10 subsamples used exactly once as the validation data. The 10 results from the folds can then be averaged (or otherwise combined) to produce a single estimation.

ANNs is applied with different parameters like different no. of hidden layers, learning rate and momentum and the best result is 96.42% of correctly classified instances with the following neural network configuration: (No. of input layer, hidden layer and output layer are: 9,2,2 respectively, learning rate:0.2, and momentum: 0.7). Table 2 shows the confusion matrix, and accuracy of the algorithm are provided in Table 3.

Table 2. Confusion matrix for ANN

	Benign	**Malignant**
Benign	441 ***(a)***	17 ***(b)***
Malignant	8 ***(c)***	233 ***(d)***

The entries in the confusion matrix have the following meaning in the context of this study: ***a***: number of correct predictions that an instance is negative, ***b***: number of incorrect predictions that an instance is positive, ***c***: number of incorrect predictions that an instance is negative, and ***d***: number of correct predictions that an instance is positive. The Breast cancer data with 699 tuples and 9 different attributes was analyzed to identify the error rates and accuracy. Table 3 shows the accuracy measures of the result.

Table 3. ANN performance measurement

	Instances	Percentage
Correctly Classified Instances	674	96.42%
Wrongly Classified Instances	25	3.57%

10 different tests have been conducted on the same dataset using the decision tree algorithm J48 and Multi-layer perception model for neural network. In J48, the dataset was split using Pareto principle ratio, 80% training set and 20% test data. As for Multi-layer perception, the data was split into 10 folds using cross validation. Both algorithms predicted at least 92% cases each test. However, Multi-layer perception model was able to correctly classify more cases on average as shown in Figure3.

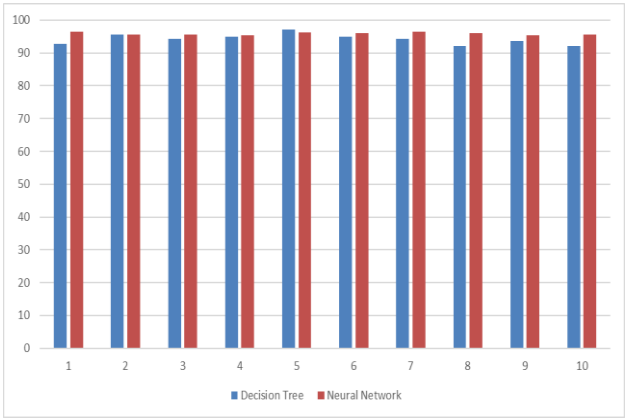

Figure 3. Performance Comparison chart of decision tree and ANN algorithms.

Calculating the mean of these tests, J48 algorithm was able to correctly classify 94.0% cases whereas Multi-layer perception algorithm was able 95.9% cases.

4. CONCLUSION

Decision tree and Neural Networks are powerful data mining techniques that can be used to classify cancerous tumors. Decision tree algorithm creates understandable rules, indicates important attributes and requires less computation compared to other algorithms such as Neural Networks. On the other hand, Neural Network algorithm is an adaptive and naturally suitable for parallel implementations. In this study, both algorithms have been used as intelligent methods for breast cancer diagnostic. Both algorithms were successful in correctly classifying more than 92% cases in the 10 experiments. However, Neural Network algorithm had a better predictive accuracy rate on average (rate of correct classification is 95.9%).

5. REFERENCES

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

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

[3] Khalilian, A., Sahamijoo, G., Avatefipour, O., Piltan, F., & Nasrabad, M. R. S. (2014). Design high efficiencyminimum rule base PID like fuzzy computed

torque controller. International Journal of Information Technology and Computer Science (IJITCS), 6(7), 77.

[4] Khalilian, A., Piltan, F., Avatefipour, O., Nasrabad, M. R. S., & Sahamijoo, G. (2014). Design New Online Tuning Intelligent Chattering Free Fuzzy Compensator. International Journal of Intelligent Systems and Applications, 6(9), 75.

[5] Sahamijoo, G., Avatefipour, O., Nasrabad, M. R. S., Taghavi, M., & Piltan, F. (2015). Research on minimum intelligent unit for flexible robot. International Journal of Advanced Science and Technology, 80, 79-104.

[6] Mokhtar, M., Piltan, F., Mirshekari, M., Khalilian, A., & Avatefipour, O. (2014). Design minimum rule-base fuzzy inference nonlinear controller for second order nonlinear system. International Journal of Intelligent Systems and Applications, 6(7), 79.

[7] Avatefipour, O., Piltan, F., Nasrabad, M. R. S., Sahamijoo, G., & Khalilian, A. (2014). Design New Robust Self Tuning Fuzzy Backstopping Methodology. International Journal of Information Engineering and Electronic Business, 6(1), 49.

[8] Shahcheraghi, A., Piltan, F., Mokhtar, M., Avatefipour, O., & Khalilian, A. (2014). Design a Novel SISO Off-line Tuning of Modified PID Fuzzy Sliding Mode Controller. International Journal of Information Technology and Computer Science (IJITCS), 6(2), 72.

[9] M, Lundin, Lundin J, Burke HB, Toikannen S, and Joensuu H. "Artificial Neural Networks Applied to Survival Prediction in Breast Cancer." US National Library of Medicine National Institutes of Health 57 (1999): 281-6.

[10] Bellaachia, Abdelghani, and Erhan Guven. "Predicting Breast Cancer Survivability Using Data Mining Techniques." Society for Industrial and Applied Mathematics: 1-4.

[11] B, Nadira, and Banu Kamal. "Automatic Classification of Mammogram MRI using dendograms." Asian Journal of Computer Science and Information Technology 4 (2012): 78-81.

[12] Yadav, Reeti, Zubair Khun, and Hina Saxena. "Chemotherapy Prediction of Cancer Patient by Using Data Mining Techniques." International Journal of Computer Applications (0975 – 8887) 76.10 (2013).

[13] 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.'

[14] "Weka 3: Data Mining Software in Java." Weka 3. N.p., n.d. Web. 25 Nov. 2015.

[15] Shrivastava, 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.

[16] "Pareto Density Estimation: A Density Estimation for Knowledge Discover."Y. N.p., n.d. Web. Nov. 2015.

[17] International Journal of Computer Science and Applications, Vol. 6 No.2 Apr 2013, Issn: 0974-1011 (Open Access), Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification (n.d.): n. pag. Web.

[18] 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).