International Journal of Emerging Engineering Research and Technology
Volume 3, Issue 7, July 2015, PP 172-178
ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)

Cancer Classification Using Neural Network

Sudip Mandal[1] and Indrojit Banerjee[2]

[1] H.O.D, ECE Department, GIMT, Krishna Nagar, India
[2]Student, ECE Department, GIMT, Krishna Nagar, India

ABSTRACT

Naturally, cells in human body grow and divide in a controlled way to produce more cells to maintain health. Cancer affects human body when abnormal cells divide without control and becomes able to invade other tissues. The genetic material (DNA) of these cells becomes damaged or changed that affects normal cell growth and division. Early diagnosis is of considerable significance of the physician's skills conducted based on their knowledge and experience yet an error might occur. A range of therapies have been provided by researchers already. Use of various Artificial Intelligence methods for medical diagnosis of diseases has recently become widespread. These intelligent systems help physicians as a diagnosis assistant. Now, various Artificial Neural Network, Rough Set, Decision Tree, Bayesian Network are very popular for this purpose. In this paper, Multi layer Feed Forward Neural Network was used to detect cancer from Microarray Data and UCI Machine Learning Data. Back Propagation Rule was used for training the model. Throughout this paper, two types of validations were performed: cross validation and new case testing for above two datasets with different combination of hidden layers and corresponding nodes. It was found that, NN model can classify the data with very good accuracy and this will lead to automated medical diagnosis system for the particular disease.

Keywords: *Cancer detection, Soft Computing, Neural Network, Microarray Data.*

INTRODUCTION

Cells divide and grow in our body throughout life. This process is controlled by different genes that bear different information about cells (*i.e.* functions, lifetimes of different cells etc.). For different reasons like excessive inhale of alcohol, smoke genetic mistakes occur and cells divide more often increasing chances of affecting cancer. Some chemicals in the smoke are also directly harmful to DNA, which further increases the chance of genetic mistakes. The genetic material (DNA) of a cell can become damaged or changed, producing mutations that affect normal cell growth and division. When this happens, cells do not die timely and new cells form when the body does not need them. The extra cells may form a mass of tissue called a tumor. In case of cancer, abnormal cells divide without control and are able to invade other tissues. Cancer cells can spread to other parts of the body through the blood and lymph systems.

A gene is the basic physical and functional unit of heredity. Genes, which are made up of DNA, act as instructions to make molecules called proteins. Deoxyribonucleic acid (DNA) is the chemical information database that carries the complete set of instructions for the cell as to the nature of the proteins produced by it, its life span, maturity, function and death. Genes are the working subunits of DNA. Normal human body has particular amount of protein nucleotide for each genes, that control functions of the genes, are called gene expression level. Any disease, like cancer, changes gene expression level. A gene can regulate or inhibit other genes. By observing the changes in gene expression levels, we can easily predict the status of the disease. Obviously all genes are not responsible or not activated for all diseases. If somehow we can control these genes we might be able to cure the disease for which early detection of cancer is necessary.

There are two ways available to detect cancer. In conventional method blood, urine or stool sample of patient is taken and tested in laboratory. This process is time consuming and requires skilled examiners. In this method the chance of error is relatively high. Soft computing is a new multidisciplinary field to construct new generation of Artificial Intelligence, known as computational

***Address for correspondence:**

sudip.mandal007@gmail.com

intelligence. The main goal of soft computing is to develop intelligent machines to provide solutions to real world problems, which are not modeled or too difficult to model mathematically. Its aim is to exploit the tolerance for approximation, Uncertainty, Imprecision and partial truth in order to achieve close resemblance with human like decision making. Several soft computing methods are already proposed like Neural Network, Fuzzy Logic, Data mining, Decision Tree, Bayesian network etc. [1, 2, 3].

Clinicians and patients need reliable information about an individual's risk of developing different disease. Ideally, they would have entirely accurate data and would be able to use a perfect model to estimate risk. Such a model would be able to categorize people with disease and others. Indeed, the perfect model would even be able to predict the timing of the disease's onset.

