Investment strategy of colleges based on BP neural network and optimization program

Dong Chen[1], Hongwei Pan[2], Yuxia Dai[3,a], Lihong Wang[1,2,b]

[1]Faculty of Science, Ningbo University, Ningbo, 315211, China

[2]Faculty of Mechanical Engineering & Mechanics, Ningbo University, Ningbo, 315211, China

[3]Faculty of Architectural, Civil Engineering and Environment, Ningbo University, Ningbo, 315211, China

[a]email: 986249833@qq.com, [b]email: wanglihong@nbu.edu.cn

Keywords: BP Neural Network, Principal Components Analysis, Cluster Analysis

Abstract: In this paper, a model to determine optimal investment strategy for foundation is introduced. Firstly, we preprocess the college data and classify, filter and fill the missing data with clustering analysis and regression method. Secondly, based on Back Propagation (BP) Neural Networks, a model of ROI was established. We find out the relationship between the variables and the comprehensive index. Thirdly, in order to obtain the maximum of ROI under fixed investment, an Optimization Model of fund allocation is built.

Introduction

In this paper we study on neural network algorithm. Meanwhile, neural network has its distinguished advantage in the process of dealing with complex nonlinear systems. With the technique of storing knowledge in connection weights and mapping all kinds of nonlinearity as well, the whole network is equipped with a powerful ability of model identification and data fitting. This paper applies the algorithm, cluster analysis and principal component analysis (PCA) to build the strategy of investment [1]. The data of the paper is extracted from the U.S. National Center on Education Statistics [2].

Data Preprocessing

All evaluation indexes in this paper are respectively classified into 62 Variable Indexes, 7 Outcome Indexes, and 53 Fixed Indexes.

Due to the fact that Fixed Indexes will not change with the school investment situation, therefore, Fixed Indexes will not be taken into consideration when screening schools. Fixed Indexes will be ignored in the following statement.

The absence of data in various databases often exists，certain values can be used to fill the missing data. And the paper does regression to evaluate factors of the same kind of schools [3]. Clustering analysis method can be used to classify data according to the structure characteristics of the data is suitable for the classification of colleges and universities in this paper.

We define a vector $z_i = (z_{i1},\ z_{i2}, \dots, z_{i62}, z_{i63}, \dots, z_{i69})$ to denote the i^{th} Variable Indexes and Evaluation Indexes.

We assume that the index data of schools obey the Normal Distribution. The differences can be described by Euclidean Distance which can calculate any schools evaluation indexes with data, for example,$(z_{i1},\ z_{i2}, \dots, z_{i69})and(z_{j1},\ z_{j2}, \dots, z_{j69})$.

The formula is as follows:

$$\text{Euclid}(i,j) = \sum_{n=1}^{69} (z_{in} - z_{jn})^2$$

The specific process of Clustering analysis is as follows:

(1) All schools will be classified firstly as one independent class (n classes,$n = 2935$), calculate the distance between each data point according to the Euclidean distance to form a distance matrix $\boldsymbol{D}$;

(2) Merge the nearest two schools into one category, and it comes m $(m < n)$ categories. Calculate the new categories and the distance between each category, forming a new distance matrix $\boldsymbol{D}_1$;

(3) According to the same principle of the second step, we merge the nearest two categories into one. If the number of classes is still greater than 1, repeat it until all data are merged into one.

In the end we divide all schools into seven categories. The classification results show as Tab.1

Tab.1 Seven categories distribution

CLASS	INSTNM	TOTAL
1	Crown College|, Shaw University, Wingate University, etc.	1240
2	Connecticut College, Juniata College, McMurry University, etc.	268
3	Chabot College, Chaffey College, Kilgore College, etc.	1153
4	Kilgore College, Crossroads College, Ecclesia College, etc.	166
5	Sitting Bull College, Dine College, Stone Child College, etc.	37
6	The New School, Villa Maria College, Pratt Institute-Main, etc.	57
7	Sterling College, Kauai Community College, Kapiolani, etc.	14

Comprehensive Index and Return on Investment

Therefore, Income Indexes chosen include the situations of undergraduates and their job incomes, which effectively reflect the investment return of education [6].

There are seven Outcome Indexes we get above. In order to make it more convenient in subsequent model calculation, we combine the seven Outcome Indexes into one Comprehensive Index. Principal component analysis is to make linear combination of the seven Outcome Indexes to form new comprehensive indexes.

