Indian Journal of Science and Technology, *Vol 7(4), 421–425, April 2014*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Tuberculosis Disease Classification using Genetic-neuro Expert System

P. V. Geetha[1*], R. A. Lukshmi[1] and P. Venkatesan[2]

[1]Department of Mathematics, Meenakshi College for Women, Chennai–600 024, India; geetha.mcw@gmail.com, lukshmi67@yahoo.com
[2]National Institutes for Research in Tuberculosis, ICMR, Chennai-600 031, India; venkaticmr@gmail.com

Abstract

This study investigates the application of the hybrid technique Genetic-Neuro approach for Tuberculosis disease classification. Evolutionary algorithms are proved to be the efficient methods for optimization problems and their primary component namely Genetic Algorithm is used to select the significant features for Disease Classification. Artificial Neural Network is used for classification and the training is done by methods like Levenberg Marquardt algorithm. The construction process of the system is illustrated by using tuberculosis disease data. The results reveal that the hybrid technique Genetic-Neural system outperforms the conventional technique Artificial Neural Network for disease classification.

Keywords: Feature Selection, Genetic Algorithm, Neural Network, Tuberculosis

1. Introduction

Tuberculosis (TB) is an infectious disease caused by Mycobacterium Bacillus. It is spread through the air by a person suffering from TB. A single patient can infect 10–15 people in a year[1]. TB was declared a global health emergency in 1993, but it has been growing unchecked. Today, TB is causing millions of deaths every year globally[2]. Like any infectious disease, TB is prevalent even in developed countries. But it is a more serious problem in the developing and populous countries. India and China together account for nearly 40 per cent of the global burden[2].

Tuberculosis is one of India's major public health problems. According to WHO estimates, India has the world's largest tuberculosis epidemic. India accounts for one-fifth of the global TB incident cases[2]. Each year nearly 2 million people in India develop TB, of which around 0.87 million are infectious cases. It is estimated that annually around 3,30,000 Indians die due to TB[3].

Within the past decade it also has become clear that the spread of Human Immunodeficiency Virus (HIV) infection and immigration of persons from areas of high incidence have resulted in increased number of TB cases.

2. Machine Learning

Machine learning algorithms offer a principled approach for developing sophisticated, automatic and objective algorithms for analysis of high dimensional multi-modal biomedical data[4]. Machine learning, a branch of Artificial Intelligence is about the construction and study of models that can learn from data. A Machine Learning model could be trained on the data set to learn to distinguish the different patterns of the data sets which then can be used to classify a similar data[5,6]. A supervised learning system that performs classification is a classifier. The performance of the classification largely depends on the input features. Feature selection is performed on the input feature set which plays an important role in classifiers such as Neural Network. Some features might be irrelevant or redundant and selecting only the significant features is desirable[7]. In multi-layer perceptron the number of layers and the total number of neurons used in each layer is estimated by the number of input attributes[8,9]. In this work the genetic algorithm is used to choose the significant features which provide input features to the Neural Network.

*Author for correspondence

2.1 Neural Network

Neural Network processes information in a similar way as the human brain. The network is composed of a large number of interconnected processing neurons working in parallel to solve a specific problem. Neural Network learns through example. It has been used to establish complex relationships between input and output data. The supervised learning model proposed in this work aims at reducing the Mean Square Error (MSE) in order to have better accuracy in prediction. The model designed is a three layer model with input, hidden and output layers. The model would, on the basis of weights, predict the output/labels by incorporating the back propagation algorithm. The number of input nodes is equal to the number of input features (nine). The number of nodes in the hidden layer is generally selected on trial and error methods or 75% of the input nodes. There is only one output node, one to denote the incidence of TB and zero for non-incidence of TB. The Neural Network also takes in parameters like learning rate, learning time, etc[10].

2.1.1 Multi-layer Perceptron

Multi-layer perceptrons have been successfully used to solve a variety of problems by training them using a very powerful and a popular algorithm called the error back-propagation algorithm which is discussed in the next section. The fully connected basic architecture of a multilayer perceptron is shown in Figure 1. It has three main layers namely: input layer, hidden layer and the output layer.

The perceptron model can be divided into three modules:

1. Each neuron in the network has a nonlinear activation function which is smooth unlike the hard-limiting function used in Rosenblatt's perceptron. The most commonly used form of non-linearity that is in practice is the sigmoidal nonlinearity given by,

$$y_j = \frac{1}{1+\exp(-v_j)} \tag{1}$$

where v_j is the weighted sum of all the synaptic inputs of neuron j and y_j is the output of the neuron.