Breast cancer is a type of cancer originating from breast tissue. Worldwide, breast cancer accounts for 22.9% of all cancers (excluding non-melanoma skin cancers) in women. The first noticeable symptom of breast cancer is typically a lump that feels different from the rest of the breast tissue. The primary risk factors for breast cancer are female sex and older age. Other potential risk factors include: smoking, genetics, lack of childbearing or lack of breastfeeding higher levels of certain hormones, certain dietary patterns, and exposure to light pollution.

Lung Adenocarcinoma has been increasing in recent years, often beginning in the outer parts of the lungs and as such well-known symptoms of Lung Cancer such as chronic cough and coughing up blood may be less common until later stages in the disease. Early symptoms of Adenocarcinoma that may be overlooked include fatigue, mild shortness of breath, backache, shoulder ache, or chest pain.

In this paper, Artificial Neural Networks (ANN) is impended to classify the data of Breast Cancer and Lung Adenocarcinoma, which is inspired by information flow in biological neuron of human, provides main features, such as: flexibility, competence, and capability to simplify and solve problems in pattern classification, function approximation, pattern matching and associative memories [4]. Among many different ANN models, the multilayer feed forward neural networks (MLFF) have been mainly used due to their well-known universal approximation capabilities [5]. The success of ANN mostly depends on their design, the training algorithm used, and the choice of structures used in training. ANN are being applied for different optimization and mathematical problems such as classification, object and image recognition, Signal processing, seismic events prediction, temperature and weather forecasting, bankruptcy, tsunami intensity, earthquake, and sea level etc. [6], [7],[8], [15].

The remaining paper is organized as follow. Section II gives theoretical knowledge on NN, Data preprocessing and classification accuracy. Section III explains simulation results for different cases. Finally, the paper is concluded in the Section V followed by references.

THEORITICAL BACKGROUND

Before going through the proposed method, let's get some preliminary idea and theoretical concept of the ANN.

Artificial Neural Network (ANN)

Artificial Neural Network (ANN) derives its origin from the working of human brain. ANN is an information processing model which consists of multiple single processing units (neurons), these neurons are massively parallel in nature which performs highly complex computations. The sole goal of ANN is to make a computer learn something so that network would adapt to a given dataset. ANN, like people learn by example. These abilities make ANN suitable for pattern recognition, speech recognition or data classification problem.

The construction of the neural network involves three different layers with feed forward architecture. This is the most popular network architecture in use today. The input layer of this network is a set of input units, which accept the elements of input feature vectors. The input units (neurons) are fully connected to the hidden layer with the hidden units. The hidden units (neurons) are also fully connected to the output layer. The output layer supplies the response of neural network to the activation pattern applied to the input layer. The information given to a neural net is propagated layer-by-layer from input layer to output layer through (none) one or more hidden layers. Following is the simplest NN model.

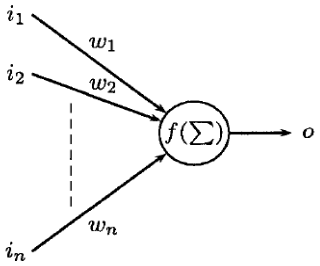


Figure1. *A simple ANN*

The factors $W_1, W_2, \ldots.. W_n$ are weights to determine the strength of input vectors $I = [I_1, I_2, \ldots.. I_n]^T$. Each input is multiplied by the associated of the neuron connection I^T W which can be given as following equation. The positive weights excite and the negative weights inhibit the node output.

$$I = I^T . W = I_1 W_1 + I_2 W_2 + \ldots\ldots + I_n W_n = \sum_{i=1}^{n} I_i\, W_i \qquad (1)$$

