Effective Classification Technique Enhanced Using Genetic Algorithm: For Data Mining Disease in the Incumbents to the Health Centre

Manaswini Pradhan[1] and Dr. Ranjit Kumar Sahu[2]

[1]Lecturer, P.G. Department of Information and Communication Technology,
Fakir Mohan University, Orissa, India
E-mail: ms.manaswini.pradhan@gmail.com

[2] Consultant, Plastic, Cosmetic and Laser Surgery, Mumbai, India
E-mail: drsahurk@yahoo.com

***Abstract**: The diagnosis of disease is a vital and intricate job in the incumbents to the health centre. The proposed method combines the learning algorithm of BP neural network with genetic algorithm to train BP network and optimize the weight values of the network in a global scale. This method is featured as global optimization, high accuracy and fast convergence. The data-mining model based on genetic neural network is selected as the enhanced classifier and has been widely applied to the procedure of data mining on case information of incumbents in the reception counter of a health centre. It achieves an excellent effect for assisting health professionals to solve cases and make good decisions. In this paper, the principles and methods of this data-mining model are described in details. A real case of its application is also presented which predicts the disease in the incumbents to the health centre. From this case we can draw a conclusion that the data-mining model we have chosen is scientific, efficient, robust and practicable.*

Keywords: *Data Mining, Data Warehouse, BP Neural Network, Genetic Algorithm, Data Cleaning, Case Analysis*

1. Introduction

Data mining is the method of extraction of information and knowledge that are hided in data, unknown by people and potentially useful from a huge amount of data with multiple characteristics that is incomplete, containing noise, fuzzy and random. As a kind of cross-discipline field that combines multiple disciplines including database technology, artificial intelligence, neural networks, statistics, knowledge acquirement and information extraction, nowadays data mining has become one of the most important research direction in the international realms of information-based decision making. Analyzing and comprehending data from different aspects, people use data mining methods to dig out useful knowledge and hidden information of prediction from a large amount of data that are stored in database and data warehouse. The methods include association rules, classification knowledge, clustering analysis, tendency and deviation analysis as well as similarity analysis. By finding valuable information from the analysis results, people can use the information to guide their business actions and administration actions, or assist their scientific researches. All of these provide new opportunities and challenges to the development of all kinds of fields related to data processing.

Data mining is applied to the procedure of data analyzing, processing, decision making and data warehouse. Data mining technologies assist in many social departments to make scientific and reasonable decisions. This has major contribution for the development of our society and economy. Data mining can be applied to various different realms. For instance, many sale departments use data mining technology to determine the distribution and the geographical position of the sale network, the purchase and stock quantities of every kind of goods, in order to find out the potential customer groups and adjust the strategies for sale. In insurance companies, stock companies, banks and credit card companies, people apply data mining technology to detect the deceptive actions of customers to reduce the commercial deceptions. Data mining has been also widely applied to medical treatment and genetic engineering and many other fields. In recent years, with the acceleration of the step of information construction in police departments and with the increment of its development level, data mining technology has also been applied to the health departments especially in the health centre to improve the hospital treatment. This paper mainly discusses the principle and the practical application of genetic neural network based data mining model in disease analysis of patients.

Data classification is a classical problem extensively studied by statisticians and machine learning researchers. It is an important problem in variety of engineering and scientific disciplines such as biology, psychology, medicines, marketing, computer vision, and artificial intelligence A.K. Jain *et al*(2000). The goal of the data classification is to classify objects into a number of categories or classes. There have been wide ranges of machine learning and statistical methods for solving classification problems. Different parametric and non-parametric classification algorithms have been studied R.O. Duda *et al*.(1973), Breitman. L *et. al*(1984), Buntine, W.L *et al*(1992),Cover, T. M. *et al*.(1967), Hanson R. *et al*.(1993), Michie,D, *et al*(1994), Richard, M.D *et al*.(1991) and Tsoi, A.C. *et al*(1991). Some of the algorithms are well suited for linearly separable problems. Non-linear separable problems have been solved by neural networks dealt by C. Bishop (1995), support vector machines V.N. Vapnik *et al*. (1971) etc.

Neural networks (NNs) are increasing in popularity due their ability to approximate unknown functions to any degree of desired accuracy, as demonstrated by Funahashi (1989) and Hornik *et al*. (1989). In addition, NNs can also do this without making any unnecessary assumptions about the distributions of the data. This makes it convenient for researchers, as it can include any input variables that they feel could possibly contribute to the NN model. Although, it is likely that irrelevant variables are introduced to the model, the NN is expected to learn sufficiently to ignore these variables during the training process. It does this by finding weights associated to these irrelevant variables that when plugged into the NN would generate values that simply zero each other out, thereby having no effect on the final output prediction.