According to the principle that accumulated contribution rate is more than 85%, we pick the first three principal components and multiply their contribution rates to obtain comprehensive indexes.

Obtaining the formula of each principal component:

$$F_1 = 0.508z_{63} + 0.687z_{64} + 0.689z_{65} + 0.809z_{66} + 0.763z_{67} + 0.830z_{68} \ + 0.829z_{69}$$

$$F_2 = -0.189z_{63} + 0.722z_{64} + 0.720z_{65} - 0.267z_{66} - 0.278z_{67} - 0.305Zz_{68} \ - 0.258z_{69}$$

$$F_3 = 0.795z_{63} + 0.010z_{64} + 0.010z_{65} - 0.069z_{66} + 0.172z_{67} - 0.263z_{68} \ - 0.330z_{69}$$

The comprehensive index F is

$$F = 0.54539F_1 + 0.19772F_2 + 0.12068F_3$$

ROI is the ratio of the increment of comprehensive index to time, to some extent, it can reflect state of school development.

We define the ROI as:

$$\text{ROI} = \frac{f_k(t + \Delta t) - f_k(t)}{\Delta t}$$

In the formula:

$f_k(t)$ represents a predicted comprehensive index of k^{th} school in t^{th} year,

$k \in \{0, 1, \dots, 100\}$;

Δt represents a time interval ($\Delta t = 1$ in this paper).

The higher a college's ROI is the greater potential it has. We could screen schools, ascertain the volume and the duration time of investment according to ROI.

School Screen

We fit the current data to create a model for each school's ROI then choose the fittest one to invest on. Apparently, a simple linear relationship can't be found between the fund and ROI. That's to say, more fund can't promise a higher ROI. Therefore, we need to find another simulation to interpret the relationship between those two variables, and we take a neural network into consideration. Because this model includes relationships among all evaluation indexes of schools so as to make an effective simulation of ROI after a possible investment.

A neural network includes a distribution of information storage so that a slight damage of the network will only cause a little influence on operation of artificial neural network. The robustness and fault tolerance it contains can also be helpful to tackle some problems like a large instability of data's influencing factors. BP neural network is a kind of multilayer feed forward neural network that calculates in accordance with the algorithm of back-error propagation [5][6].

Because the original data follows the normal distribution, we simplify the calculation in MATLAB.

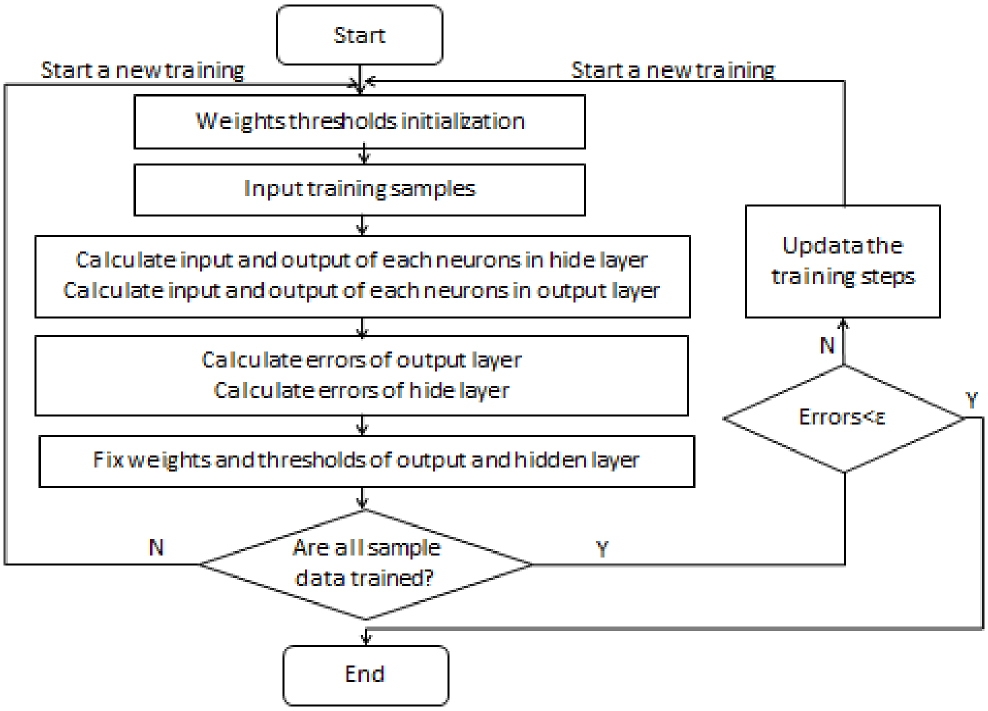
Start
Start a new training
Start a new training
Weights thresholds initialization
Input training samples
Calculate input and output of each neurons in hide layer
Calculate input and output of each neurons in output layer
Calculate errors of output layer
Calculate errors of hide layer
Fix weights and thresholds of output and hidden layer
Updata the training steps
N
Errors<ε
Y
N
Are all sample data trained?
Y
End