The nodes interval threshold ϕ is the magnitude offset. It affects the activation of node output *O* as:

$$O = f(I) = f\{ \sum_{i=1}^{n} I_i W_i - \phi_k \} \qquad (2)$$

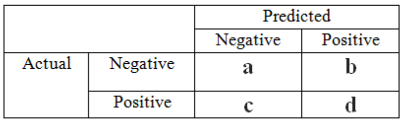
For Classification task, ANN needs to be trained for the networks to be able to produce the desired input output mapping. For training purpose a set of example data are feed to the network and connection weights, which is also called synaptic weight, are adjusted by using a learning algorithm. The objective of a neural network system is to give an output due some input signals. Before the training of the neural network, the system is initialized to its defaults values, and all the outputs (possible answers of the system) have the same probability. While the network is trained, the weights that define the connection between notes modified the value, and depending on the input and hidden values, the structure can be also changed. That implies that it is possible to optimize the neural networks modifying the structure of the solution and modifying the way that the weights are calculated. Here, FA approach has been used for the training of NN. For *m* number of training data the squared error can be given as

$$E = \sum_{i=1}^{m} (t_i - o_i)^2 \qquad (3)$$

Where *t* is the target output and *o* is the calculated output from training data. Different techniques are used in the past for optimal network performance for training ANNs such as Back Propagation Algorithm or Delta Rule [9], [10], [14].

Classification Accuracy

Accuracy of a network indicates how much correct output that network can produce. In other words, accuracy indicates reliability of a network in case of real world problem solving. Confusion Matrix:- In predictive analytics, a table of confusion or confusion matrix is a table with two rows and two columns that reports the number of false positives, false negatives, true positives, and true negatives. This allows more detailed analysis than mere proportion of correct guesses (accuracy). Accuracy is not a reliable metric for the real performance of a classifier, because it will yield misleading results if the data set is unbalanced (that is, when the number of samples in different classes vary greatly).



		Predicted	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

Figure2. *Confusion Matrix*

The entries in the confusion matrix have the following meaning in the context of our study: *a* is the number of correct predictions that an instance is negative, *b* is the number of incorrect predictions that an instance is positive, *c* is the number of incorrect predictions that an instance is negative, *d* is the number of correct predictions that an instance is positive. The accuracy (AC) is the proportion of the total number of predictions that were correct. It is determined using the equation:

$$AC = \frac{a+d}{a+b+c+d} \qquad (4)$$

EXPERIMENTAL RESULTS

Preprocessing of Data

The healthcare industry collects huge amounts of healthcare data and that need to be mined to discover hidden information for effective decision making. Discover of hidden patterns and relationships often go unexploited. There are several online databases available for the researchers. These databases are supported and contributed by different cancer institute and government organization.

The UCI Machine Learning Repository is a collection of databases, domain theories and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. For, the first phase of this work, we choose Breast Cancer data [13] with different status of the cells like Benign and Malignant. Objective is to learn the weight of the network so that maximum classification can be obtained. Initially the dataset consist of 699 instances and 10 attributes which are given as: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, Class: (0 for benign, 1 for malignant)

Another type of recently popular database is DNA microarray (known as DNA chip or biochip) which is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. In this paper we use the microarray dataset of Lung Adenocarcinoma are downloaded from National Centre for Biotechnology Information (NCBI) website with Geo Accession Number GSE10072 [11]. This dataset consist Gene expression signature of cigarette smoking and its role in Lung Adenocarcinoma development and survival of Homo sapiens. Initially the dataset consist of resulting in 107 final normalized expression values from 58 tumor and 49 non-tumor tissues from 20 never smokers, 26 former smokers, and 28 current smokers. Earlier study [12] shows that only 15 genes are responsible for the cancer. Therefore, we have microarray dataset of corresponding 15 genes for the construction of Neural Network. Following are the responsible gene's name.