Although this works fine for training data, when applied to observations that it has not seen (out-of-sample or testing data), these weights are going to generate values that are unlikely to zero each other out, causing additional error in the prediction. If, on the other hand, the research could identify the irrelevant variables, these variables could be excluded from the NN model and eliminate the possibility of introducing additional error when applied to out of-sample data. Although, it is convenient for researchers to be able to include all available input variables into the model to extract a good solution, it also has the detrimental effect of making the NN a ‘black box’ where they throw everything into the model but do not know why or how the network produces its output. Additional information about the problem can be obtained by identifying those inputs that are actually contributing to the prediction. For example, we could train a NN that predicts the disease in the incumbent to the health centre. As inputs, we could include patient’s information such as gender, age, education of the incumbent, disease history, salary level and bad habits etc. A NN model that can accurately predict this outcome as well as indicating the relevant inputs to the model would be very beneficial in identifying disease. By using the proposed algorithm to determine those variables that are relevant to prediction, additional information about the problem can be learned.

The next section includes a background of literature on back propagation and the genetic algorithm and describes the problem. The third section describes the Data mining model of enhanced supervised classifier. The fourth section describes the GA method used in this study, which includes the base algorithm. The fifth section describes the problem, how the data were generated and outlines how the GA determines the number of hidden nodes (architecture) and the training process. Reports of the results of the application of Data mining in the Health Centre to diagnose the disease in the incumbent are performed. The last section concludes with final suggestions.

2. Background Literature

Since the majority of NN research is conducted using gradient search techniques, such as back propagation, which require differentiability of the objective function, the ability for researchers to identify relevant variables, beyond trial and error, is eliminated. In this paper, a modified genetic algorithm is used for training a NN, which does not require differentiability of the objective function that will correctly distinguish relevant from irrelevant variables and simultaneously search for a global solution.

2.1 Backpropagation

Back propagation (BP) is currently the most widely used search technique for training NNs. BP’s original development is generally credited to Werbos (1993), Parker (1985) and LeCun (1986) and was popularized by Rumelhart et al. (1986a,b). Although many limitations to this algorithm have been shown in the literature (Archer and Wang, 1993; Hsiung *et al*., 1990; Kawabata, 1991; Lenard *et al*., 1995;Madey and Denton, 1988; Masson and Wang, 1990; Rumelhart et al., 1986a; Subramanian and Hung, 1990; Vitthal *et al*., 1995; Wang, 1995;Watrous, 1987;White, 1987), its popularity continues because of many successes. An additional limitation to BP, which this paper deals with, is its inability to identify relevant variables in the NN model. This inability stems from the gradient nature of BP, which requires the objective function (usually the sum of squared errors (SSEs)) to be differentiable. This requirement prevents any attempt to identify weights in the models that are unnecessary, beyond pruning of the network. Pruning the network is simply eliminating connections that have basically no effect on the error term. This can be done by trial and error, saliency of weights, and node pruning (Bishop, 1995). Also, there have been approaches to network construction, such as Cascade Correlation (Fahlman and Lebiere, 1990), which attempts to build parsimonious network architectures. A better approach might be to use an alternative algorithm, such as the GA, that is not dependent on derivatives to modify the objective function to penalize for unneeded weights in the solution. By doing so, the GA can search for an optimal solution that can identify those needed weights and corresponding relevant variables.

2.2 The Genetic Algorithm

The GA is a global search procedure that searches from one population to another for solutions, focusing on the area of the best solution as far as practicable, while continuously sampling the total parameter space. Unlike back propagation, the GA starts at multiple random points (initial population) when searching for a solution. Each solution is then evaluated based on the objective function. Once this has been done, solutions are then selected for the second generation based on how well they perform. Once the second generation is drawn, they are randomly paired and the crossover operation is performed. This operation keeps all the weights that were included in the previous generation but allows for them to be rearranged. This way, if the weights are good, they still exist in the population. The next operation is mutation, which can randomly replace any one of the weights in the population in order to find a solution so as to escape local minima. Once this is complete, the generation is ready for evaluation and the process continues until the best solution is found. The GA works well searching globally because it searches from many points at once and is not hindered by only searching in a downhill fashion like gradient techniques. Schaffer *et al*. (1992) found more than 250 references in the literature for research pertaining to the combination of genetic algorithms and NNs. In this research, the GA has been used for finding optimal NN architectures and as an alternative to BP for training. This paper combines these two uses in order to simultaneously search for a global solution and a parsimonious NN architecture. Schaffer (1994) found that most of the research using the GA as an alternative training algorithm has not been competitive with the best gradient learning methods. However, Sexton et al. (1998) found that the problem with this research is in the implementation of the GA and not its inability to perform the task. The majority of past implementations of the GA encode each candidate solution of weights into binary strings. This approach works well for optimization of problems with only a few variables but for neural networks with a large number of weights, binary encoding results in extremely long strings. Consequently, the patterns that are essential to the GA's effectiveness are virtually impossible to maintain with the standard GA operators such as crossover and mutation. It has been shown by Davis (1991) and Michalewicz (1992) that this type of encoding is not necessary or beneficial. A more effective approach is to allow the GA to operate over real valued parameters (Sexton, Dorsey, and Johnson, 1998). The alternative approach described in Sexton *et al*. (1998) also successfully outperformed back propagation on a variety of problems. This line of research in based on the algorithm developed by Dorsey and Mayer (1995) and Dorsey *et al*. (1994). The GA and its modifications used in this study follows in the next section. Since the GA is not dependent on derivatives, a penalty value can be added to the objective function that allows us to search not only for the optimal solution but also for one that identifies relevant inputs to the model.

