IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 1, Ver. II (Jan.-Feb. 2017), PP 84-89
www.iosrjournals.org

Application of Hybrid Genetic Algorithm Using Artificial Neural Network in Data Mining for the Diagnosis of Stroke Disease

Vijay Kumar Tiwari[1], Abhay Kumar Agarwal[2]
[1] *CSED, KNIT, INDIA.*
[2]*CSED, KNIT, INDIA.*

Abstract: *The main purpose of data mining is to extract knowledge from large amount of data. Artificial Neural network (ANN) has already been applied in a variety of domains with remarkable success. This paper presents the application of hybrid model for stroke disease that integrates Genetic algorithm and back propagation algorithm. Selecting a good subset of features, without sacrificing accuracy, is of great importance for neural networks to be successfully applied to the area. In addition the hybrid model that leads to further improvised categorization, accuracy compared to the result produced by genetic algorithm alone. In this study, a new hybrid model of Neural Networks and Genetic Algorithm (GA) to initialize and optimize the connection weights of ANN so as to improve the performance of the ANN and the same has been applied in a medical problem of predicting stroke disease for verification of the results.*

Keywords: *ANN, Back Propagation algorithm, Data mining, Feed Forward Network, Genetic algorithm, Hybrid model, Neuron.*

I. Introduction

Data mining is the process of discovering useful knowledge in data and also finding the inter-relation pattern among the data [7]. It is an automated discovery of strategic hidden patterns (useful information) in large amounts of raw data using intelligent data analysis methods [8]. For the past few years, there have been a lot of studies focused on the classification problem in the field of data mining [9, 10]. The general goal of data mining is to extract knowledge from large amount of data. The discovered knowledge should be predictive and comprehensible classification is given an equal importance to both predictive accuracy and comprehensibility. Neural Network is used for the classification of diseases based on the features of the patients.

A stroke is the sudden death of brain cells in a localized area due to inadequate blood flow. The sudden death of brain cells due to lack of oxygen, caused by blockage of blood flow or rupture of an artery to the brain. Sudden loss of speech, weakness, or paralysis of one side of the body can be symptoms. In medical diagnosis, the information provided by the patients may include redundant and interrelated symptoms and signs especially when the patients suffer from more than one type of disease of same category. The physicians may not able to diagnose it correctly. So it is necessary to identify the important diagnostic features of a disease and this may facilitate the physicians to diagnosis the disease early and correctly. The expert go for computer aided diagnosis (CAD) for confirming their prediction. The CAD helps to improve their prediction efficiency and accuracy. It should also be user friendly, so that expert can have the classification with explanation.

II. Related Work

In August 2012, Dharmistha. D.Vishwakarma, presented paper [1] on Genetic based weight optimization of Artificial Neural network. This papers shows the weights in different layers of the network are optimized using genetic algorithm comparison results for the ANN trained without GA and GA based ANN. In July 2012, P.Venkateshan and V. Premlatha, presented paper [2] on Genetic–Nero approach for disease classification ,which shows genetic neuro classification system performs better than the conventional neural network. In June 2012, Kafka Khan and Ashok Sahai presented a paper [3] in which the comparison of BA, BP, GA, PSO and LM algorithms for training feed forward neural network in e-Learning context was done. In May 2011, Asha Karegowada, A..S.Manjunath, M.A.Jayaram presented a paper [4] in which the Application of Hybrid model that integrates Genetic algorithm and Back Propagation network. In 2009, D. Shanthi, Dr. G. Sahoo, Dr. N. Saravanan presented the paper [5] on Evolving Connection Weights of Artificial Neural Networks Using Genetic Algorithm with Application to the Prediction of Stroke Disease , the real output and desired output of ANN and Hybrid ANN-GA were compared and it shows the classification accuracy for all surfaces were improved. In December 2008, D. Shanthi, Dr. G. Sahoo, and Dr. N. Saravanan presented a paper [6] on Input Feature Selection using Hybrid Neuro-Genetic Approach in the Diagnosis of Stroke Disease. In this paper, they proposed a neuro-genetic approach to feature selection in disease classification in this paper, a new hybrid neuro-genetic approach has been proposed and the same has been used for the selection of input features for the neural network. Experimental results showed the performance of ANN can be improved by selecting good combination of input variables.

