APPLICATION OF GENETIC ALGORITHM OPTIMIZED NEURAL NETWORK CONNECTION WEIGHTS FOR MEDICAL DIAGNOSIS OF PIMA INDIANS DIABETES

Asha Gowda Karegowda[1] , A.S. Manjunath[2] , M.A. Jayaram[3]

[1,3] Dept. of Master of Computer Applications ,Siddaganga Institute of Technology, Tumkur, India

[1]ashagksit@gmail.com
[3]ma_jayaram@rediffmail.com

[2] Dept. of Computer Science, Siddaganga Institute of Technology, Tumkur India

[2]asmanju@gmail.com

Abstract

Neural Networks are one of many data mining analytical tools that can be utilized to make predictions for medical data. Model selection for a neural network entails various factors such as selection of the optimal number of hidden nodes, selection of the relevant input variables and selection of optimal connection weights. This paper presents the application of hybrid model that integrates Genetic Algorithm and Back Propatation network(BPN) where GA is used to initialize and optmize the connection weights of BPN . Significant feactures identified by using two methods :Decision tree and GA-CFS method are used as input to the hybrid model to diagonise diabetes mellitus. The results prove that, GA-optimized BPN approach has outperformed the BPN approach without GA optimization. In addition the hybrid GA-BPN with relevant inputs lead to further improvised categorization accuracy compared to results produced by GA-BPN alone with some redundant inputs.

KEYWORDS

Back Propagation Network, Genetic algorithm, connection weight optimisation.

1. INTRODUCTION

With the computerization in hospitals, a huge amount of data is collected. Although human decision-making is often optimal, it is poor when there are huge amounts of data to be classified. Medical data mining has great potential for exploring hidden patterns in the data sets of medical domain. These patterns can be used for clinical diagnosis. Neural Networks are one of many data mining analytical tools that can be utilized to make predictions for medical data. BPN uses the gradient based approach which either trains slowly or get struct with local minimum. Instead of using gradient-based learning techniques, one may apply the commonly used optimization methods such as Genetic Algorithms (GAs), Particle swarm optimization (PSO), Ant Colony optimization to find the network weights. GA is a stochastic general search method, capable of effectively exploring large search spaces, and has been used with Back Propagation Network (BPN) for determining the various parameters such as number of hidden nodes and hidden

layers, select relevant feature subsets, the learning rate, the momentum, and initialize and optimize the network connection weights. This paper presents the application of hybrid model that integrates Genetic Algorithm and BPN for diagnosis of Pima Indians Diabetes Database by finding the optimal network connection wieghts. For sake of completeness, BPN and GA have been explained in section 2 and 3 respectively. Section 4 elloborates the hybrid GA-BPN model and its applications in diverse fields. The high dimension data not only confuse the classifier but also increases testing and training time of BPN. Hence significant feature have been identified by two different methods: Decision tree and Correlation based feature selection. Feature selection has been explained in section 5 and the data model used for the experiment has been discussed in section 6 followed by results and conclusion in section 6 and 7 respectively.

2. BACKPROPOPAGATION NETWORKS (BPN)

BPN is an information-processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. Developing a neural network involves first training the network to carry out the desired computations. The feed-forward neural network architecture is commonly used for supervised learning. Feed-forward neural networks contain a set of layered nodes and weighted connections between nodes in adjacent layers. Feed-forward networks are often trained using a back propagation-learning scheme. 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. Neural networks have been criticized for their poor interpretability, since it is difficult for humans to interpret the symbolic meaning behind the learned weights. Advantages of neural networks, however, include their high tolerance to noisy data as their ability to classify patterns on which they have not been trained [1-4]

3. GENETIC ALGORITHMS (GA)

GA[5] is an optimization techniques inspired by natural selection and natural genetics. Unlike many search algorithms, which perform a local, greedy search, GA is a stochastic general search method, capable of effectively exploring large search spaces. A genetic algorithm is mainly composed of three operators: reproduction, crossover, and mutation. As a first step of GA, an initial population of individuals is generated at random or heuristically. The individuals in the genetic space are called chromosome. The chromosome is a collection of genes where genes can generally be represented by different methods like binary encoding, value encoding, permutation encoding and tree encoding. Gene is the basic building block of the chromosome. Locus is the position of particular gene in the chromosome.