2.3 Data Mining In Reception Counter of a Health Centre

2.3.1 The Meaning of Data Mining in Reception Counter of a Health Centre

Every day in the reception counter of any health centre, patients arrive with different diseases. These large numbers of disease cases are received with various approaches. The information has been input into database to form a large amount of disease case information. These disease case information has been archived annually and periodically to form a plenty of historical case resources. By inducing and analyzing these historical cases, physician and people can get some experiences and learn some lessons that can help them to solve cases and make decisions in the future for getting better and improved health facilities. Therefore, in order to assist health departments to solve cases rapidly and make decisions efficiently, we should synthesize and organize these historical data, use proper data mining models to discover the potential and useful knowledge behind the data, and then predict and analyze the important factors in the data including the rate of disease, the constitution of disease population, the disease age structure, the area distribution of disease, the developing tendency of disease, the means and approaches of disease, the hidden areas of disease and so on. At present all of these have become urgent tasks that need our health centres to accomplish in the procedure of data processing.

2.3.2 Steps of Data Mining

The data mining steps in the health centers mainly include two steps:

(1) Filtering, selecting, cleaning and synthesizing the archived historical case information, and then performing transformation, if necessary, and finally, loading data into data warehouse after the above processing.

(2) Choosing appropriate models and algorithms of data mining to find out the potential knowledge in data. By a number of analysis and comparison among various data mining models, we select the back propagation (BP) error neural network as the general-purpose calculation model in our data mining. We train the neural network with a supervised learning method and combine BP algorithm with genetic algorithm to optimize the values of weights. Further, we apply the trained model to the prediction, classification and rule extraction of the case information.

3. Data Mining Model of Enhanced Supervised Classifier

3.1 General Methods of Data Mining

Now-a-days data mining methods include statistical method, association discovery, clustering analysis, classification and regression, OLAP(On Line Analytical Processing), query tool, EIS(Executive Information System), neural network, genetic algorithm and so on. Because of its high sustenance to noise data, good ability of generalization, high accuracy and low error rate, neural network model possesses great advantages among data mining methods. Hence, it has become a popular tool in data mining.

3.2 Data Mining Model of BP Neural Network

BP neural network is a kind of feed forward network that is now in most common use. Generally it has a multi-layer structure that consists of at least three layers including one input layer, one output layer and one or more hidden layers. There are full connections between neurons in adjacent layers and no connection between neurons in the same layer. Based on a set of training samples and a set of testing data, BP neural network trains its neurons and complete the procedure of learning. The application of BP algorithm is suitable for data mining environment in which it is impossible to solve problems using ordinary methods. Therefore, we need the use of complex function of several variables to complete non-linear calculation to accomplish the semi-structural and non-structural decision-making supporting procedure. So in the procedure of data mining in the reception counter of a health centre, we choose the BP neural network model.
The basic structure of BP neural network is as follows:

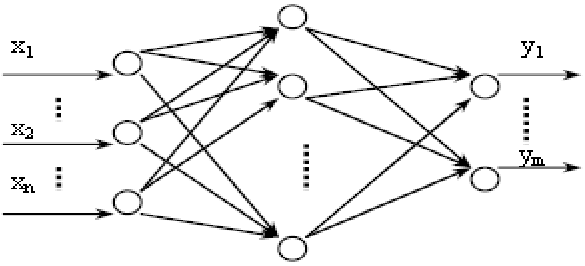

Fig. 1 The Structure of BP Neural Network

The learning procedure of neural network can be divided into two phases:

(1) The first one is a forward propagation phase in which a specified input pattern has been passed through the network from input layer through hidden layers to the output layer and becomes an output pattern.

(2) The second one is an error back propagation phase. In this phase, BP algorithm compares the real output and the expected output to calculate the error values. After that, it propagates the error values from output layer through hidden layer to input layer in the opposite direction. The connection weights will be altered during this phase.

These two phases proceed repeatedly and alternately to complete the memory training of network until it tends to convergence and the global error tends to minimum.

3.3 Learning Algorithm of Proposed Enhanced Classifier

In practical application of data mining, we use the three-layer BP neural network model that includes a single hidden layer and select differentiable Sigmoid function as its activation function. The function is defined as formula (1):

$$f(x)=1/(1+e^{-x}) \qquad (1)$$

The learning algorithm of BP neural network is described as follows:

(1) Setting the initial weight values *W (0)*: Generally we generate random nonzero floating numbers in [0, 1] as the initial weight values.