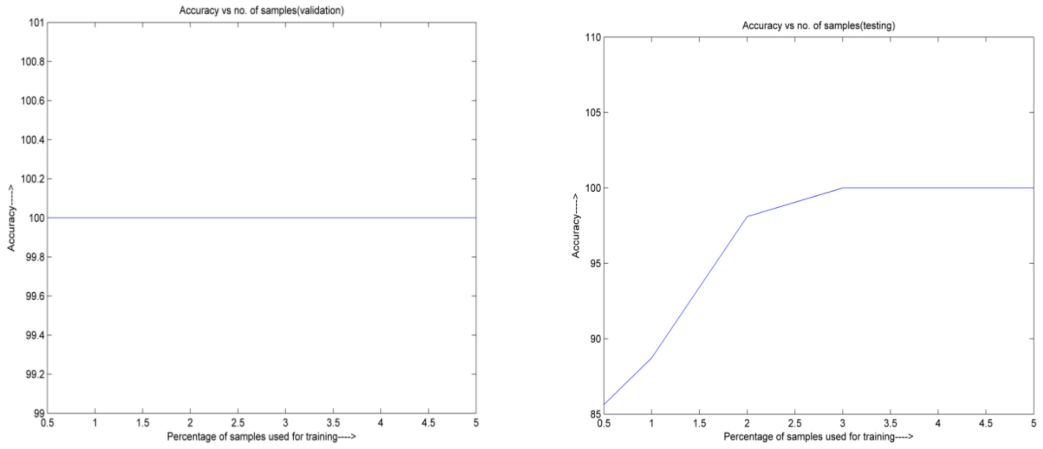
Table1. *Attributes or Genes-Id for reduced microarray dataset of Lung Adenocarcinoma*

201591_s_at	201772_at	201938_at	202295_s_a	203065_s_at
203091_at	203249_at	205261_at	206068_s_at	208056_s_at
209072_a	209613_s_at	218918_at	222313_at	49452_a

In this research, we have used two types of analysis, one is cross validation for the dataset that is used for training and another is testing for new dataset that is not used in training. All experiments were performed using Matlab 7.6 in Windows7 platform.

Case Study 1: Breast Cancer Classification

In the first phase of this research, we observe accuracy of the NN model for the Breast Cancer for the different amount of training data.



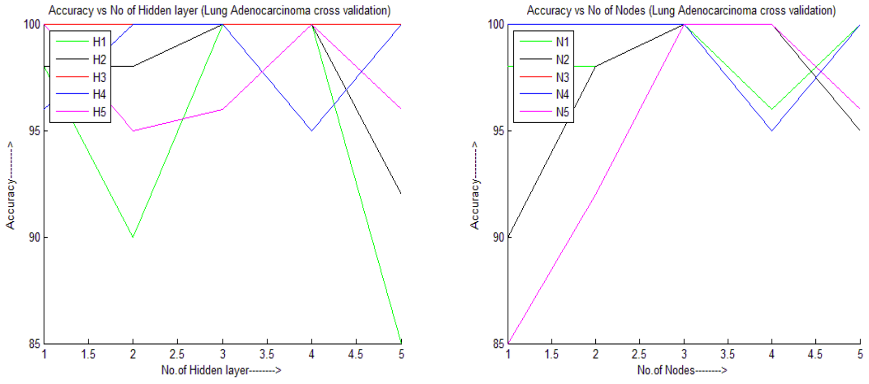
3(a) **3(b)**

Figure3(a). *Accuracy vs. percentage of samples for cross validation.* **3(b)** *Accuracy vs. no. of samples for testing new cases*

For this purpose, we used a single layer NN model for training. Initially we got the accuracy of 100% for validation and 85.63% for testing without any hidden layer by taking 0.5% of total data for Breast Cancer. But as more numbers of samples were used in training, the accuracy for testing increased up to 100% and became saturated for testing new case. The accuracy in validation was always constant to 100% which was expected and independent of number of samples used for training. The variations of accuracy with the percentage of training sample are given below. The reason behind this good classification accuracy was uniformity in the value of UCI data, small number of attributes/inputs and the large number of available samples for training.