In each generation, the population is evaluated using fitness function. Next comes the selection process, where in the high fitness chromosomes are used to eliminate low fitness chromosomes. The commonly used methods for reproduction or selection are Roulette-wheel selection, Boltzmann selection, Tournament selection, Rank selection and Steady-state selection. But selection alone does not produce any new individuals into the population. Hence selection is followed by crossover and mutation operations. Crossover is the process by which two-selected chromosome with high fitness values exchange part of the genes to generate new pair of chromosomes. The crossover tends to facilitate the evolutionary process to progress toward potential regions of the solution space. Different types of crossover by and large used are one point crossover, two-point crossover, uniform crossover, multipoint crossover and average crossover. Mutation is the random change of the value of a gene, which is used to prevent

premature convergence to local optima. Major ways that mutation is accomplished are random bit mutation, random gene mutation, creep mutation, and heuristic mutation. The new population generated undergoes the further selection, crossover and mutation till the termination criterion is not satisfied. Convergence of the genetic algorithm depends on the various criterions like fitness value achieved or number of generations [6-7].

4. THE HYBRID MODEL OF GA AND BPN

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[8-13]. 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 [14].

The hybrid GA-ANN has been used in the diverse applications. GA has been used to search for optimal hidden-layer architectures, connectivity, and training parameters (learning rate and momentum parameters) for ANN for predicting community-acquired pneumonia among patients with respiratory complaints [15]. GA has been used to initialize and optimize the connection weight of ANN to improve the performance ANN and is applied in a medical problem for predicting stroke disease[16]. GA has been used to optimize the ANN parameters namely: learning rate, momentum coefficient, Activation function, Number of hidden layers and number of nodes for worker assignment into Virtual Manufacturing Cells(VMC) application[17]. GA-ANN model has been experimented for of study of the heat transport characteristics of a nanofluid thermosyphon in a magnetic field where, GA is used to optimize the number of neurons in the hidden layer, the coefficient of the learning rate and the momentumof ANN[18].

The current paper illustrates the application of GA for initializing and optimizing the connection weights of BPN and has been experimented for PIMA dataset. The foremost step of the GA is representation of the chromosome. For the BPN with single hidden layer with m nodes, n input nodes and p output nodes the number of weights to be computed is given by (n+p)* m. Each chromosome is made up of (n+p)*m number of genes.

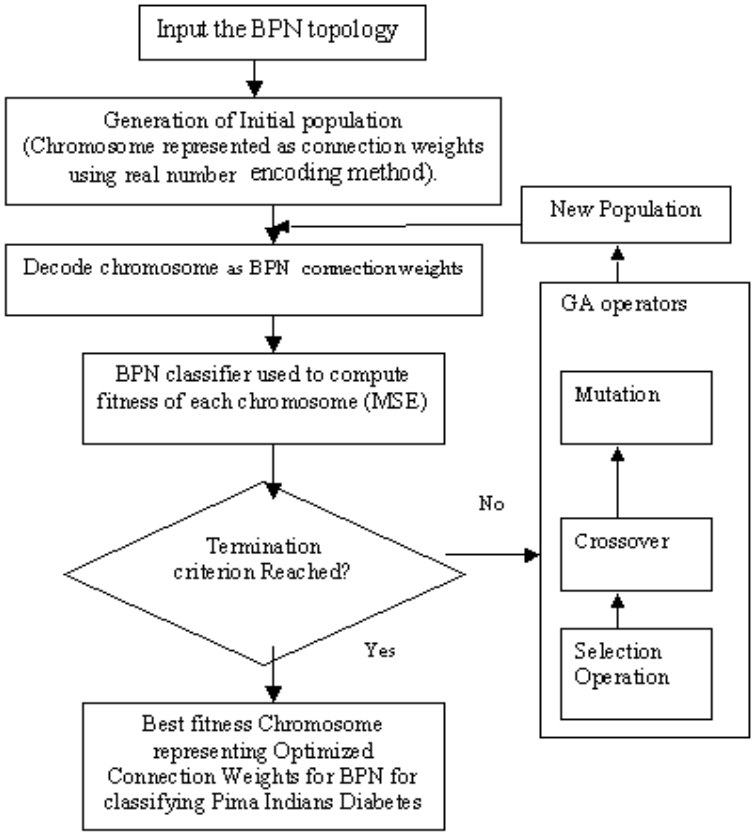

Figure 1: Working of hybrid GA-BPN for optimizing network connection weights

Genes are represented by real number encoding method. The original population is a set of N chromosome, which is generated randomly. Fitness of each chromosome is computed by minimum optimization method. Fitness is given by Fitness (Ci) = 1/ E for each chromosome of the population, where E is the error computed as root mean square error at the output layer as shown in equation 1 , where summation is performed over all output nodes pj and tj is the desired or target value of output o_j for a given input vector.