(2) Choosing certain numbers of pairs of input and output samples and calculate the outputs of network. The input samples are $X_{s=}(x_{1s,}x_{2s,\ldots,}x_{ns})$. The output samples are $t_s=(t_{1s,}t_{2s,\ldots,}t_{ms})$, s=1,2,…L. L is the number of input samples. When the input sample is the s^th sample the output of the i^th neuron is y_{is}:

$$y_{is}(t)=f\ (\sum_j w_{ij}(t)x_{js}) \qquad (2)$$

(3) Calculating the global error of network. When the input sample is the sample is the error of network. The calculating formula of is the s^th *sample E_s* is the error of network. The calculating formula of E_s is:

$$E_s(t)= \frac{1}{2\sum_k (t_{ks}-y_{ks}(t))^2}$$

$$= \frac{1}{2\sum_k e_{ks}{}^2(t)} \qquad (3)$$

where *k* represents the *k^th* neuron of output layer. $y_{ks}(t)$ the output of network when input sample is the sample *S^th* sample and the weight values has been adjusted *t* times. After training network *t* times based on all of the *L* groups of samples, the global error of all of these samples is:

$$G(t)=\sum_s E_s(t) \qquad (4)$$

(4) Determining if the algorithm ends.

$$G(t) \le \varepsilon \qquad (5)$$

When the condition of formula (5) is satisfied the algorithm ends. ε is the limit value of error that is specified beforehand. ε >0.

(5) Calculating the error of back propagation and adjusting the weights. The gradient descent algorithm has been used to calculate the adjustment values of weights. The calculating formula is as follows:

$$W_{ij}(t+1)=W_{ij}(t)-\ \eta \frac{\partial G(t)}{\partial w_{ij}(t)}$$

$$=W_{ij}(t)-\ \eta \sum_s \frac{\partial E_s(t)}{\partial w_{ij}(t)} \qquad (6)$$

where η is learning rate of network and also the step of weight adjustment.

3.4 Difficulties of the BP Network and the appropriate Solution

Because we use the gradient descent algorithm to calculate the values of weights, BP neural network still encounters problems such as local minimum, slow convergence speed and convergence instability in its training procedure. We combine two methods to solve these problems. One solution is to improve the BP network algorithm. By adding steep factor or acceleration factor in activation function, the speed of convergence can be accelerated. In addition, by compressing the weight values when they are too large, the network paralysis can be avoided. The improved activation function is defined with formula (7):

$$f_{a,b,\lambda}(x)=\frac{1}{1+e^{(x-b)/\lambda}}+\alpha \qquad (7)$$

where is α deviation parameter, *b* is a position parameter and *λ* is the steep factor.

Another solution to this is that the Genetic algorithm is a concurrence global search algorithm. Because of its excellent performance in global optimization, we can combine the genetic algorithm with BP network to optimize the connection weights of BP network. And finally we can use the BP algorithm for accurate prediction or classification.

4. Genetic Algorithm to Enhance BP Neural Network

4.1 The Principle of Genetic Algorithm

Genetic algorithm is a kind of search and optimization model built by simulating the lengthy evolution period of heredity selection and natural elimination of biological colony. It is an algorithm of global probability search. It doesn't depend on gradient data and needn't the differentiability of the function that will be solved and only need the function can be solved under the condition of constraint. Genetic algorithm has powerful ability of macro scope search and is suitable for global optimization. So by using genetic algorithm to optimize the weights of BP neural network we can eliminate the problems of BP network and enhance the generalization performance of the network.

The individuals in genetic space are chromosomes. The basic constitution factors are genes. The position of gene in individual is called locus. A set of individuals constructs a population. The fitness represents the evaluation of adaptability of individual to environment.

The elementary operation of genetic algorithm consists of three operands: selection, crossover and mutation. Selection is also called copy or reproduction. By calculating the fitness f_{i} of individuals, we select high quality individuals with high fitness, copy them to the new population and eliminate the individual with low fitness to generate the new population. Generally used strategies of selection include roulette wheel selection, expectation value selection, paired competition selection and retaining high quality individual selection. Crossover puts individuals in population after selection into match pool and randomly makes individuals in pairs to form parent generation. Then according to crossover probability and the specified method of crossover, it exchanges part of the genes of individuals that is in pairs to form new pairs of child generation and finally to generate new individuals. Generally used methods of crossover are one point crossover, multi point crossover and average crossover. According to specified mutation rate, mutation substitutes genes with their opposite genes in some loci to generate new individuals.

4.2 The Calculating Steps of Genetic Algorithm

The methods and steps of utilizing genetic algorithm to optimize the weights of BP network are described as follows:

(1) First, *k* groups of weights are given at random and assigned to *k* sets of BP networks. By training the networks, *k* groups of new weights has been calculated and adjusted. They constitute the original solution space.

(2) Using real number coding method these weights are coded to decimals and used as chromosomes. *k* groups of chromosomes comprise a population. So the original solution space has been mapped to search space of genetic algorithm. The length of gene string after coding is L=m×h+h×n . Where *m* is the number of neutrons in input layer, η is the number of neurons in hidden layer and *n* is the number of neurons in output layer.

(3) Using minimum optimization method the fitness function can be determined. The formula of fitness function is as follows:

$$f=\frac{1}{2G}=\frac{1}{\sum_{i=1}^{s}\sum_{j=1}^{m}(t_{ij}-y_{ij})^2} \qquad (8)$$

where is *S* the total number of samples, *m* is the number of neurons in output layer, *G* is the global error of all of *S* numbers of samples and y_{ij} is the output of network.