III. Artificial Neural Network

A Neural Network (NN) consists of many Processing Elements (PEs), loosely called "neurons" and weighted interconnections among the PEs. Each PE performs a very simple computation, such as calculating a weighted sum of its input connections, and computes an output signal that is sent to other PEs. The training (mining) phase of a NN consists of adjusting the weights (real valued numbers) of the interconnections, in order to produce the desired output. The Artificial Neural Network (ANN) is a technique that is commonly applied to solve data mining applications. The previous neuron doesn't do anything that conventional computers don't do already. A more sophisticated neurons the McCulloch and Pitts model (MCP) as shown in figure 1. The difference from the previous model is that the inputs are "weighted", the effect that each input has at decision making is dependent on the weight of the particular input. The weight of an input is a number which when multiplied with the input gives the weighted input. These weighted inputs are then added together and if they exceed a pre-set threshold value, the neuron fires. In any other case the neuron does not fire as shown in figure 2.

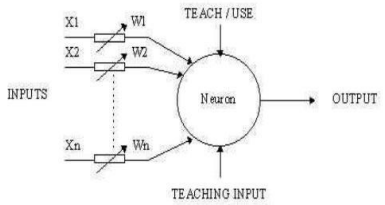

Figure 1: A MCP neuron

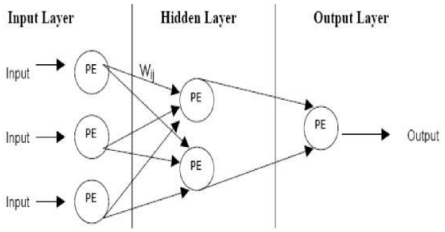

Figure 2: An example of a simple feed-forward network

IV. Back Propagation Algorithm

The back propagation algorithm (BP) [10] is a classical domain-dependent technique for supervised training. It works by measuring the output error, calculating the gradient of this error, and adjusting the ANN weights (and biases) in the descending gradient direction. Hence, BP is a gradient-descent local search procedure (expected to stagnate in local optima in complex landscapes). The squared error of the ANN for a set of patterns is the actual value of the previous expression depends on the weights of the network. The basic BP algorithm calculates the gradient of E (for all the patterns) and updates the weights by moving them along the gradient-descendent direction. This can be summarized with the expression $\Delta w = -\eta \nabla E$, where the parameter $\eta > 0$ is the learning rate that controls the learning speed. The pseudo-code of the BP algorithm is shown in figure 3.

Initialize Weights;
While not Stop-Criterion do
  For all i, j do
    $w_{ij} = w_{ij} - \eta \frac{\partial E}{\partial w_{ij}}$
  End For
End While

Figure 3: Pseudo code of back propagation algorithm

V. Genetic Algorithm

A GA is a stochastic general search method. It proceeds in an iterative manner by generating new populations of individuals from the old ones. Every individual is the encoded (binary, real, etc.) version of a tentative solution [11]. The canonical algorithm applies stochastic operators such as selection, crossover, and mutation on an initially random population in order to compute a new population. In generational GAs all the population is replaced with new individuals. In steady-state GAs (used in this work) only one new individual is created and it replaces the worst one in the population if it is better.

The total process is described as follows and pseudo code is given in figure 4:

1- Generate randomly an initial population;

2- Evaluate this population using the fitness function;

3- Apply genetic operators such selection, crossover, and mutation;
4- Turn the process "Evaluation Crossover mutation" until reaching the stopped criteria fixed in prior.

$t = 0$
Initialize: $P(0) = \{a_1(0), \ldots, a_\mu(0)\} \in I^\mu$
Evaluate: $P(0)$: $\{\Phi(a_1(0)), \ldots, \Phi(a_\mu(0))\}$
While $\mathcal{L}(P(t)) \neq$ true //Reproductive loop
    Select: $P'(t) = s_{\theta z}\{P(t)\}$
    Recombine: $P''(t) = \otimes_{\theta z}\{P'(t)\}$
    Mutate: $P'''(t) = m_{\theta m}\{P''(t)\}$
    Evaluate: $P'''(t)$: $\{\Phi(a_1'''(t)), \ldots, \Phi(a_\lambda'''(t))\}$
    Replace: $P(t+1) = r_{\theta r}(P'''(t) \cup Q)$
    $t = t + 1$
End While