Case Study 2: Lung Cancer Classification

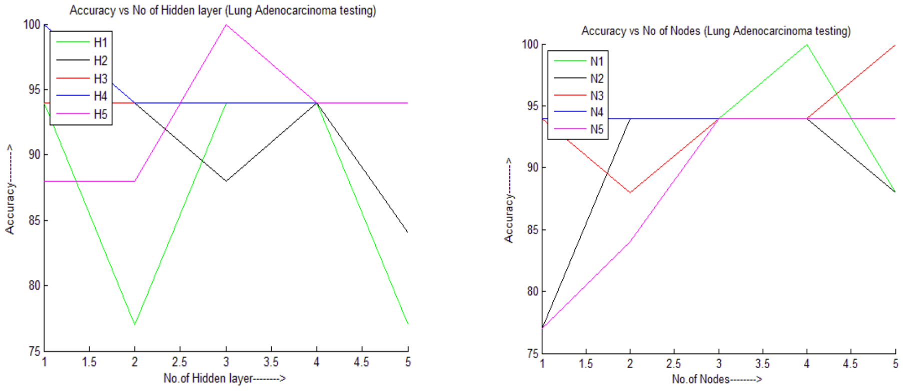
In next phase of this work, we have observed the effect on accuracy for classification of Lung Adenocarcinoma with the change of NN structure (number of hidden layer and the nodes in it). For both type of experiment, 80% of data was used for training purpose and rest of the data i.e. 20% was used for testing the network. As the microarray data is very noisy and number of available sample for training is also very small, therefore we got the initial accuracy of 96% for cross validation and 94 % for testing of new dataset with a single hidden layer which is quite satisfactory. The objective should be to increase the classification accuracy of the NN model. Therefore, one can increase the number of hidden layers and corresponding nodes to deal with more non linearity and noise of dataset. So, we gradually increased the complexity of the structure and kept record of the consequent accuracy. Fig. 4(a) and 4(b) shows accuracy vs. no. of hidden layers (while number of nodes is fixed) and accuracy vs. no. of nodes in each hidden layer (while number of hidden layer is fixed) respectively for cross validation. Fig 5(a) and 5(b) denote the same but for testing new cases.



4(a) 4(b)

Figure4. *Accuracy for cross validation vs.* **(a)** *No. of hidden layers.***(b)** *No. of nodes*

It was observed that as the number of hidden layers was increased, the accuracy was increased up to a certain number of hidden layers. If number of hidden layers was increased further the accuracy might decrease. Similarly, if the number of nodes was increased the accuracy was increased up to a certain number of nodes, but after that the number accuracy may decreases with increase in nodes or may vary due to increase in complexity of the structure.



5(a) 5(b)

Figure5. *Accuracy for testing new cases vs.* **(a)** *No. of hidden layers*. **(b)** *No. of nodes*

Moreover, if the nodes and number of hidden layers are increased, complexity of NN structure is also increased, hence computation time or epoch is also increased. In general, optimal number of hidden layer and node are required for efficient performance of the NN model with respect to the accuracy as well as computation time. By observing the graphs it is seen that for 3 hidden layers and 3 nodes the network becomes efficient in terms of both accuracy and number of epochs.

CONCLUSION

In this paper, a very popular soft computing technique Neural Network is implemented and tested on the two dataset of Breast Cancer and Lung Adenocarcinoma. Throughout the study, cross validation and percentage split for training & testing is performed using Matlab 7.6 software. From the analysis it is concluded that, NN techniques with conventional Back propagation training plays a major role in disease classification and prediction with great accuracy and efficiency.

In case of UCI dataset (breast cancer), we increase the percentage of training data from 0.5% to certain value and we observe that with the increase in number of samples of training data the accuracy increases gradually. So, it is concluded that the accuracy of the network increases with increased number of samples used for training. Also, the accuracy for a network trained with large number of dataset is high with respect to that which is trained with less number of datasets. Hence, by considering all these things it can be said that the accuracy is functions of dataset.