(4) The weights are optimized using genetic algorithm. We calculate the fitness and perform the selection with method of roulette wheel selection. After that, we copy the individuals with high fitness into next generation of the population. The next step is crossover. We crossover the individuals after selection with probability $P_{c.}$ Because we use real number coding method to code weights into decimals, the algorithm of crossover should be altered. If the crossover is performed between the i_{th} individual and the (i+1)th individual, the operand is as follows:

$$x_i^{i+1}=c_{i_i}*x_i^t+(1-c_i)*x_{i+1}^t$$

$$x_{i+1}^{i+1} = (1 - c_{i_i}) * x_i^t + c_i * x_{i+1}^t \qquad (9)$$

where is x_i^t, x_{i+1}^t a pair of individuals before crossover, x_i^{t+1}, x_{i+1}^{i+1} is a pair of individuals after crossover. c_i is a random datum of uniform distribution in [0,1] . With probability P_m, we mutate the individuals after crossover. If we mutate the ^{ith} individuals, the operand is

$$x_i^{i+1} = x_i^t + c_i \qquad (10)$$

where x_i^t is an individual before mutation, x_i^{t+1} is an individual after mutation, c_i is a random datum of uniform distribution in [$u_{\min} - \delta_1 - x_i^t, u_{\max} + \delta_2 + x_i^t$]. After once of these operations, a new population is generated. By repeating the procedure of selection, crossover and mutation, the weight combination is adjusted close enough to the most optimized weight combination.

(5) Finally, through the BP networks the weights can be adjusted delicately. Till now, the whole procedure of optimization ends.

With respect to every kind of prediction and analysis problems in the course of data mining, we extract proper sets of training samples and testing data, train mature neural network models with above-mentioned methods and apply the models to the future case analysis and prediction.

5. Application of the Data mining in the Health Centre to Diagnose the Disease

Finally, we give a real application of data mining in the reception counter of health centre as example. In this example we analyze patient's gender, age, educational degree, history of disease, chronic/acute, personal features, social relations and economical incomes. We find that to some extent these factors affect patient's social actions and habits that may lead patient to suffer a disease. Using these factors as input variables, a genetic neural network can be utilized to predict the present disease possibility of the patients..

5.1 Clean the Data in Database

In the first step, we fill up the missing data, smooth the noise data in database and solve the problems of same name for different meaning and different name for same meaning. And then, we load related data into data warehouse.

5.2 Select Training Samples of BP Networks

As in case archive databases, the case information are arranged in order of time, representative data can be obtained by random sampling. So we select samples by random sampling. To obtain the training sample set of BP networks, we select 5000 records from data warehouse. In addition, we extract other 2000 records as the testing sample set.

5.3 Normalize Samples

The most important input variables of BP network include gender, age, education degree, disease history, salary level and bad habits. The output of samples is the status (Yes or No) of whether these people suffer a disease at present. The output of BP network is the probability of people's present status (Percentage) of disease. Table 1 gives a list of first 10 samples of the total 5000 training samples.

By normalizing above input and output variables, the range of values of these variables has been mapped to the range of [0, 1]. The mapping relationship is given as follows:

5.3.1 Gender

Male: 1.0;
Female: 0.0

5.3.2 Age

0: 0.00;
1: 0.01;
2: 0.02; …;
100 and above: 1.0

No.	Sex	Age	Education Degree	Disease History	Salary Level	Bad Habits	Present Status of disease
1	M	25	Secondary school	Yes	3000--8000	No	Yes
2	M	32	Secondary school	No	1--3000	Yes	Yes
3	M	40	Primary School	Yes	3000--8000	Yes	Yes
4	F	30	Primary School	No	50000--80000	No	No
5	F	27	Secondary School	Yes	3000--8000	Yes	Yes
6	M	28	University	No	15000-30000	No	No
7	M	50	Junior University	No	8000--15000	No	No
8	M	38	Post-graduate	No	50000-80000	No	No
9	M	70	Primary School	No	1--3000	Yes	No
10	F	35	High School	No	3000--800 0	No	No

Table 1: Values of Input Variables

5.3.3 Education Degree

Illiterate: 0.0;
Graduate of Primary School: 0.125;
Graduate of Secondary School: 0.25;
Graduate of High School: 0.375;
Graduate of Junior University: 0.5;
Graduate of University: 0.625;
Postgraduate: 0.75;
Doctor: 0.875;
Post doctor: 1.0

5.3.4 Disease History

Yes: 1.0;
No: 0.0

5.3.5 Salary Level

None: 0.0;
Below 3000 Rupees: 0.125;
3000—8000 Rupees: 0.25;
8,000—15,000 Rupees: 0.375;
15,000—30,000 Rupees: 0.5;
30,000—50,000 Rupees: 0.625;
50,000—80,000 Rupees: 0.75;
80,000—1, 50,000 Rupees: 0.875;
1, 50,000 Rupees and above: 1.0

5.3.6 Bad Habits

Yes: 1.0; No: 0.0

5.3.7 Present Status of Disease

Yes: 1.0; No: 0.0

Table 1 gives the value list of first 10 samples of the total 5000 training samples after normalization.