In this work, the sigmoidal function for the neurons in the input layer and the hyperbolic tangent function for the neurons in the hidden layer were chosen. The hyperbolic tangent function is given below.

$$\varphi_j(v_j(n)) = a\tanh(bv_j(n)), (a,b) > 0 \tag{2}$$

where, a and b are constants

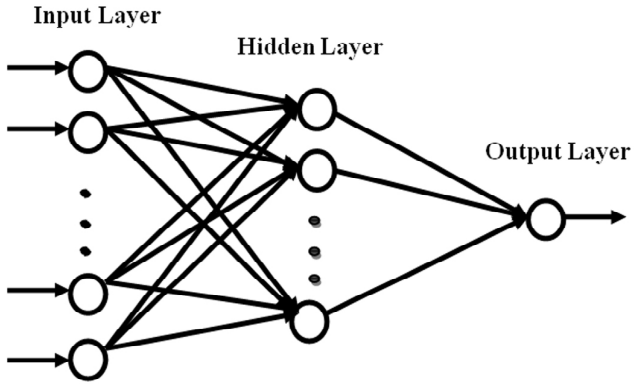

Figure 1. Architecture of MLP.

2. The network can contain one or more hidden layers of neurons that are not a part of the input or output layers. The hidden neurons provide a way to learn complex tasks by extracting meaningful features from input patterns.
3. The connectivity of the network is determined by the synapses of the network. A change in the connectivity of the network demands a change in the population of synaptic connections or their weights.

A combination of these characteristics results in the computing power of the multilayer perceptrons. The back-propagation algorithm provides a computationally efficient method for training the multilayer perceptron network[11–14].

2.1.2 Back-propagation Algorithm

The error signal from the output of neuron j for the nth training sample is defined by

$$e_j(n) = d_j(n) - y_j(n) \tag{3}$$

where j is an output node. The instantaneous value of the error energy for neuron j can be defined as $\frac{1}{2}e_j^2(n)$. The total instantaneous energy obtained by summarizing over all the neurons of the output layer is defined by $E(n)$. This can be written as:

$$E(n) = \frac{1}{2}\sum_{j\in C} e_j^2(n) \tag{4}$$

C denotes all the neurons in the output layer. If there are N patterns in the training set, the average squared error energy is obtained by summing E(n) overall n and normalized with respect to the set size N which is given by

$$E_{av} = \frac{1}{N}\sum_{n=1}^{N} E(n) \tag{5}$$

3. Genetic Algorithms

Genetic Algorithms are search algorithms based on the mechanics of the natural selection process of biological evolution. The most basic concept is that 'the strong tend to adapt and survive, while the weak tend to die out'. In the Genetic Algorithmic approach, optimisation is based on evolution and survival of the fittest concept. It uses techniques inspired by Darwin's Evolutionary Theory such as Inheritance, Selection, Crossover and Mutation[15,16].

Genetic Algorithm is a particular class of evolutionary computation which is a subfield of Artificial Intelligence. Genetic Algorithm solves optimisation problems by manipulating a population of chromosomes (encoded solution) to the problem. Each chromosome is assigned a fitness that is related to the success in solving the problem. Genetic Algorithm proceeds by choosing chromosomes to serve as parents and then replacing members of the current population with new chromosomes obtained after applying the Genetic Algorithm operators Crossover and Mutation. Chromosomes are chosen as parents for the reproduction process based on their fitness. The process will be repeated until a solution has been found which satisfies pre-defined termination criteria. Some of the stopping criteria may be number of generations, almost stable values of the chromosomes of the population and also the almost stable values of the average fitness of the population.

The use of fitness based reproduction generally leads to an improvement in the solution. Mutation operator ensures the entire state space will be searched and can lead the solution out of local minima. The most important parameters of Genetic Algorithm are population size, evaluation or fitness function, crossover method, mutation rate. Determining the size of the population is a critical factor. Choosing a population size too small increases the risk of converging prematurely to the local minima and a larger population size will make the convergence rate slow. Mutation rate is always low to ensure no bit position is stuck to single values. The basic principles of Genetic Algorithm and its applications are discussed in detail by many authors[7, 16, 17–20].