On the other hand, in case of NCBI dataset (lung cancer) we see that with the increase of no. of nodes and hidden layers of NN, the accuracy increases up to a certain level. It is obvious that the accuracy for cross validation is better than that for new testing cases as the same data is used for training during cross validation. However, it is observed that accuracy can be increased if we insert hidden layer between input and output layer as the data set is noisy and non-linear. The accuracy does not increase infinitely if the hidden layers or nodes or both are increased simultaneously. After an optimum value, if we further increase the number of hidden layers and the number of node for experiment purpose it is seen that the accuracy does not increase and the complexity of the network becomes high and it takes more time to be trained. Hence, by considering all these factors it can be said that the accuracy and training time are functions of dataset, hidden layers and number of nodes.

Therefore we can construct the automated diagnosis system using Neural Network that can predict the status of human body very accurately. Moreover Neural Network can be applied in any field of Engineering or Science where there is problem of classification. In future, different advanced algorithm may incorporate to get the more accuracy and efficiency.

REFERENCES

[1] J. Han and M. Kamber, "Data Mining; Concepts and Techniques," Morgan Kaufmann Publishers‖, 2000.

[2] T. Mitchell, "Machine Learning," McGraw Hill, 1997.

[3] R. Brachman, T. Khabaza, W.Kloesgan, G.Piatetsky Shapiro and E. Simoudis, "Mining Business Databases‖,Comm." ACM, Vol. 39, no. 11, pp. 42-48, 1996.

[4] J. E. Dayhoff, "Neural-Network Architectures: An Introduction," 1st ed., Van Nostrand Reinhold Publishers, New York. (1990).

[5] S. Haykin," Neural Networks, a Comprehensive Foundation," Prentice Hall, New Jersey, 1999.

[6] C. Guojin, Z. Miaofen et al.: Application of Neural Networks in Image Definition Recognition," Signal Processing and Communications, ICSPC. pp.1207-1210, 2007.

[7] M. Romano, S. Liong, et al.: Artificial neural network for tsunami forecasting, J. Asian Earth Sciences, vol. 36, pp. 29-37, 2009.

[8] M. Hayati, and Z. Mohebi," Application of Artificial Neural Networks for Temperature forecasting," J. World Academy of Science, Engineering and Technology, vol. 28 (2), pp.275-279, 2007.

[9] C. Leung, and C. Member, "A Hybrid Global Learning Algorithm Based on Global Search and Least Squares Techniques for back propagation neural network Networks," International Conference on Neural Networks, pp. 1890 -1895, 1994.

[10] N.M. Nawi, R.S. Ransing, M.N.M. Salleh, R.Ghazali, and N.A. Hamid, "An improved back propagation neural network algorithm on classification problems," J. Communications in Computer and Information Science (CCIS), vol.118, pp. 177-188, 2011.

[11] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10072.

[12] S. Mandal, G. Saha, R.K. Pal, “Reconstruction of Dominant Gene Regulatory Network from Microarray Data Using Rough Set and Bayesian Approach,” Journal Computer Science System Biology, vol. 6, issue 5 6, pp. 262-270, 2013.

[13] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29.

[14] S. Mandal, G. Saha and R. K. Pal; “Neural Network Training Using Firefly Algorithm,” Global Journal on Advancement in Engineering and Science, vol. 1, Issue 1, pp. 07-11, March 2015.

[15] S. Mandal, G. Saha & R. K. Pal ; “A Comparative Study on Disease Classification Using Different Soft Computing Techniques” published in “The SIJ Transactions on Computer Science Engineering & its Applications (CSEA)” Volume2, Issue 3, pp.59-66, August-2014.

AUTHORS’ BIOGRAPHY

Sudip Mandal received M.Tech. Degree in ECE from Kalyani Govt. of Engineering College. Now, he held the position of Head of Electronics and Communication Engineering Department in GIMT, Krishnanagar, India. He is also pursuing Ph.D. degree from University of Calcutta. His current research work includes Bioinformatics, Soft Computing and Tomography. The author is also member Computational Intelligence Society and Man, System & Cybernetics Society of IEEE. He published 3 National Conference Paper, 3 International Conference Paper and 7 International Journal so far.

Indrojit Baneerjee is final year student (B.Tech) of Electronics and Communication Engineering Department in GIMT, India. His current research work includes Artificial Intelligence.