No.	Sex	Age	Education Degree	Disease History	Salary Level	Bad Habits	Present Status of Disease
1	1	0.25	0.25	0	0.25	0	1
2	1	0.32	0.25	1	0.125	1	1
3	1	0.40	0.125	1	0.25	1	1
4	0	0.45	0.125	0	0.75	0	0
5	0	0.27	0.25	1	0.25	1	1
6	1	0.28	0.625	0	0.5	0	0
7	1	0.50	0.5	0	0.375	0	0
8	1	0.38	0.75	0	0.75	0	0
9	1	0.70	0.125	0	0.125	1	0
10	0	0.35	0.375	0	0.25	0	0

Table 2 Normalized Values of Input Variables

5.4 Build BP Neural Networks and Begin to Train

As it is observed, including above 6 important variables the total number of input variables is 10, we determined that the number of neurons in input layer is 10 and the number of neurons in output layer is 1. According to our experience and conforming to the principle of simplifying the network structure, we set the number of neutrons in hidden layer to 16. With above parameters we build 10 BP networks that have same structure. Then we generate 10 sets of small random numbers as initial weights of these networks and use the extracted 5000 samples as input and output samples of these networks. After that, we utilize BP algorithm to train the networks and get 10 sets of trained weights. The training times are 8000. After training we test the networks with our testing sample set. The generalization ability of our first network is shown as follows:

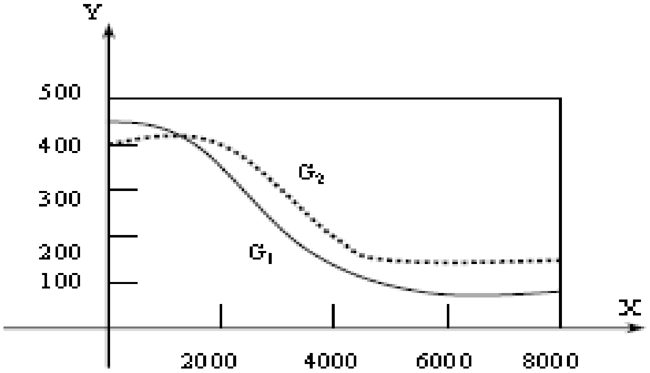

Fig. 2 The Generalization Ability of BP Network

where X is times of training, Y is the value of error, G_1 is global error of training sample set and G_2 is global error of testing sample set.

5.5 Utilize Genetic Algorithm to Optimize the Weight Values

We code the 10 sets of trained weights by real number coding method and use the weights after coding as chromosomes. 10 groups of chromosomes consist of a population. Then we optimize these weights using genetic algorithm until the weights, after decoding, are adjusted close enough to the most optimized weight combination.

5.6 Use the Optimized Weights to Train the BP Network Again

Finally, we use one of the BP network to adjust the optimized weights delicately. The training times for this adjustment are 4000. As a result, the generalization ability of the network is shown below:

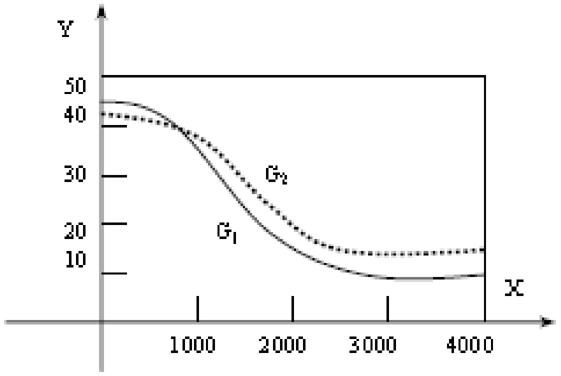

Fig. 3 The Generalization Ability of BP Network

5.7 Apply the Trained Network to Prediction and Analysis

We use the finally adjusted weights as the running weights of BP network to predict the probability of disease that people may suffer at present. The probability is the output of the BP network and is a float point number representing the occurrence probability of events. The prediction result by this process is highly accurate. In real terms of the disease diagnose of health centre, this prediction result can be used to guide the monitoring and tracing against the former attack of the disease to a person. At the same time, it can assist the lock and confirmation of suspects in case detection. Hence, it is highly useful and fool-proof method for case solving and decision-making.

6. Conclusion

GA was found to be an appropriate alternative to BP for training neural networks that not only finds better solutions with a parsimonious structure but can also identify relevant input variables in the data set. By using the GA in this manner, researchers can now determine those inputs that contribute to estimation of the underlying function. This can help with analysis of the problem, improved generalization, and network structure reduction. These results have demonstrated that a NN can be more than just a 'black box'. A complex chaotic time series problem as well as real-world problems could be solved that outperformed traditional NN training techniques as well as discovering relevant input variables in the model. Based on these results, future research is warranted for additional experiments and comparisons using the GA for NN training. BP neural network that has been applied to data mining possesses characteristics of high ability of memory, high adaptability, accurate knowledge discovery, none restriction to the quantity of data and fast speed of calculation. Based on using genetic algorithm to optimize the BP network can effectively avoid the problem of local minimum. Therefore, enhanced supervised classifier which is the proposed data-mining model has many advantages over other data mining models. In the real practice of data mining in the disease diagnosis in health centre the advantages have been fully embodied. This method has its own usefulness and is an effective prediction system to detect any type of diseases and at the same time has its beneficial effect upon the society.