$$E = \frac{1}{2}\sum_{p}\sum_{j}(t_{pj} - o_{pj})^2 \qquad (1)$$

Once fitness is computed for the all the chromosomes, the best-fit chromosomes replace the worst fit chromosomes. Further crossover step is experimented using single point crossover, two-point crossover and multi point crossover. In addition a new type of crossover called mixed crossover has been used where for the given number of generation M, first 60% generation we applied multipoint crossover, followed by next 20% generation using two point crossover and remaining using one point crossover. Finally mutation is applied as the last step to generate the new population. The new population is given as input to PN to compute the fitness of each chromosome , followed by process of selection, reproduction, cross over and mutations to

generate the next population. This process is repeated till more or less all the chromosomes converge to the same fitness value. The weights represented by the chromosome in the final converged population are the optimized connection weights of the BPN. The working of hybrid GA-ANN for optimizing connection weights is shown in figure 1.

5. FEATURE SELECTION

Feature subset selection is of great importance in the field of data mining. The high dimension data makes testing and training of general classification methods difficult. Feature selection is an essential pre-processing method to remove irrelevant and redundant data. It can be applied in both unsupervised and supervised learning. In supervised learning, feature selection aims to maximize classification accuracy. The goal of feature selection for unsupervised learning is to find the smallest feature subset that best uncovers clusters form data according to the preferred criterion [19]. Authors have used two filters approaches namely Gain ratio and Correlation based feature selection for identifying relevant features. Decision tree is a simple tree like structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes. The non-terminal nodes are taken as relevant features. The basic decision tree induction algorithm ID3 was enhanced by C4.5. The WEKA classifier package has its own version of C4.5 known as J4.8. In the first method adopted for attribute selection, the authors have used J4.8 to identify the significant attributes [20]. In second method, the authors have used GA and Correlation based feature selection (CFS) in a cascaded fashion, where GA rendered global search of attributes with fitness evaluation effected by CFS. Genetic algorithm is used as search method with Correlation based feature selection as subset evaluating mechanism[21]. Experimental results show that the feature subsets selected by CFS filter resulted in marginal improvement for back propagation neural network classification accuracy when compared to feature subset selected by DT for PIMA dataset [22].

6. DATA FOR THE MODEL

6.1 Background

Diabetes can occur in anyone. However, people who have close relatives with the disease are somewhat more likely to develop it. Other risk factors include obesity, high cholesterol, high blood pressure and physical inactivity. The risk of developing diabetes also increases, as people grow older. People who are over 40 and overweight are more likely to develop diabetes, although the incidence of type-2 diabetes in adolescents is growing. Also, people who develop diabetes while pregnant (a condition called gestational diabetes) are more likely to develop full-blown diabetes later in life. Poorly managed diabetes can lead to a host of long-term complications among these are heart attacks, strokes, blindness, kidney failure, blood vessel disease [23].

6.2 Pima Indians Diabetes Database (PIDD)

The data used for the model is PIDD available http://www1.ics.uci.edu/~mlearn/ MLSummary.html. PIDD includes the following attributes (1-8 attributes as input and last attribute as target variable) number of times pregnant, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), Body mass index (weight in kg/ (height in m)^2), Diabetes pedigree function and Age (years)

Class to be predicted is patient is tested-positive or tested-negative. A total of 768 cases are available in PIDD. 5 patients had a glucose of 0, 11 patients had a body mass index of 0, 28 others had a diastolic blood pressure of 0, 192 others had skin fold thickness readings of 0, and 140 others had serum insulin levels of 0. After deleting these cases there were 392 cases with no missing values (130 tested positive cases and 262 tested negative). The data set was divided into Training and Test using 60-40 ratio.

7.Results

The number of generations, population size, number of nodes in input layer used with different number of hidden nodes in the hidden layer experimented for PIMA dataset is shown in Table 1. Among the various topologies experimented, the best performance of GA-BPN with 4 types of crossover with all 8 inputs is for 8-20-1 topology, with 5 inputs (Plasma, diastolic blood pressure, Body mass index, diabetes pedigree function and age) identified by DT (Plasma, insulin, Body mass index and Age) is for 5-15-1 topology and with 4 inputs (Plasma, insulin, Body mass index and Age) identified by GA-CFS is 4-10-1 topology is shown in table 2,3 and 4 respectively. For 8 inputs, the two points and multiple crossover resulted in slight improved accuracy compared to single point and mixture cross over. With 5 and 4 inputs, the single point and mixture cross over resulted in slight improved accuracy compared to two points and multiple crossover. Further table 3 and 4 shows the significance of relevant inputs identified by

DT and GA-CFS for improving the classfication of GA-BPN when compared with 8 inputs shown in table 2. Results of GA-BPN were compared with BPN alone with all 8 inputs, 5 inputs identified by DT and with 4 inputs identified by GA-CFS shown in table 5.

