The Application of BP Neural Network principal component analysis in the Forecasting the Road Traffic Accident

He Ming, GuoXiucheng &LuGuangming

Transportation College of Southeast University
1107 room,YiFu Building, Southeast University
210096 Nanjing, China
Phone: (0)13814108366
Fax: 025-83795528
E-mail: heming9302@163.com, seuguo@163.com, gml0926@sohu.com

Abstract

According to complexity and comprehensibility of factors which affect road traffic safety, we use the method of principal component analysis to refine new factors which are linearly independent, then we forecast road traffic accident according to principal component by BP neural network simulation, analyse the relationship between traffic accident evaluating index and the causes of traffic accident, including people, vehicles, road and environment. At last we applied the method to a case, from the simulated result we can infer that the method of BP neural network simulation principal component analysis is superior to multinomial fitting and BP neural network simulation in the efficiency and precision.

Key words: road traffic accident, forecasting, BP neural network, principal component analysis

1 .Introduction

In the planning of road safety, traffic forecasting is one of the most basic tasks in the planning process, and it is the most important issue. Understanding future traffic accident Scientifically and accurately is of great significance for the overall grasp of road safety in order to prefabricate corresponding measures.

Existing traffic forecasting methods can be grouped into three general categories[1]: First extrapolation method, that is, they use the past to predict the future state of information, such as time series; Second, causality, that is, based on available information, to identify the relationship between the variables to predict the future state, such as regression analysis; Third, judge analysis, Experts predict the future state rely on past experience and the ability of the comprehensive analysis methods. Although the forecasting methods have their advantages, but due to the complex nature of the transport system and diversity of traffic accidents, most do not fit the data exist very well, extrapolation is not enough, and the forecasting results may deviate from the actual results and so on. Such as time series prediction, it uses the longitudinal data of the number of traffic accidents in the past to predict its movements over time. The process does not involve any other relevant factors. Although the regression analysis can forecast according to transverse and longitudinal data, but it establishes regression equation using just some historical data, the regression equation often considers only part of affecting factors. Therefore, the model is not accurate enough. Judge analysis is qualitative, it based on subjective experience, so the forecasting may not accurate.

In recent years, the rapid development of computer and artificial intelligence technology provide traffic modeling and forecasting with new methods. Artificial neural network is composed of neurons with different functions, it can be used to simulate, operate and reason the complicated nonlinear system through neural network interaction. It has extensive adapting ability, learning ability, mapping ability, and can approach to any nonlinear function in theory. In multivariable nonlinear system modeling, it has made remarkable achievements. BP neural network structure is intuitive, and it is the most widely used neural network. While Using BP neural network system to simulate, the first is to identify the factors. Traffic is very complex, we often adapt qualitative analysis to find all accident factors so as to avoid major factor be missed .However, when input variables are too many, it will obviously add to the complexity of the network, reduce network performance, greatly increase the calculation of operating time, and decrease the precision.

To solve the problem of too many input variables, this paper proposes the use of BP neural network traffic forecasting model combined with principal component analysis decreases original input variables through principal component analysis, obtains linearly independent new factors which include the information of original input variables. Then it uses these new factors as input variables so as to simply input variables. Finally, the paper uses actual traffic data for traffic forecasting.

2.Traffic forecasting model based on BP Neural Network principal component analysis

2.1 The structure and principle of BP Neural Network

BP Neural Network is a one-way transmission to the multi-forward network, and except for input and output nodes, there are also one or more layers of hidden nodes, the nodes of the same layer is out of couple with each other, the input signal passes from the input layer nodes to the hidden nodes followed by the transfer function, then spread to the output nodes. The output of each node only influenced the next output nodes, as shown in figure 1:

