An Algorithm to Reconstruct the Missing Values for Diagnosing the Breast Cancer

F. Paulin, A.Santhakumaran

GJCST Computing Classification
I.2.6, F.1.1 & H.2.m

***Abstract*- The treatment of incomplete data is an important step in pre-processing data prior to later analysis. The main objective of this paper is to show how various methods can be used in such a way that they are able to process dataset with missing values. Computer–aided classification of Breast cancer using Back propagation neural network is discussed in this paper. The classification results have indicated that the network gave the good diagnostic performance of 99.06%.**

Keywords- Artificial Neural Networks, Back propagation, Breast cancer, Successive Iteration.

I INTRODUCTION

Cancer begins in cells, the building blocks that make up tissues. Tissues make up the organs of the body. Normally, cells grow and divide to form new cells as the body needs them. When cells grow old, they die and new cells take their place. Sometimes, this orderly process goes wrong. New cells form when the body does not need them, and old cells do not die when they should. These extra cells can form a mass of tissue called growth or tumor. Tumors can be either cancerous (malignant) or non-cancerous (benign). Malignant tumors penetrate and destroy healthy body tissues. Cancer that forms in tissues of the breast, usually the ducts (tubes that carry milk to the nipple) and lobules (glands that make milk) are called Breast cancer. Breast cancer is one of the leading cancers for women worldwide. It is the second most common cause of cancer death in white, black, Asian/Pacific Islander and American Indian/Alaskan native women [1], [2]. Early detection and improved therapy planning are crucial for increasing the survival rates of cancer patients. To aid clinicians in the diagnosis of breast cancer, recent research has looked into the development of computer aided diagnostic tools.

Neural networks have been widely used for breast cancer diagnosis [3], [5]. The effectiveness of breast cancer classification by training neural networks using a linear programming technique is demonstrated in [7].

Manuscript received "26/02/2010"

F. Paulin is with the MCA department as Senior Lecturer, CMS College of Science and commerce, Coimbatore, Tamilnadu, India. She is doing her ph.d in computer science from Mother Teresa university, Kodaikannel, India. Her research interest is artificial neural network.
(Telephone: 98422 67441 email: paulinrex@rediffmail.com)
Santhakumaran is currently a Reader in the department of statistics, Salem Sowdeswari College, affiliated to Periyar University, Salem, and TamilNadu. He received his PhD in Mathematics-Statistics from the Ramanujam Institute for advanced study in Mathematics, University of Madras. His research interests statistical Quality Control and Stochastic Processes and their applications.
(Telephone: 9443995082 m)

Artificial Neural Network (ANN) has made a significant mark in the domain of health-care applications. The brain learns from experience, in ANN, learning is typically achieved through progressive adjustment of the weighted interconnections of neurons and other network parameters, guided by learning algorithm. Feed forward neural networks have been trained with standard Back propagation algorithm [8]. They are supervised networks so they require a desired response to be trained. They have been shown to approximate the performance of optimal statistical classifiers in difficult problems.

There is much research on medical diagnosis of breast cancer with Wisconsin Breast Cancer Data (WBCD) in neural network literature [9]-[13]. In real world applications, missing values often abound. Therefore, there is a need for algorithms that can cope with missing values. Missing values in a datum mean that the values for some of the attributes of that datum are unknown. In [4], the 16 missing value instances have been left out while using WBCD for Breast Cancer diagnosis. The constructed feed forward neural network has been evaluated for breast cancer detection without replacing missing values [16]. Eliminating some instances will affect the diagnosis accuracy. The seventh attribute called Bare Nuclei of WBCD has 16 missing values. This paper presents a result of direct classification of data after replacing missing values using various methods for the WBCD dataset with a given number of classes.