7.Conclusions

In this paper, application of hybrid GA_BPN has been experimented for classification of PIMA dataset. Back propagation learns by making modifications in weight values by using gradient method starting at the output layer then moving backward through the hidden layers of the network and hence is prone to lead to troubles such as local minimum problem, slow convergence pace and convergence unsteadiness in its training procedure. The optimal network connection weights can be obtained by using hybrid GA-BPN. GA is a stochastic general search method, capable of effectively exploring large search spaces, is used with BPN for determining the optimized connection weights of BPN. The hybrid GA_BPN shows substantial improvement in classfication accuracy of BPN. Significant features selected by DT and GA-CFS further enhanced classfication accuracy of GA-BPN.

Table 1. BPN Topologies with range of generation and population size.

Number of Inputs	Number of hidden nodes	Number of Generations	Number of Chromosomes
8 (all inputs)	8-20	100-200	20-60
5 (Decision Tree)	5-15	100-200	20-60
4 (GA_CFS)	4 - 10	100-200	20-60

Table 2. Classfication accuracy of GA-BPN with all inputs.

Topology	Number of Generations	Number of Chromosomes	Crossover type	GA_BPN Classification Accuracy (%)
8-20-1	175	60	Single Point	77.069
8-20-1	100	60	Two Point	77.707
8-20-1	175	60	Multiple	77.706
8-20-1	175	60	Mixture	77.068

Table 3. Classfication accuracy of GA-BPN with inputs identified by DT.

Topology	Number of Generations	Number of Chromosomes	Crossover type	GA_BPN Classification Accuracy (%)
5-15-1	150	60	Single Point	84.076
5-15-1	150	60	Two Point	83.439
5-15-1	150	60	Multiple	82.803
5-15-1	150	60	Mixture	84.076

Table 4. Classfication accuracy of GA-BPN with inputs identified by GA-CFS.

Topology	Number of Generations	Number of Chromosomes	Crossover type	GA_BPN Classification Accuracy (%)
4-10-1	175	60	Single Point	84.713
4-10-1	100	60	Two Point	84.712
4-10-1	175	60	Multiple	84.712
4-10-1	100	60	Mixture	84.713

Table 5. Comparison of Classfication accuracy of GA-BPN vs BPN alone

Attribute Selection Method	Number of Attributes	BPN Topology	BPN Accuracy (%)	GA-BPN Topology	GA-BPN Accuracy (%)
With All Attributes	8	8-24-1	72.88	8-20-1	77.707
Decision Tree	5	5-15-1	78.21	5-15-1	84.076
GA-CFS	4	4-8-1	79.5	4-10-1	84.713

REFERENCES

[1] S. Haykin,(1994), *Neural Networks- A comprehensive foundation,* Macmillan Press, New York.

[2] D.E. Rinehart, G.E. Hinton, and R. J. Williams, (1986), Learning internal representations by errorpropagation, In D.E. Rumelhart and J.L. McClelland, editors, Parallel Distributed Processing, Cambridge, MA: MIIT Press.

[3] H. Lu, R. Setiono and H. Liu, (1996),"Effective data mining using neural networks", IEEE Trans. On Knowledge and Data Engineering.

[4] A. Roy, (2000),Artificial neural networks – a science in trouble, SIGKDD Explorations.

[5] D. Goldberg, (1989)Genetic Algorithms in Search, Optimization , and Machine learning, Addison Wesley.

[6] http://www.myreaders.info/09_Genetic_Algorithms.pdf

[7] Berson Alex, Smith Stephen J. (1999) ,Data Warehousing, Data Mining, &OLAP., McGraw-Hill Book Co.

[8] Jihoon Yang, Vasant G. Honavar , (1998)"Feature Subset Selection Using a Genetic Algorithm", Journal IEEE Intelligent Systems, Volume 13 Issue 2.

[9] Brill, F., Brown, D., & Martin, W. (1992). "Fast Genetic Selection of Features for Neural Network Classifiers", IEEE Transactions on Neural Networks, 3(2), pp324-328.