References

1. Aoying Zhou, 2005. A Genetic-Algorithm-Based Neural Network Approach for Short-Term Traffic Flow Forecasting. Advances in Neural Networks, 3498, pp. 965-969.

2. Archer NP, Wang S. 1993. Application of the back propagation neural network algorithm with monotonicity constraints for two-group classification problems. Decision Sciences 24(1): 60–75.

3. Berson Alex, Smith Stephen J. Data Warehousing, Data Mining, & OLAP. McGraw-Hill Book Co, 1999

4. Bishop CM. 1995. Neural Networks for Pattern Recognition. Clarendon Press: Oxford.

5. Center for Computational Research in Economics and Management Science, MIT, Cambridge, MA.

6. D.E.Goldberg, 1989. Genetic Algorithms in Search, Optimization and Machine. Leaning, Addison-Wesley.

7. D.E.Goldberg, 1992. Genetic Algorithms: A Bibliography, IlliGAL Technical Report , 920008.

8. David Hard, 2003. Principles of Data Mining. Machine Industry Publisher, Beijing.

9. Davis L (ed.). 1991. Handbook of Genetic Algorithms.Van Nostrand Reinhold: New York.

10. Dorsey RE, Johnson JD, Mayer WJ. 1994. A genetic algorithm for training feed forward neural networks. In *Advances in Artificial Intelligence in Economics, Finance and Management* (Vol. 1), Johnson JD, Whinston AB (eds). JAI Press Inc.: Greenwich, CT; 93–111.

11. Dorsey RE, MayerWJ. 1995. Genetic algorithms for estimation problems with multiple optima, nondifferentiability, and other irregular features. *Journal of Business and Economic Statistics* 13(1): 53–66.

12. Fahlman SE, Lebiere C. 1990. The cascade-correlation learning architecture. In Advances in Neural Information Processing Systems (Vol. 2), Touretzky DS (ed.). Morgan Kaufmann: San Mateo, CA; 524–532.

13. Funahashi KI. 1989. On the approximate realization of continuous mappings by neural networks. Neural Networks 2(3): 183–192.

14. Guo Zhimao, 2003. An Extensible System for Data Cleaning. Computer Engineer, 29(3), pp. 95-96, 183

15. Heckerling Paul S, Gerber Ben S, 2004. Use of Genetic Algorithms for Neural Networks to Predict Community-Acquired Pneumonia. Artificial Intelligence in Medicine, 30 (1), pp. 71-75.

16. Hornik K, Stinchcombe M, White H. 1989. Multilayer feed-forward networks are universal approximators. Neural Networks 2(5): 359–366.

17. Hsiung JT, SuewatanakulW, Himmelblau DM. 1990. Should backpropagation be replaced by more effective optimization algorithms? Proceedings of the International Joint Conference on Neural Networks (IJCNN) 7: 353–356.

18. Kawabata T. 1991. Generalization effects of k-neighbor interpolation training. Neural Computation 3: 409–417.

19. LeCun Y. 1986. Learning processes in an Asymmetric threshold Network. Disordered Systems and Biological Organizations. Springer-Verlag: Berlin; 233–240.

20. Lenard M, Alam P,Madey G. 1995. The applications of neural networks and a qualitative response model to the auditor's going concern uncertainty decision. Decision Sciences 26(2): 209–227.

21. Li Mingqiang, 2002. The Principle and Application of Genetic Algorithm. Science Publisher, Beijing.

22. Li Yang, 2004. A Data Mining Architecture Based on ANN and Genetic Algorithm. Computer Engineer, 30(6), pp. 155-156.

23. Madey GR, Denton J. 1988. Credit evaluation with missing fields. *Proceedings of the INNS*, Boston, 456.
24. Masson E, Wang Y. 1990. Introduction to computation and learning in artificial neural networks. *European Journal of Operational Research* 47: 1–28.

25. Michalewicz Z. 1992. Genetic Algorithms + Data Structures = Evolution Programs. Springer: Berlin.

26. Parker D. 1985. Learning logic. Technical report TR-87. Parallel Distributed Processing: Exploration in the Microstructure of Cognition. MIT Press: Cambridge MA, 318–362.

27. Prechelt L. 1994. PROBEN1—A set of benchmarks and benchmarking rules for neural network training algorithms. Technical Report 21/94, Fakultat fur Informatik,Universit¨atKarlsruhe,Germany.Anonymous FTP:/pub/papers/techreorts/1994/1994-21.ps.gzon ftp.ira.uka.de.

28. Qing Guofeng, 2003. Acquirement of Knowledge on Data Mining. Computer Engineer, 29(21), pp. 20-22.

29. Rumelhart DE, Hinton GE, Williams RJ. 1986b Learning representations by back propagating errors.. Nature 323: 533–536.

30. Rumelhart DE, Hinton GG, Williams RJ. 1986a. Learning nternal Representations by Error Propagation. Parallel Distributed Processing: Exploration in the Microstructure of Cognition. MIT Press: Cambridge MA, 318–362.

