GENETIC NEURAL NETWORK BASED DATA MINING AND APPLICATION IN CASE ANALYSIS OF POLICE OFFICE

LIU Han-li, LI Lin, ZHU Hai-hong

School of Resource and Environment Science, Wuhan University, 129 Luoyu Road, Wuhan, P.R.China, 430079
Tel: 86-27-87882209 E-mail: liuhl000@sohu.com; lilin@telecarto.com

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

ABSTRACT:
This paper puts forward a method that 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 has been widely applied to the procedure of data mining on case information in the command centre of police office. It achieves an excellent effect for assisting people 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. From this case we can draw a conclusion that the data-mining model we have chosen is scientific, efficient and practicable.

1. INTRODUCTION

1.1 The Definition of Data Mining

Data mining is a procedure of extraction of information and knowledge that are hided in data, unknown by people and potentially useful from a large quantity of data with multiple characteristics that is uncompleted, containing noise, fuzzy and random. As a kind of cross-discipline field that syncretizes multiple disciplines including database technology, artificial intelligence, neural networks, statistics, knowledge acquirement and information extraction, nowadays data mining has becomes one of the most front research direction in the international realms of information-based decision making. Analysing 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.

1.2 The Application of Data Mining

Applied to the procedure of data analysing, processing and utilization as well as the procedure of decision making in many social departments, data warehouse and data mining technologies assist these departments to make scientific and reasonable decisions. This has profound meaning 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 police departments especially the command centre of police office. This paper mainly discusses the principle and the practical application of genetic neural network based data mining model in case analysis of police office.

2. THE MEANING AND METHOD OF DATA MINING IN COMMAND CENTRE OF POLICE OFFICE

2.1 The Meaning of Data Mining in Command Centre of Police Office

Every day in the command centre of police office, people receive a large number of information about cases received with various approaches. The information has been input into database to form a large amount of case information. These case information has been archived annually and periodically to form a plenty of historical case resources. By inducing and analysing these historical cases, people can get some experiences and learn some lessons that can help them solve cases and make decisions in the future. Therefore, in order to assist police 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 analyse the important factors in the data including the rate of crime, the constitution of crime population, the crime age structure, the area distribution of crime, the developing tendency of crime, the means and approaches of crime, the hidden areas of criminals and so on. At present all of these have become urgent tasks that need our police office to accomplish in the procedure of data processing.

2.2 Two Steps of Data Mining

The data mining procedure in the command centre of police office mainly includes two steps:

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

(2) Choosing appropriate models and algorithms of data mining to dig out the potential knowledge in data. By plenty of analysis and comparison among various data mining models, we select the error back propagation (BP) 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 NEURAL NETWORK

3.1 Common Methods of Data Mining

Present data mining methods include statistics 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 durability against noise data, good ability of generalization, high accuracy and low error rate, neural network model possesses great advantages among data mining methods. Now it becomes a popular tool in data mining.

3.2 Data Mining Model of BP Neural Network

BP neural network is a kind of feedforward 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 use 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 command centre of police office, we choose the BP neural network model.

The basic structure of BP neural network is as follows:

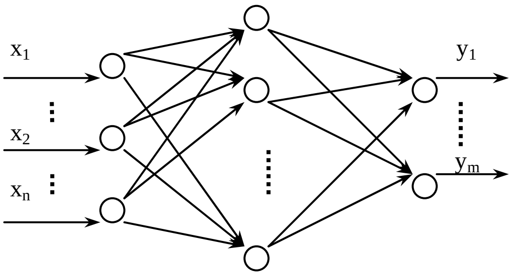

Fig. 1 The Structure of BP Neural Network

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

(1) The first one is a forward propagation phase in which a specified input pattern has been past 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 Neural Network

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):

$$f(x)=\frac{1}{1+e^{-x}} \tag{1}$$

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 $x_s=(x_{1s},x_{2s},\cdots,x_{ns})$. The output samples are $t_s=(t_{1s},t_{2s},\cdots,t_{ms})$, $s=1,2,\cdots,L$. L is the number of input samples. When the input sample is the s^{th} sample the output of the i^{th} neutron is y_{is} :

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

Where x_{js} is the j^{th} input of neutron i when the input sample is the s^{th} sample.

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

$$E_s(t) = \frac{1}{2\sum_k (t_{ks} - y_{ks}(t))^2} = \frac{1}{2\sum_k e_{ks}{}^2(t)} \quad (3)$$

Where k represents the k^{th} neutron of output layer. $y_{ks}(t)$ is the output of network when input sample is the 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:

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

(4) Determining if the algorithm ends.

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

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

(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:

$$w_{ij}(t+1) = w_{ij}(t) - \eta \frac{\partial G(t)}{\partial w_{ij}(t)} = w_{ij}(t) - \eta \sum_s \frac{\partial E_s(t)}{\partial w_{ij}(t)} \quad (6)$$

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

3.4 The Problem of BP Network and The 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):

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

Where a is a deviation parameter, b is a position parameter and λ is the steep factor.