4. Receiver Operator Characteristic (ROC) Curve

One of the ways of evaluating the performance of a classifier is through its accuracy. Accuracy is defined to be the ratio of total number of correctly classified examples to the total number of available examples for a given operating point of a classifier. In disease classification problems with two predefined classes as positive diagnosis and as negative diagnosis, the classified test cases are categorized into four categories namely,

i. True Positives (TP) - Correctly Classified Positive Cases
ii. True Negatives (TN) - Correctly Classified Negative Cases
iii. False Positives (FP) - Misclassified Negative Cases
iv. False Negatives (FN) - Misclassified Positive Cases

$$\text{Therefore the accuracy is } \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

But using accuracy as a parameter for evaluating the classifier's performance has a drawback in clinical applications when the number of cases of prevalence class and the number of cases of the non-prevalence class differ largely. To overcome this problem ROC analysis is commonly used in clinical applications[21]. ROC curve describes the relation between two indices namely (i) True Positive Fraction (TPF) and (ii) False Positive Fraction (FPF). TPF is defined as

$$TPF = \frac{TP}{TP+FN} \quad (7)$$

$$FPF = \frac{FP}{TN+FP} \quad (8)$$

ROC curve plots TPF (sensitivity) vs. FPF (1-Specificity) for every possible decision threshold imposed on the decision variable. Generally the Area Under the ROC Curve (AUC) is used as a measure of performance.

5. Application-Tuberculosis Classification

For the purpose of model development, a database is created with the information gathered through a well-structured questionnaire to extract important features for the identification of tuberculosis. The database has 539 patients each with 9 features (input) –Age, Gender, Marital Status, Place Of Living, Literacy Level, Family Income, History Of Tuberculosis in the Family, habits such as smoking and drinking and Food Habits. Each record also has a binary output that reveals the presence or absence of the tuberculosis disease.

In this study a Multi-layer Neural Network structure with one hidden layer is developed for classification purpose, considering all the 9 features. The network is trained with Levenberg Marquardt Back Propagation Algorithm.

Genetic Algorithm is used to extract importance of individual feature in the data set[22]. The main objective of the feature selection problem is in finding a set in the space of all possible subsets of the feature set. Binary Chromosomes are defined for each case with a 1 to indicate the inclusion of the feature and 0 for the exclusion of the feature. The GA fitness function is chosen to be the neural network test error (MSE). The network is trained with 70% of the data. After the training is over, based on the Mean Square Error of the test data more successful chromosomes are selected to create new generations. A standard single point crossover is applied. Subsequently the mutation operator is applied on the chosen chromosome to obtain diversity with possible successful generations. By using genetic algorithm, the important features extracted are Age, Income, Location of residence, History of TB, Food, Habits, Education.

The Neural Network toolbox in MATLAB, Version 7.12.0.635 (R2011a)[23] has been used for the development of models.

The data set is divided into training set, validation set and test set using random data partition method given in Table 1. The Neural Network model correctly classified 86.4% of the test data whereas the GA-ANN model correctly classified 96.3% of the test data.

5.1 Classification Accuracy

The classification accuracy using Neural Network and Genetic Algorithm–Neural Network are given in Table 2. The MSE obtained in ANN and GA-ANN Models are tabulated in Table 3. The Figure 2 gives ROC for testing with respect to ANN and the Figure 3 gives the ROC for testing for GA-ANN.

6. Conclusion

In this paper, application of the hybrid technique GA–ANN has been proposed for classification of tuberculosis data set. Initially all the nine symptoms are taken into consideration and the ANN is trained with all the nine

Table 1. Data partition for training and testing

S.No	Data Partition Set	Records	Percentage
1	Training Set	377	69.96
2	Validation Set	81	15.02
3	Test Set	81	15.02
	Total	539	100

Table 2. Classification accuracy

Approach	Training	Validation	Testing
ANN	93.9%	86.4%	86.4%
GA-ANN	94.2%	90.1%	96.3%

Table 3. Mean square error rate

Approach	Training	Validation	Testing
ANN	0.0489	0.0305	0.0731
GA-ANN	0.0315	0.0316	0.0486

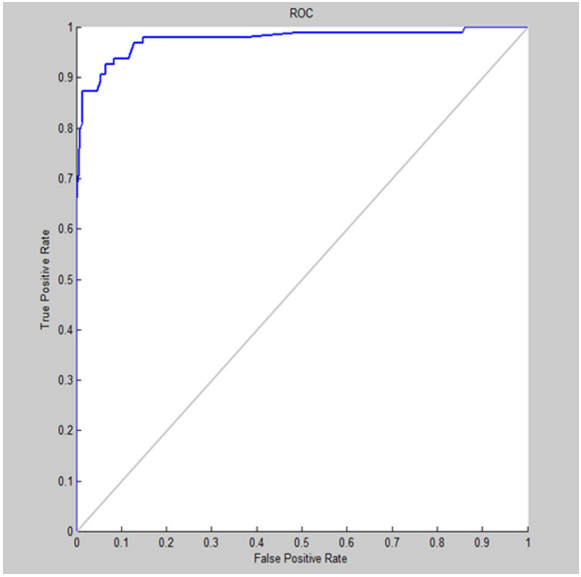

Figure 2. ROC for Testing-ANN.