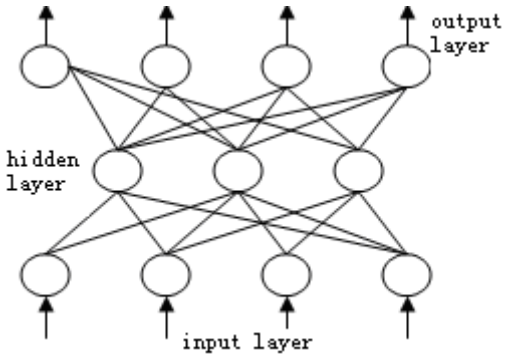


Figure 1 the structure of BP Neural Network

BP Neural networks can be viewed as a highly nonlinear mapping from input to output namely F $R^n : R^m$,f(x)->Y For pattern: input $x_i \in R^n$ output $y_i \in R^m$ $g(x_i) = y_i$ (i= 1, 2,...n) The neural network is approximate to complex function after a number of simple nonlinear functions, and it can obtain the output using input at will.

1.transfer function often is 0 1 S function

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

2.error function

The pth pattern error computing formula

$$E_p = \frac{\sum_i (t_{pi} - O_{pi})^2}{2} \quad 2$$

t_{pi}, O_{pi} are expected output and network's computing output.

Through correcting the weights of network w_{ij}, T_{ij} and threshold θ ,to make error function E descend following the direction of minimum local gradient [3],[5]

BP network nodes include: input nodes x_j, hidden nodes y_i, output nodes O_l. The weight of network between input nodes and hidden nodes is w_{ij}, the weight of network between hidden nodes and output nodes is T_{li}. When the desired output of the output nodes is t_l, the calculation formula of BP model is:

3. the formula of output O_l of output nodes:

the input of input nodes: x_j

output of hidden nodes: $y_i = f\left(\sum_j w_{ij} x_j - \theta_i\right)$ 3

output of output nodes $O_l = f\left(\sum_i T_{ij} y_i - \theta_l\right)$ 4

connecting weight is w_{ij} nodes threshold O_l

4. output layer's correcting formula:

desired output of output nodes: t_l

all the patterns'error: $E = \sum_{k=1}^{P} e_k < \varepsilon$ 5

one pattern's error: $e_k = \sum_{l=1}^{n} | t_l^{(k)} + O_l^{(k)} |$ 6

p is the number of patterns n is the number of output nodes.

error formula: $\delta_l = t_l \quad O_l \quad * O_l * 1 \quad O_l$ 7

correcting weight: $T_{li}^{(k+1)} = T_{li}^{(k)} + \eta \delta_l y_i$ 8

k is the number of number of iterations.

correction of threshold $\theta_l^{(k+1)} = \theta_l^{(k)} + \eta' \delta_l'$ 9

5.the correction formula of hidden node:

error formula $\delta_i' = y_i (1 - y_i) \sum_l \delta_l T_{li}$ 10

the correction formula of weight $w_{ij}^{(k+1)} = w_{ij}^{(k)} + \eta' \delta_i' x_j$ 11

correction of threshold $\theta_i^{(k+1)} = \theta_i^{(k)} + \eta' \delta_i'$ 12

2.2 Traffic accident forecasting model

2.2.1 The major influencing factors of accidents

Traffic accidents happen because of co-ordination of various factors such as cars, roads, climate and environment. In consideration of the many factors that impact on traffic, we selected population, drivers, the population of large vehicles, the population of small cars, the mileage of arterial road, the mileage of minor arterial roads, rain and snow as seven factors.

2.2.2 Traffic accident forecasting model

2.2.2.1 Principal component analysis

Principal component analysis is the use of dimension reduction by constructing the appropriate linear combination of the original index, to produce a series of uncorrelated comprehensive indexes, and to select some of these indexes, which include as much information as old indexes, so as to use these new indexes to reflect individual. Because the method reduces dimension by eliminating the correlation between indexes, it has been bringing in concern in recent years and becoming a unique multi-evaluation of technical indexes.

Based on the analysis of the main influencing factors of accident, the paper selects seven factors of accident are $x_1, x_2, x_3, x_4, x_5, x_6, x_7$, adopts principal component analysis first, and analyses these seven factors, as follows:

Step 1 Normalization of factors

Because every index has different concept of magnitude, before analyzing principal component, we need to normalize data, making values range from 0 to 1.