Another solution is that: Genetic algorithm is a concurrence global search algorithm. Because of its excellent performance in global optimization, so 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. UTILIZING GENETIC ALGORITHM TO OPTIMIZE 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. Select 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 $L = m \times h + h \times n$. Where m is the number of neutrons in input layer, h is the number of neutrons in hidden layer and n is the number of neutrons 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 s is the total number of samples, m is the number of neutrons 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:

$$\begin{cases} X_i^{t+1} = c_i \cdot X_i^t + (1 - c_i) \cdot X_{i+1}^t \\ X_{i+1}^{t+1} = (1 - c_i) \cdot X_i^t + c_i \cdot X_{i+1}^t \end{cases} \qquad (9)$$

Where X_i^t, X_{i+1}^t is a pair of individuals before crossover, X_i^{t+1}, X_{i+1}^{t+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 i^{th} individual, the operand is:

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

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. A REAL INSTANCE OF APPLICATION

Finally, we give a real application of data mining in the command center of police office as example. In this example we analyse people's gender, age, education degree, history of crime, experiences, personal features, social relations and economical incomes. And we find that to some extent these factors affect people's social actions that may lead people to commit a crime. Using these factors as input variables, a genetic neural network can be utilized to predict the present crime possibility of these people.

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

Because in case archive database the case information is 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, crime history, salary level and bad habits. The output of samples is the status (Yes or No) of whether this people commit a crime at present. The output of BP network is the probability of people's present status (Percentage) of crime. Table 1 gives a list of first 10 samples of the total 5000 training samples.

No.	Sex	Age	Education Degree	Crime History	Salary Level	Bad Habits	Present Status of Crime
1	M	25	Secondary school	Yes	300--800	No	Yes
2	M	32	Secondary school	No	1--300	Yes	Yes

3	M	40	Primary School	Yes	300--800	Yes	Yes
4	F	30	Primary School	No	5000--8000	No	No
5	F	27	Secondary School	Yes	300--800	Yes	Yes
6	M	28	University	No	1500--3000	No	No
7	M	50	Junior University	No	800--1500	No	No
8	M	38	Post-graduate	No	5000--8000	No	No
9	M	70	Primary School	No	1--300	Yes	No
10	F	35	High School	No	300--800	No	No

Table 1 Values of Input Variables

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

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 Crime History
Yes: 1.0; No: 0.0

5.3.5 Salary Level
None: 0.0; Below 300 Yuan: 0.125; 300—800 Yuan: 0.25;
800—1500 Yuan: 0.375; 1500—3000 Yuan: 0.5;
3000—5000 Yuan: 0.625; 5000—8000 Yuan: 0.75;
8000—15000 Yuan: 0.875; 15000 Yuan and above: 1.0

5.3.6 Bad Habits
Yes: 1.0; No: 0.0

5.3.7 Present Status of Crime
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	Crime History	Salary Level	Bad Habits	Present Status of Crime
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

Because including above 6 important variables the total number of input variables is 10, we determined that the number of neutrons in input layer is 10 and the number of neutrons 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:

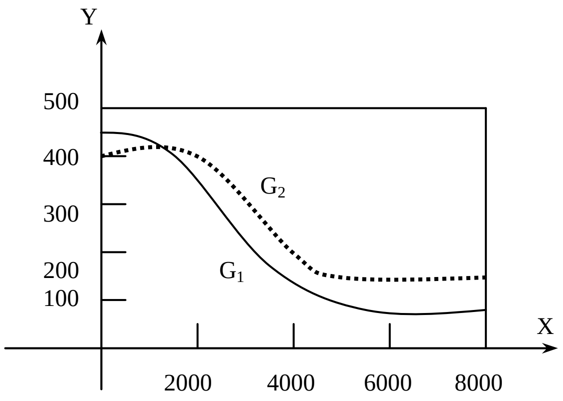

Fig. 2 The Generalization Ability of BP Network

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:

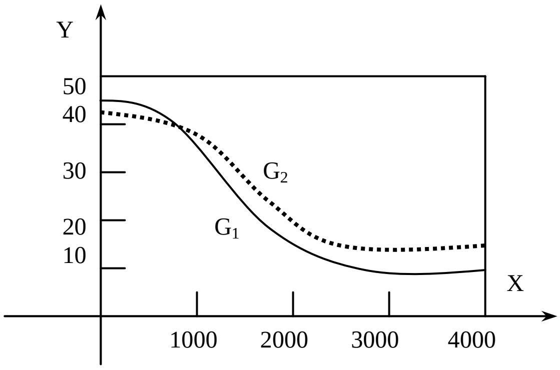

Fig. 3 The Generalization Ability of BP Network

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 crime that people may commit 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 is highly accurate. In real work of the command centre of police office, this prediction result can be used to guide the monitoring and tracing against the former criminals. At the same time, it can assist the lock and confirmation of suspects in case detection. So it is highly useful for case solving and decision-making.

6. CONCLUSION

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. Using genetic algorithm to optimize the BP network can effectively avoid the problem of local minimum. Therefore, the genetic neural network based data-mining model has many advantages over other data mining models. In the real practice of data mining in the command centre of police office, the advantages have been fully embodied.

References from Journals:

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

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.

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.

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

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

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

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

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

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

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

References from Books:

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

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

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

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

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