Fig.1 BP neural network algorithm flowchart

BP neural network algorithm flowchart is shown in Fig.1， follows are descriptions:

1) Initialize BP neural network model. Connection weights $w_{ij}, v_{ji}$and thresholds θ_j, γ_i of the neural network are given numerical interval [0, 1].

2) Calculate input u_j and output h_j of neural node in the hidden layer of BP neural network.

$$u_j = \sum_{i=1}^{n} w_{ij} x_{ij} - \theta_j$$

$$h_j = f(u_j) = \frac{1}{1 + \exp(-u_j)}$$

3) Calculate input $l_i$and output y_i of neural node in the output layer of BP neural network.

$$l_i = \sum v_{ji} h_j - \gamma_i$$

$$y_i = \frac{1}{1 + \exp(-l_i)}$$

4) Calculate weights error ε_i of neural node in output layer connected to BP neural network.

$$\varepsilon_i = (c_i - y_i)y_i(1 - y_i)$$

5) Calculate weights error ε_j of neural node in hidden layer connected to BP neural network.

$$\varepsilon_j = \sum_{i=1}^{q} \varepsilon_i \, v_{ji} h_j(1 - h_j)$$

6) Update weights v_{ji} and thresholds γ_i of BP neural network.

$$w_{ji}(N+1) = w_{ji}(N) + \partial \varepsilon_j x_i$$

$$\theta_j(N+1) = \theta_j(N) + \beta \varepsilon_j$$

7) Calculate error of output value and the expectations. If they meet the setting accuracy, the neural network training learning is over. If not, then turn to step two to continue training study [7] [8].

After training neural network, stable relationship of ***X*** and ***F*** can be obtained, namely, the appropriate network structure being established, such as shown in Fig.2.

Train model is selected for network training, and the learning algorithm is Levenberg-Marquadt method.

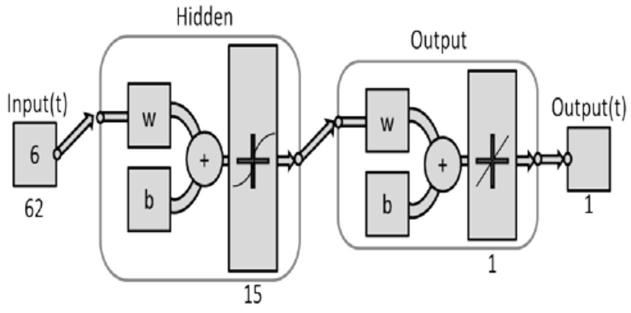

Fig.2 Neural Network

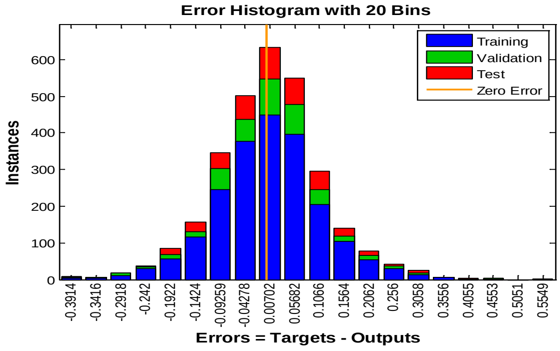

Fig.3 Error histogram

Fig.3 shows the distribution of the errors. The yellow vertical bar in the middle represents zero error. As you can see, error is roughly evenly distributed on both sides of the zero, which presents the law of normal distribution. Training results are great. And the network structure is relatively perfect [9].