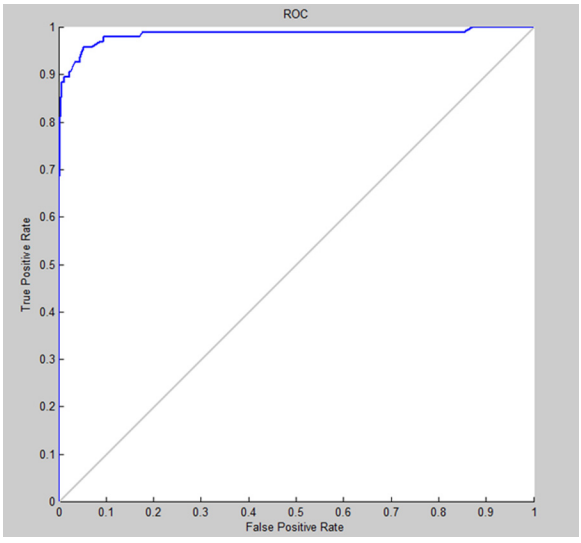

Figure 3. ROC for Testing-GA-ANN.

symptoms as input features for classification of tuberculosis disease. In GA-ANN approach, Genetic Algorithm is used for selecting the most relevant features by exploring the space of all possible sub-sect with the objectives of improving predict to accuracy and also to reduce the complexity of the problem. It is shown that employing feature selection in ANN classifier improves the classification accuracy. In view of the improved results obtained for classification accuracy, it is concluded that the approach based on GA–ANN with Levenberg Marquardt algorithm is capable of achieving better performance for determining the significant features for classification of tuberculosis disease.

7. References

1. WHO. Media centre Fact sheet N0104. Reviewed 2013 Feb.
2. Geneva WHO, World Health Organization. Global tuberculosis report12.WHO/HTM/TB/2012.6;2012. Available from: http://www.who.int/tb/publications/global_report/en/
3. WHO. Global tuberculosis control: surveillance, planning, financing; 2008.
4. Kayaer K, Yıldırım T. Medical diagnosis on Pima Indian diabetes using general regression neural networks. Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP). 2003 Jun. Istanbul. p. 181–84.
5. Dy JG, Brodley CE. Feature selection for unsupervised learning. J Mach Learn Res. 2004; 5:845–889.
6. Mitchell TM. Machine learning. Burr Ridge, IL: McGraw Hill; 1997.
7. Brill FZ, Brown DE, Martin WN. Fast generic selection of features for neural network classifiers. IEEE Trans Neural Network. 1992; 3(2):324–28.
8. Boger Z, Guterman H. Knowledge extraction from artificial neural network models. 1997 IEEE International Conference on Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation. Orlando, FL; 1997 Oct 12–15. 3030–3035.
9. Fogel DB, Wasson III EC, Boughton EM. Evolving neural networks for detecting breast cancer. Cancer letters. 1995; 96(1):49–53.
10. Alpaydin E. Introduction to machine learning. MIT press; 2004.
11. Haykin S. Neural Networks: A comprehensive foundation 2. NJ, USA: Prentice Hall PTR; 2004.
12. Hoang A. Supervised classifier performance on the UCI database. University of Adelaide; 1997.
13. Hornik K, Stinchcombe M, White H. Multilayer feed forward networks are universal approximators. Neural Network. 1989; 2(5):359–66.
14. Kecman V. Learning and soft computing: support vector machines, neural networks, and fuzzy logic models. Cambridge, MA: MIT press; 2001.
15. Beasley D, Martin RR, Bull DR. An overview of genetic algorithms: Part 1, Fundamentals. University computing. 1993; 1592):58–69.
16. Melanie M. An introduction to genetic algorithms. Cambridge, MA, USA: MIT Press; 1999.
17. Davis LD. Handbook of genetic algorithms; New York: Van Nostrand Reinhold; 1991.
18. Goldberg DE. Genetic algorithms in search. Optimization, and Machine Learning, Addison-Wesley, Reading, MA; 1989.
19. Harp SA, Smad T. Optimising neural network with genetic algorithm–Honeywell–SSDC. 1993:1138–43.
20. Holland JH. Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. U Michigan Press; 1975.
21. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition. 1997; 30(7):1145–59.
22. Kermani BG, White MW, Nagle HT. Feature extraction by genetic algorithms for neural networks in breast cancer classification. IEEE 17th Annual Conference Engineering in Medicine and Biology Society; 1995 Sep 20–25; Montreal, Que. p. 831–32.
23. MATLAB R2011a Student Version.