[10] Arena P, Caponetto R, Fortuna L, Xibilia M G (1992), "Genetic algorithm to select optimal neural network topology", Proceedings of the 35th Midwest Symposium on Circuits and Systems 2: pp1381–1383.

[11] Maniezzo V,(1994) ,Genetic evolution of the topology and weight distribution of neural networks. IEEE Neural Network. 5: pp39–53

[12] Sexton R S, Dorsey R E, Johnson J D (1998) Toward global optimization of neural networks: A comparison of the genetic algorithm and back propagation. Decis. Support Syst. 22: pp171–185

[13] Rajasekaran, S and G. A Vijayalakshmi Pai (1996), Genetic Algorithm based Weight Determination for Backpropogation Networks, Proc of the Fourth Int Conf on Advanced Computing, pp 73-79).

[14] Osman Ahmed, Mohd Nord, Suziah Sulaiman, Wan Fatimah,(2009), "Study of Genetic Algorithm to Fully-automate the Design and Training of Artificial Neural Network",International Journal of Computer Science and Network Security, VOL.9 No.1.

[15] H. Paul S., G. Ben S., T. Thomas G., W. Robert S.,(2004), " Use of genetic algorithms for neural networks to predict community-acquired pneumonia", Artificial Intelligence in Medicine, Vol. 30, Issue 1, pp.71-84.

[16] D.Shanti, G. Sahoo , N. Saravanan, (2009), " Evolving Connection Weights of ANN using GA with application to the Prediction of Stroke Disease", International Journal of Soft Computing 4(2):pp95-102, Medwell Publishing.

[17] R.V. Murali, Member, IAENG, A.B.Puri, and G.Prabhakaran ,(2010), "GA-Driven ANN Model for Worker Assignment into Virtual Manufacturing Cells",Proceedings of the World Congress on Engineering 2010 Vol III, London, U.K.

[18] H. Salehi, S. Zeinali Heris*, M. Koolivand Salooki and S. H. Noei,(2011)," Designing a NN for closed Themosyphon with Nanofluid using a GA", Brazilian Journal of ChemicalEngineering ,Vol. 28, No. 01, pp. 157 – 168.

[19] Jennifer G. Dy, (2004),Feature Selection for Unsupervised Learning, Journal of Machine Learning, pp845-889.

[20] M.A.Jayaram, Asha Gowda Karegowda,(2007)," Integrating Decision Tree and ANN for Categorization of Diabetics Data", International Conference on Computer Aided Engineering, IIT Madras, Chennai, India.

[21] Asha Gowda Karegowda and M.A.Jayaram, (2009),"Cascading GA & CFS for feature subset selection in Medial data mining", IEEE International Advance Computing Conference, Patiyala, India

[22] Asha Gowda Karegowda, A. S. Manjunath & M.A.Jayaram,(2010), Comparative study of attribute selection using Gain ratio and correlation based feature selection, International Journal of Information Technology and Knowledge Management, Volume 2, No. 2, pp. 271-277.

[23] Editorial, (2004),Diagnosis and Classification of Diabetes Mellitus, American Diabetes Association, Diabetes Care, vol 27, Supplement 1.

AUTHORS

Asha Gowda Karegowda received her MCA degree and M.Phil in Computer Science in 1998 and 2008 from Bangalore University and Madurai Kamraj University, India respectively. She is currently pursuing her Ph.D under Visvesvaraya Technological University, Belgaum,India. She is working as Associate Professor in the Dept of Master of Computer Applications, Siddaganga Instituteof Technology, Tumkur, India. Her research interests are soft computing, image analysis and medical data mining. She has published few papers in International conferences and International Journals.

A.S. Manjunath received his M.Tech and PhD in Computer Science 1988 and 2003 from Mysore University and Bangalore University, India respectively. He is working as Professor in the Dept of Computer Science and Engineering, Siddaganga Institute of Technology,Tumkur, India. Her research interests are Embedded Systems and solutions, Networking and communications and soft computing. He has published few papers in International conferences and International Journals.

M.A. Jayaram received his M.Tech in Civil ,MCA degree and PhD from Bangalore University , IGNOU University , and Visvesvaraya Technological University, Belgaum, India in the year 1987, 2002 and 2008 respectively. He is working as Director in the Dept of Master of Computer Applications, Siddaganga Institute of Technology,Tumkur, India. His research interests are soft computing, image analysis and medical data mining. He has published few papers in International conferences and International Journals.