A. Data Set

This breast cancer database is downloaded from the UCI machine-learning repository [14], which was collected by Dr. William H. Wolberg from the University of Wisconsin Hospitals, Madison [6]. The dataset is comprised of elements that consist of various scalar observations. The total number of the original samples is 699 with 16 samples contain missing values. The dataset contains two classes referring to benign and malignant samples. There are 458 samples in the dataset that are assigned to benign and the other 241 samples are malignant. The original dataset contains 11 attributes including both sample id number and class label, which are removed in the actual dataset that are used in our experiments. The remaining 9 attributes represent 9 cytological characteristics of breast fine-needle aspirates (FNAs), as shown in Table 1. The cytological characteristics of breast FNAs were valued on a scale of one to ten, with one being the closest to benign and ten the most malignant.

Table 1
Attribute Information

No.	Attribute	Domain
1	Clump thickness	1-10
2	Uniformity of cell size	1-10
3	Uniformity of cell shape	1-10
4	Marginal Adhesion	1-10
5	Single Epithelial cell size	1-10
6	Bare Nuclei	1-10
7	Bland Chromatin	1-10
8	Normal Nucleoli	1-10
9	Mitoses	1-10
10	Class	2 for benign and 4 for malignant

Number of instances : 699
Missing Attributes : 16
Benign : 458

II METHODOLOGY

In ANN, the assigned weights for each connector of node resemble the long term memory. They contain information of the input's importance and ANN learns by repeated adjustments of these weights. The weight adjustments are carried out according to the mathematical functions known as learning or activation function, which will be compared to the threshold value of the network. A feed forward back propagation artificial can learn a function of mapping inputs to outputs by being trained with cases of input-output pairs. Back propagation neural network (BPNN) is actually a descending slope method to minimize the total square of the output, calculated by the network [15]. There are three phases in the training process: first is to send the signal pattern forward, second is to calculate the propagated error and the last is to update all weights in the network. In addition BPNN also has the advantages of faster learning in multilayer Neural Network, especially sigmoidal activation function is represented by hyperbolic tangent. The neurons in feed forward networks can be any transfer function of the designer wishes to use. The usually used transfer function is the sigmoid function with threshold defined as in equation (1).

$$f(\sum_{i=1}^{n} w_i x_i - \theta) = 1/(1+\exp(-(\sum_{i=1}^{n} w_i x_i - \theta))) \qquad (1)$$

where x_i is the input to the node and w_i is the corresponding input weight, θ is a value which is usually called the threshold, n is the number of inputs to the node.
The network performance and convergence depends on many parameters like initial weights, learning rate and momentum used during the training process. Fig1. illustrates the flowchart of the overall processes used in this research.

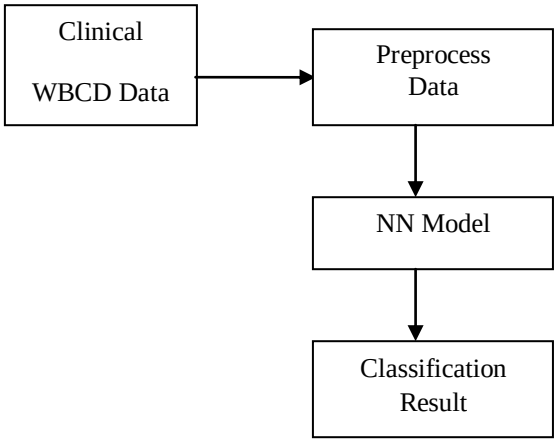

Fig.1. Overall Processes Flow Chart.

The proposed algorithm used in this research is follows.

i. Load Data set.
ii. Replace missing values by any one of the missing values replacement method (mean or median or mode or successive iteration).
iii. Normalize each variable of the data set, so that the values range from 0 to 1. We call this data set as normalized data set.
iv. Create a separate training set and testing set by randomly drawing out the data for training and for testing.
v. Create an initial ANN architecture consisting of three layers, an input, an output and a hidden layer. The number of nodes in the input layer is the same as the number of inputs of the problem. Randomly initialize the nodes of the hidden layer. The output layer contains 1 node. Randomly initialize all connection weights within a certain range.
vi. Train the network on the training set by using Back propagation algorithm until the error is almost constant for a certain number of training epochs, this is specified by the user.
vii. Present the test data to the trained network and evaluate the performance.