Figure 4: Pseudo code of genetic algorithm

VI. Hybrid Model Of Genetic And Back Propagation Algorithm

Back propagation learning works by making modifications in weight values starting at the output layer then moving backward through the hidden layers of the network. BPN uses a gradient method for finding weights and is prone to lead to troubles such as local minimum problem, slow convergence pace and convergence unsteadiness in its training procedure. Unlike many search algorithms, which perform a local, greedy search, GAs performs a global search. GA is an iterative procedure that consists of a constant-size population of individuals called chromosomes, each one represented by a finite string of symbols, known as the genome, encoding a possible solution in a given problem space. The GA can be employed to improve the performance of BPN in different ways. GA is a stochastic general search method, capable of effectively exploring large search spaces, which has been used with BPN for determining the number of hidden nodes and hidden layers, select relevant feature subsets, the learning rate, the momentum, and initialize and optimize the network connection weights of BPN. GA has been used for optimally designing the ANN parameters including, ANN architecture, weights, input selection, activation functions, ANN types, training algorithm, numbers of iterations, and dataset partitioning ratio [11]. The new hybrid Neuro-Genetic approach is depicted in figure 5 and the same has been applied for the diagnosis of stroke disease. The result show this hybrid approach has the potential to eventually improve the success rate better than traditional ANN monolithic design.

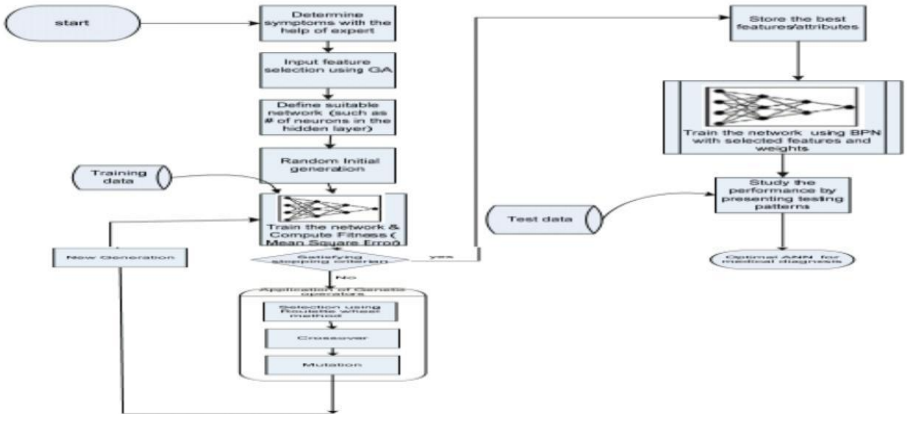

Figure 5: Hybrid Neuro - Genetic Approach

The process of GA-NN algorithm is presented below:

1. Determine the symptoms
2. Initialize the values of count=0, fitness=0 and number of cycles
3. Generate Initial Population. The chromosome of an individual is formulated as a sequence of consecutive genes, each one coding an input.

4. Design suitable network (input layer, hidden Layer, output layer)
5. Assign weights for each link
6. Train the network using BP algorithm
7. Find cumulative error and the fitness value on the basis of the fitness function.
8. If (previous fitness<current fitness value) Then store the current features
9. count = count +1
10. Selection: Two parents are selected using the roulette wheel mechanism
11. Genetic Operations: Crossover, Mutation and Reproduction to generate new features set (Apply new weights to each link)
12. If (number of cycles <= count) go to 4
13. Train the network with the selected features
14. Study the performance with test data.

VII. Study Of Previous Results And Discussion

For Diagnosis of Stroke Disease

The data for this work have been collected from 150 Patients who have symptoms of stroke disease. The data have been standardized so as to be error free in nature. Table 1 below shows the various input parameters for the Prediction of stroke disease.

Table 1: Input Parameters

Sr..No	Attributes
1	Hypertensive
2	Diabetes
3	Myocardial
4	Cardiac failure
5	Atrial fibrillation
6	Smoking
7	Blood cholesterol
8	Left arm and leg
9	Right arm and leg
10	Slurring
11	Giddiness
12	Headache
13	Vomiting
14	Memory deficits
15	Swallowing
16	Vision
17	Double vision
18	Vertigo
19	Numbness
20	Dizziness

7.1 Neural Network based Feature Selection

Data are analysed in the dataset to define column Parameters and data anomalies. Data analysis information needed for correct data pre-processing. After data analysis, the values have been identified as missing, wrong type values or outliers and which columns were rejected as unconvertible for use with the neural network. Feature selection methods are used to identify input columns that are not useful and do not contribute significantly to the performance of neural network. The removal of insignificant inputs will improve the generalization performance of a neural network. In this study, first backward stepwise method is used for input feature selection. This method begins with all inputs and it works by removing one input at each step. At each step, the algorithm finds an input that least deteriorates the network performance and becomes the candidate for removal from the input set. The architecture of the neural network designed with 20 input nodes, 10 hidden nodes, and 10 output nodes. This ANN is trained using Back propagation algorithm and tested with the data and overall predictive accuracy was shown in table 10 against different datasets.