The method is as follows:

$$x_i^{'} = \frac{x_i}{\max(x_i)} \qquad 13$$

Step 2: Using standardized data to calculate correlation matrix

$$R = (r_{ij})_{7\times 7} \; (r_{ij} = \frac{1}{n}\sum_{l=1}^{n} x_{li}^{'} x_{lj}^{'} \text{ ,i,j=1,2.....7}) \qquad 14$$

Step 3: Calculate eigenvalue and eigenvector of correlation matrix R to get principal component

Make $|R - \lambda I| = 0$ calculate 7 eigenvalue such as λ_i (i=1,2.....7), they are variances of principal components, rank them from small to large: $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_7 \geq 0$,so the expression of principal component is

$$Y_i = l_i^{'} X = l_{i1} X_1 + l_{i2} X_2 + ... + l_{i7} X_7 \qquad 15$$

i=1,2,.....7.

Step 4 Select m principal components to make sure variance contribution rates $a = \sum_{i=1}^{m} \lambda_i / \sum_{i=1}^{7} \lambda_i$ >0.995.

Through step 1 to 4,we can calculate their principal components are $Y_1, Y_2, ...Y_m$ m<7,and Y_i (i=1,2,...,m) is linear combinations with $x_i^{'}$ (i=1,2,...,7) and make $Y_1, Y_2, ...Y_m$ as input of BP Neural Network to get forecasting results by study of BP network algorithm.

2.2.2.2 BP Neural Network simulation using principal components as input factors

The theory has been proved that the three-tier Network system is a better model for nonlinear modeling, every continuous function can be realized through one three-layer neural network. In the neural network forecasting model, we use a three-tier network. That is, one input layer, one hidden layer, one output layer. Because the network's ability to express is increasing with the number of input layer and output layer increasing, and also convergence rate is increasing, so the model is heoretical workable. In the condition of considering factors of traffic accident, the paper uses m principal components as the input layer, the input layer has 7 neurons, the hidden layer has 10 neurons, the output layer has 1 neurons. The output layer is object variable, namely the number of traffic accidents.

Iteration process of BP neural network is as follows:

1. Give initial value for weight coefficient w_{ij} of all layers.

2.Get all nodes' output according to 1 3 4 .

3.Get error e_k according to 6 ,if it meets 5 ,or go to step 4.

4.Get errors(δ_l , $\delta_l^{'}$) of principal components and hidden layers according to the formula (7) and (10), then correct weights according to 8 and 11 ,then go to step 2. (2-4 is the iterative process).

3. Example

Take a city as an example, describe the method of the paper, the data can be seen from table 1:

Table 1 The traffic accident data of a city

year	population ten thousand	drivers (ten thousand	the number of large vehicle	the number of small vehicle	the mileage of arterial road (km	the mileage of minor arterial road km	rain or snow d	The number of traffic accident
1997	87.9	306867	16520	40935	4549	1149	10	58041
1998	89.5	358603	19170	44605	4950	1281	12	59621
1999	90.2	437264	23210	53951	5662	1544	14	62713
2000	91.2	627291	32960	71199	7196	1981	18	67725
2001	92.8	851636	44190	90305	8515	2735	24	68930
2002	93.9	962001	49460	93104	9539	3008	28	69802
2003	96.3	1137364	57620	106006	13388	3934	32	71311
2004	98.6	1280130	63880	121128	15424	4335	34	73521
2005	101.8	1430164	69970	140310	17321	4739	37	76308
2006	104.9	1564299	75280	142527	18980	5167	42	79532

3.1 Data processing

The factors of accidents include population, drivers, the population of large vehicles, the population of small cars, the mileage of arterial road, the mileage of minor arterial roads, rain and snow, namely $x_1, x_2, x_3, x_4, x_5, x_6, x_7$,the change of accidents with years is as figure 2.