Replacement of missing values using Mean Method:

i. Find mean (average) for the Bare Nuclei (This attribute contains missing values).
ii. All the missing value of this attribute replaced by this mean value.

Replacement of missing values using Median Method:

i. Find median (middle value) for the Bare Nuclei (This attribute contains missing values).
ii. All the missing value of this attribute replaced by this median value.

Replacement of missing values using Mode Method:

i. Find the highest value for the Bare Nuclei (This attribute contains missing values).
ii. All the missing value of this attribute replaced by this highest value.

Replacement of missing values using Successive Iteration Method:

i. Find mean for the Bare Nuclei (This attribute contains missing values).
ii. Replace this mean for the first missing value.
iii. Again, find the mean for the entire attribute.
iv. If the new mean and old mean are same then replace this mean value for all missing values and stop the iteration
v. Else, perform step 1.

III AN APPLICATION

Preprocessing the input data set for a knowledge discovery goal using the neural network approach usually consumes the biggest portion of the effort devoted in the entire work. A simple analysis shows that the WBCD data set has missing information in the field of Bare Nuclei for 16 records. In this research, these missing values have been replaced by the calculated value using various replacement methods.
The following pre-classification rule have adopted in this work. In which three fields are included: Clump thickness, Bare Nuclei, and Mitoses as given below.
If (Clump thickness < 7 and Uniformity of cell size<8 and Uniformity of cell shape<3 and Normal Nucleoli<9) then

Benign

Else

Malignant

A. Normalize the Data

One of the most common tools used by designers of automated recognition systems to obtain better results is to utilize data normalization. Data normalization can also speed up training time by starting the training process for each feature within the same scale.
Input data has been normalized by the formulae as in (2), in the range between 0 and 1:

$$\bar{X} = (x - x_{\min}) / (x_{\max} - x_{\min}) \qquad (2)$$

Where, $\bar{X}$ is standard value of input,
x is Observed value,
$x_{\max}, x_{\min}$ are minimum and maximum actual observed values.

The above mentioned approaches for the proposed algorithm have been implemented and tested with the breast cancer dataset from the University of Wisconsin Hospitals, Madison, collected by Dr. W. H. Wolberg. To analyze the data neural network toolbox, which is available in MATLAB, software is used.
The feed forward neural network (FFNN) consists of an input layer, an output layer and one hidden layer. With 9 features in each input vector and 2 values in each output vector, we select 5 nodes for the hidden layer. The training algorithm is a standard back-propagation with a set of initialized parameters. The non-linearity in the hidden layer is the sigmoid function. The Levenberg-Marquardt (trainlm) algorithm was applied to increase the training speed. This trainlm algorithm appears to be the fastest method for training moderate-sized feed forward neural networks (up to several hundred weights). In this application, 80% of the data were selected randomly and used to train and construct the network. The remaining 20% of the data were then used to test capability of the resulting network. The proposed algorithm was executed 10 times each with a different set of missing value replacement methods. In Table 3 the percentage of correct classification indicates the percentage of the patterns that were correctly classified by the constructed networks. Table 2 shows the accuracies of different replacement methods.

Table 2
Performance of the Replacement methods

S. No.	Missing Value Replacement Methods	Percentage of Correct Classification
1	Mean	98.92%
2	Median	99.06%
3	Mode	98.56%
4	Successive Iteration	98.63%

IV CONCLUSION

In this research, a feed forward neural network is constructed and the Back propagation algorithm is used to train the network. The proposed algorithm is tested on a real life problem, the Wisconsin Breast Cancer Diagnosis problem. In a paper four missing value replacement methods are used, among these four methods, Median method gave the good result of 99.06%. Preprocessing using min-max normalization is used in this diagnosis. Further work is needed to increase the accuracy of classification of breast cancer diagnosis.