7.2 GA Based Feature Selection

In this neuro-genetic approach all the 20 symptoms are taken in to account. GA optimizes the 20 inputs in to 14. The genotype is represented by a sequence of symptoms. The number of individual in the initial population is 20. The fitness function is represented by means of root mean square error. The maximum numbers of generations are fixed at 20 as shown in table 2.

Table 2: Parameters used in GA

Search Method	Genetic Algorithm
Population size	20
Number of generations	20
Probability of crossover	0.6
Probability of mutation	0.033
Random number seed	1

In this work, the probability of crossover is 0.6 and the probability of mutation is 0.033. These probabilities are chosen by trial and error through experiments for good performance. The data is partitioned randomly and the table 3 shows the no of records in the training set, validation set and test set.

Table 3: Data Partition Set

Sl.No	Data Partition set	Records	Percentage
1.	Training set	104	69.33%
2.	Validation set	23	15.33%
3.	Test set	23	15.34%
4.	Ignored set	0	0%
	Total	150	100%

The average prediction accuracy by the traditional Neural Network approach and the new hybrid Neuro-Genetic approaches are depicted in table 4 below:

Table 4: Average Prediction Accuracy

Approach	Training	Validation	Testing
Neural Networks	78.52%	82.43%	90.61%
GA-NN	79.17%	83.88%	98.67%

The result shows clearly that new hybrid neuro-genetic method provides better accuracy and faster convergence due to the complexity of the network. The prediction accuracy is 98.67% with the reduced features. Sometimes reduction of features yields the drop in accuracy. So the input parameters should be chosen without compromising the accuracy.

VIII. Conclusion

In this paper, hybrid neuro genetic approach has been used for the selection of input features for the neural network and results shows the performance of ANN can be improved by selecting good combination of input variables. GA-NN approach gives better average prediction accuracy than the traditional ANN.

References

[1] Ms. Dharmistha D.Vishwakarma "Genetic Algorithm based Weights Optimization of Artificial Neural Network" International Journal of Advanced Research in Electrical ,Electronics and Instrumentation Engineering(2278-8875),Volume1,Issue 3,August 2012.

[2] P.Venkatesan ,V.Premlatha," Genetic Neuro Approach for classification", International Journal of Science and Technology (2224-3577)Volume 2 No.7,July 2012

[3] Kafka Khan ,Ashok Sahai ," A Comparison of BA,GA,PSO,BP and LM for Training Feed Forward Neural Networks in e - Learning Context"I.J.Intelligent System and applications,2012,7,23-29.

[4] Asha Gowanda Karegowada,A.S.Manjunath ," Application of Genetic algorithm optimized Neural Network Connection Weights for medical diagnosis of PIMA Indian Diabetes "International Journal on SoftComputing(IJSC),Volume2,No.2,May 2011

[5] D.Shanthi, Dr.G.Sahoo, Dr.N.Saravanan "Evolving Connection Weights of Artificial Neural Networks Using Genetic Algorithm with Application to the Prediction of Stroke Disease", International Journal of Soft Computing Year: 2009 | Volume: 4 | Issue: 2 | Page No.: 95-102

[6] D.Shanthi, Dr.G.Sahoo, Dr.N.Saravanan," Input Feature Selection using Hybrid Neuro-Genetic Approach in he Diagnosis of Stroke Disease ",IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.12, December 2008

[7] S.Kalaiarasi Anbananthen, Fabian H.P. Chan, K.Y. Leong. Data Mining Using Decision Tree Induction of Neural Networks, 3rd Seminar on Science and Technology. Kota Kinabalu: SST. 2004.

[8] Kriegel, H.-P., et al.: Future Trends in Data Mining. Data Mining and Knowledge Discovery (5): 113-134, 2008.

[9] Kalaiarasi Anbananthen, Fabian H.P. Chan, K.Y. Leong.Data Mining Using Decision Tree Induction of Neural Networks, 3rd Seminar on Science and Technology. Kota Kinabalu: SST. 2004.An introduction to neural computing. Aleksander, I.and Morton, H. 2nd edition.

[10] Koohang, A. Foundations of Informing Science. T. Grandon Gill & Eli Cohen (eds.), (5): 113-134, 2008.

[11] Larranaga, P., Sierra, B., Gallego, M.J., Michelena,M.J., Picaza, J.M.,(1997). Learning Bayesian networks by genetic algorithms: a case study in the rediction of survival in malignant skin melanoma. Proc. Artificial Intelligence in Medicine Europe .