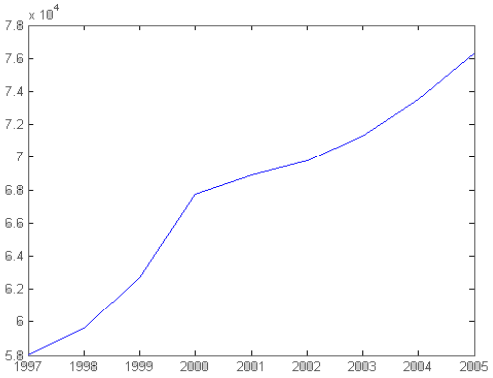


Figure 2 the change of accidents with years

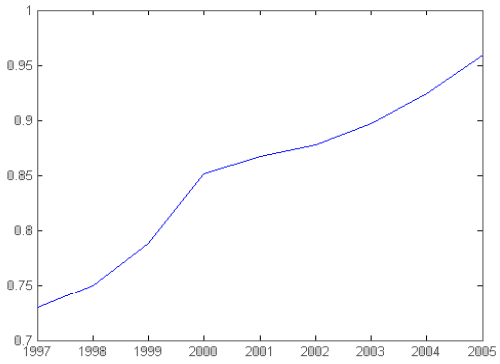


Figure 3 the change of accidents with years after normalization

Because the difference between factors is too large, we normalize them according to the data of 2006, as can been seen from figure 3 and table 2.

Table 2 The traffic accident data after normalization

year	population ten thousand	drivers (ten thousand	the number of large vehicle	the number of small vehicle	the mileage of arterial road (km	the mileage of minor arterial road km	rain or snow d	The number of traffic accident
1997	0.8379	0.1962	0.2194	0.2872	0.2397	0.2224	0.2451	0.7298
1998	0.8532	0.2292	0.2546	0.3130	0.2608	0.2479	0.2880	0.7496
1999	0.8599	0.2795	0.3083	0.3785	0.2983	0.2988	0.3386	0.7885
2000	0.8694	0.4010	0.4378	0.4995	0.3791	0.3834	0.4270	0.8515
2001	0.8847	0.5444	0.5870	0.6336	0.4487	0.5293	0.5684	0.8667
2002	0.8951	0.6150	0.6570	0.6532	0.5026	0.5822	0.6674	0.8777
2003	0.9180	0.7271	0.7654	0.7438	0.7054	0.7614	0.7599	0.8966
2004	0.9399	0.8183	0.8486	0.8499	0.8127	0.8390	0.8143	0.9244
2005	0.9704	0.9143	0.9295	0.9844	0.9126	0.9172	0.8804	0.9595
2006	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

3.2 Principal component analysis

Do principal component analysis, get 3 principal components in the condition of variance contribution rates is over 0.995, as can been seen from table 3:

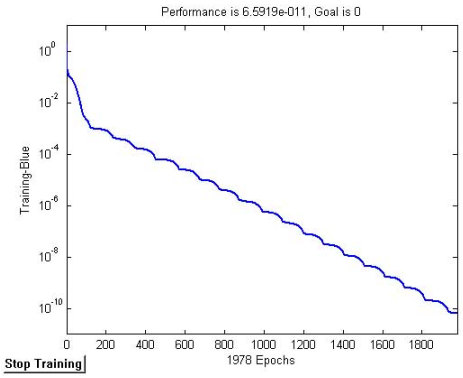
Table 3 The new components

factors	factor 1	factor 2	factor 3
x_1	0.074705	0.015684	-0.06927
x_2	0.50227	-0.47181	0.38855
x_3	0.22619	-0.26258	0.25926
x_4	0.40321	-0.41789	-0.787
x_5	0.69495	0.70893	-0.01341
x_6	0.17561	-0.04852	0.21459
x_7	0.12087	-0.16921	0.33379
contribution rate	99.07%	0.63%	0.27%

3.3 BP Neural Network combining with principal component analysis

Take factor 1 as input and traffic accident afer normalization as output, we set up three-layer neural network model, the input layer has 7 neurons, the hidden layer has 10 neurons, the output layer has 1 neurons.