V REFERENCES

1) American Cancer Society, Cancer Facts and Figures 2007, 2007th ed. American Cancer Society, 2007.
2) U. S. Cancer Statistics Working Group, "United states cancer statistics 2003 incidence and mortality (preliminary data)," National Vital Statistics, vol. 53, no. 5, 2004.
3) Tuba Kiyan And Tulay Yildirim, "Breast Cancer Diagnosis Using Statistical Neural Networks" Istanbul University, *Journal Of Electrical And* Electronics Engineering, Year 2004, vol. 4, Number 2, pp.1149-1153
4) Anupam Shukla, Ritu Tiwari and Prabhdeep Kaur, "Knowledge Based Approach for Diagnosis of Breast Cancer" IEEE International Advance Computing Conference, Patiala, India, March 2009, pg 6-12
5) Sudhir D. Swarkar, Ashok Ghatol, Amol P. Pande, "Neural Network Aided Breast Cancer Detection and Diagnosis Using Support Vector Machine" Proceedings of the International conference on Neural Networks, Cavtat, Croatia, June 12-14, 2006, pp. 158-163.
6) W. H. Wolberg and O.L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," in Proceedings of the National Academy of Sciences, vol. 87, pp. 9193-9196, U.S.A., December 1990.
7) K. Bennett and O. L. Mangasarian, "Neural Network Training via Linear Programming," Advances in Optimization and Parallel Computing. Elsevier Science Publishers, 1992.
8) Renato De Leone, Rosario Capparuccia and Emanuela Marelli, "A Successive Overrelaxation Backpropagation Algorithm for Neural-Network Training" IEEE Transactions on Neural Networks, vol. 9, No. 3, May 1998, pg 381-388
9) Jun Zhang MS, Haobo Ma Md MS, "An Implementation of Guildford Cytological Grading System to diagnose Breast Cancer Using Naïve Bayesian Classifier", MEDINFO 2004, M.Fieschi et al. (Eds),Amsterdam:IOS Press
10) Punitha, C.P.Sumathi and T. Santhanam, "A Combination of Genetic Algorithm and ART Neural Network for Breast Cancer Diagnosis" Asian Journal of Information Technology 6 (1):112-117, 2007, Medwell Journals, 2007.
11) S.M. Kamruzzaman and Md. Monirul Islam, "Extraction of Symbolic Rules from Artificial Neural Networks" Proceedings of world Academy of science, Engineering and Technology, vol. 10, Dec. 2005, ISSN 1307-6884
12) Rudy Setiono and Huan Liu, "Neural-Network Feature Selector" IEEE Transactions On Neural Networks, vol. 8, No. 3, May 1997, pg 664-662
13) Wlodzislaw Duch and Rafal Adamczak and Krzysztof Grabczewski, "A New methodology of Extraction, Optimization and Application of Crisp and Fuzzy Logic Rules" IEEE Transactions On Neural Networks, vol. 12, No. 2, March 2001, pg 227-306
14) UCI Machine Learning Repository. [http://archive.ics.uci.edu/ml/]. Irvine, CA: University of California, Center for Machine Learning and Intelligent Systems.
15) D.W. Ruck,, S.K. Rogers., M.Kabrisky., P. S. Meibeck., and M. E. Oxley., "Comparatives Analysis of Backpropagation & the extended Kalman Filter for Training Multilayer perceptrons", IEEE Transactions on Pattern Analysis and Machine Intelligence, June 1992, Vol 14, No 6, pg 686-691
16) F.Paulin and A.Santhakumaran, "Extracting Rules from Feed Forward Neural Networks for Diagnosing Breast Cancer" CiiT International Journal of Artificial Intelligent Systems and Machine Learning, vol. 1, No. 4, July 2009, pg 143-146