31. Schaffer JD, Whitley D, Eshelman LJ. 1992. Combinations of Genetic Algorithms and Neural Networks: A survey of the state of the art, COGANN-92 Combinations of Genetic Algorithms and Neural Networks, *IEEE Computer Society Press*: Los Alamitos, CA; 1–37.

32. Schaffer JD. 1994. Combinations of genetic algorithms with neural networks or fuzzy systems. In Computational Intelligence: Imitating Life, ZuradaJM,

33. Schuster H. 1995. Deterministic Chaos: An Introduction. VCH: Weinheim, New York.

34. Sexton RS, Dorsey RE, Johnson JD. 1998. Toward a global optimum for neural networks: A comparison of the genetic algorithm and backpropagation. Decision Support Systems 22(2): 171–186.

35. Srinivas M., Lalit M.Patnaik, 1994. Genetic Algorithms: *A Survey. IEEE Computer*, 27(6), pp. 17-26.

36. Subramanian V, Hung MS. 1990. A GRG-based system for training neural networks: Design and computational experience. *ORSA Journal on Computing* 5(4): 386–394.

37. Vitthal R, Sunthar P, Durgaprasada Rao Ch. 1995. The generalized proportional-integral-derivative (PID) gradient decent back propagation algorithm. Neural Networks 8(4): 563–569.

38. Wang S. 1995. The unpredictability of standard back propagation neural networks in classification applications. Management Science 41(3): 555–559.

39. Wang Yu, 2005. Predictive Model Based on Improved BP Neural Networks and it's Application. Computer Measurement & Control, 13(1), pp. 39-42.

40. Watrous RL. 1987. Learning algorithms for connections and networks: Applied gradient methods of nonlinear optimization. *Proceedings of the IEEE Conference on Neural Networks* 2, San Diego, 619–627.

41. Werbos P. 1993. The roots of backpropagation: From ordered derivatives to neural networks and political forecasting. JohnWiley: New York.

42. White H. 1987. Some asymptotic results for backpropagation. *Proceedings of the IEEE Conference on Neural Networks* 3, San Diego, 261–266.

43. Xu Lina, 2003. Neural Network Control. *Electronic Industry Publisher*, Beijing.

44. Xu Zezhu, 2004. A Data Mining Algorithm Based on the Rough Sets Theory and BP Neural Network. Computer Engineer and Application, 31, pp. 169-175.

45. Zhang Liming, 1993. The Model and Application of Artificial Neural Network. Fudan University Publisher, Shanghai.

46. A.K.Jain, R.P.W. Duin, and J.Mao, Statistical Pattern Recognition: *A Review, IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22(1), January 2000, pp.4-37.

47. Breitman,L.,Friedman,J.H.,Olshen,R.A.,C.J.,Classification and Regression trees, Wadsworth, Belmont, CA, 1984.

48. Buntine,W.L., Learning classification trees, Statistics and Computing, 1992,pp. 63-73.

49. C. Bishop, Neural Networks for Pattern Recognition. New York: Oxford Univ. Press, 1995.

50. Cover, T.M., Hart,P.E., Nearest neighbors pattern classification, *IEEE Trans on Information Theory*, vol. 13, ,1967,pp. 21-27.

51. Hanson R.,Stutz,J.,Cheeseman,P., Bayesian classification with correlation and inheritance, Proceedings of the 12th *International Joint Conference on Artificial Intelligence 2*, Sydney,Australia,Morgan KaufSANN, 1992,pp. 692-698.

52. Michie,D. et al , Machine Learning, Neural and Statistical Classification, Ellis Horwood,1994.

53. R.O.Duda and P.E.Hard, Pattern classification and Scene Analysis, John wiley & Sons, NY, USA, 1973. Richard,M.D, LippSANN,R.P., Neural network classifiers estimate Bayesian a-posterior probabilities, Neural Computation ,vol.3, ,1991,pp. 461-483

54. Tsoi, A.C et al, Comparison of three classification Techniques, CART, C4.5 and multilayer perceptrons , *Advances in Neural Information Processing Systems*, vol. 3, 1991 pp.963-969.

55. V.N.Vapnik, A.Y. Chervonenkis, On the uniform convergence of relative frequencies of their probabilities, Theory of Probability and its Application, 1971, pp. 264-280.

Authors Profile

Manaswini Pradhan received the B.E. in Computer Science and Engineering, M.Tech in Computer Science from Utkal University, Orissa, India. She is into teaching field from 1998 to till date. Currently she is working as a Lecturer in P.G. Department of Information and Communication Technology, Orissa, India. She is currently persuing the Ph.D. degree in the P.G. Department of Information and communication Technology, Fakir Mohan University, Orissa, India. Her research interest areas are neural networks, soft computing techniques, data mining, bioinformatics and computational biology.

Dr Ranjit Kumar Sahu, M.B.B.S, M.S. (General Surgery), M.Ch. (Plastic Surgery). Worked as an Assistant Surgeon in post doctoral department of Plastic and Reconstructive Surgery, S.C.B. Medical College, Cuttack, Orissa, India. Presently working as a Consultant, Plastic, Cosmetic and Laser Surgery, Mumbai, India, He has five years of research experience in the field of surgery and published many national and international papers in medical field.