The process of simulation can been seen as figure 3,in order to make the precision is high enough(<10e-4),the result is 0.9995,namely the number of traffic accident is 79495.If put the principal component of years from 1997 to 2005,the result is [0.7298 0.7496 0.7885 0.8515 0.8667 0.8777 0.8966 0.9244 0.9595].(the comparison between simulation result and actual data can been seen from figure 4)



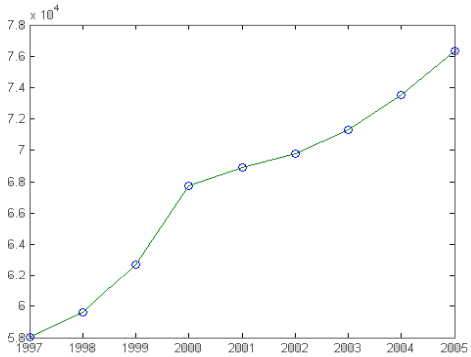


Figure 3 the result of BP Neural NetworkFigure 4 the comparison between combining with principal component analysis simulation result and actual data

3.4 The comparison of results

Compare the result of BP Neural Network combining with principal component analysis, the result of BP Neural Network and the result of nonlinear regression by putting factor 1 as independent variable,we find the result of BP Neural Network combining with principal component analysis is better than the one of BP Neural Network and nonlinear regression.

Table 4 the comparison of results

year	1997	1998	1999	2000	2001
actual data	58041	59621	62713	67725	68930
nonlinear regression	58042	59593	62775	67610	69217
BP Neural Network	57708	59991	62965	67332	68835
BP Neural Network combining with principal component analysis	58042	59617	62711	67721	68930
year	2002	2003	2004	2005	2006
Actual data	69802	71311	73521	76308	79532
nonlinear regression	69519	71428	73464	76319	80415
BP Neural Network	70044	71189	73663	76239	79389
BP Neural Network combining with principal component analysis	69805	71308	73519	76311	79495

The result indicates that BP Neural Network combining with principal component analysis is better than BP Neural Network in the efficiency and precision[4][6]. BP Neural Network combining with principal component analysis is very practicable.

4.Summary

We can conclude from the paper:

The merit of the model:

1.The model of BP Neural Network combining with principal component analysis is feasible, the result is better than BP Neural Network and regression.

2.The model can consider all factors affecting traffic accidents, through principal component analysis, we get less new factor as input, it is better than BP Neural Network in the efficiency and precision.

But these are also some shortages:

1.The theory about the stability of BP Neural Network is faultiness, we can only try to calculate hidden layers and nodes, and the method is limited in the use.

2.The model is the same with forecasting traffic accidents in the stable condition, not considering the change of exterior conditions, such as the change of transport policy and so on,this is also an important factor of the traffic accident.

In a word, The model of BP Neural Network combining with principal component analysis is feasible,considering the complexity and randomicity of urban transportation, the apply range could be developed except accident forecasting, and the model needs to studied and improved.

Reference

Wang Wei.Research on Sustainable Development Planning Theory of Urban Transportation [J].Journal of southeast university,2001,31:(1:6)

Awtani, Modeling of Material Behavior Data in a Functional Forms Uitable for Neural Network Representation[J].Computtional Materials Science,1999,15:493 502

Martin E W,Defferyes D W,Hoffer J A etal. Managing information technology. New York:MacMillan,1991

Su Jinming ZhangLianhua LiuBo. The Apply of Matlab. BeiJing: The China Electronic Industry Publishing House,2000.100-130

H.I.CHOI.F Abrication of High Conductivity Copper Alloys by Rod Milling. Journal of Materials Science Letters,1997,16:1600 1602

I.A.Basheer.Artificial, neural network: Fundamentals, Computing, Design, and Application. Journal of Microbiological Methods,2000,43:3 31