When a school receives investment funds, variable indexes ***X*** change, which will stimulate the ***F*** value corresponding to delta ***F*** changes. In turn, ROI can be gotten to simulate the school development after the receipt of funds.

If the investment funds are too scattered, it will reduce investment effect. We selected 100 schools which have the highest comprehensive indexes. ROI of some schools sorted in Tab.2.

Tab.2 ROI of sorted schools(part)

UNITID	INSTNM	ROI
447582	New River Community and Technical College	11.620%
217989	Denmark Technical College	10.698%
383190	Hawaii Community College	9.882%
144157	City Colleges of Chicago-Kennedy-King College	9.133%

The Optimal Distribution of Funds

The allocation of funds should be under the most rational and the most optimal condition, which means that the sum of all colleges and universities' comprehensive indexes should be the maximum after capital is invested.

We assume that the amount of the fund is 100 million. It is divided into 100 portions. Let every single portion expresses as x_k. The 100 portions are distributed to 100 colleges and universities respectively.

$$
\begin{aligned}
&\text{Max} && f\,(\boldsymbol{N}) \\
&\text{S.T.} && \sum_{k=1}^{100} x_k = 10^8 \\
& && 0 \le x_k \le 10^7 \\
& && 0 \le x_k(x_k - 10^6) \le 9 \times 10^{13} \\
& && x_k = 10^6 + 10^5 p \qquad p \in \{1,2,3,4,5\}
\end{aligned}
$$

In the formula:

$\boldsymbol{N}$ is a matrix of 100 dimensions. It represents the investment amount every college and university of 100 colleges and universities obtains.

$f(\boldsymbol{N})$ is the relationship between investment amount and return on investment of all colleges and universities which is determined by BP neural network.

x_k is the investment amount that the k^{th} college and university obtains.

Finally, we get the distribution scheme of every college and university. Tab.3 has listed specific distribution scheme of some colleges and universities.

Tab.3 Each colleges and universities' specific distribution scheme (part)

MONEY	1.3million	1.2million	1.1 million	1.0 million	0 million
UNITID	447582	194611	198729	144193	138840
	217989	198640	175573	413617	217837
	383190	226204	175786	117733	372958
	etc.	etc.	etc.	etc.	etc.
Total	12	31	33	12	12

Conclusion

Our models in this paper are based on bulk samples and indexes. The relationship between indexes is complex, so it's difficult to describe it with ordinary linear relationship. However, neural network has a strong ability of non-linear fitting and thus could map any complex non-linear relationship. Besides, the learning rules are easy to be understood and possess great robustness, good memory and strong self-learning ability [10]. Through processing the principal components, we combine some indexes which are strongly relative to ROI into one comprehensive index. It's a method which describes ROI more visually. through which we could get the most effective and reasonable investment strategy more easily.

Acknowledgement

In this paper, the research was sponsored the Science and Technology Projects in Zhejiang Province Department of Education (Project No. Y201430940).

References

[1] X. H. Shi, Y. C. Liang, H. P. Lee, et al. Improved Elman networks and applications for controlling ultrasonic motors [J]. Applied Artificial Intelligence, 2004, 18(7):603-629.

[2] http://nces.ed.gov/ipeds/datacenter/

[3] Lee S J, Lee C H, Hong J S, et al. Bibliography (1) A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society , Series B. 39(1):13, 1977. (2) Ambrose, R. O[J]. Ahc, 2008.

[4] Ling Wu. Research on Gap of Higher Education Returns in the United States in the Age of Knowledge Economy[D]. ShangDong University, 2015.

[5] Sadeghi B H M. A BP-neural network predictor model for plastic injection molding process[J]. Journal of Materials Processing Technology, 2000, 103(3):411-416.

[6] Olson D, Mossman C. Neural network forecasts of Canadian stock returns using accounting ratios[J]. International Journal of Forecasting, 2003, 19(02):453-465.

[7] Wang Y S, Shen G Q and Xing Y F 2014 A sound quality model for objective synthesis evaluation of vehicle interior noise based on artificial neural network Mech. Syst. Signal Process. 45 255.

[8] Wang Y S, Lee C M, Kim D G and Xu Y 2006 Sound-quality prediction for nonstationary vehicle interior noise based on wavelet pre-processing neural network model J. Sound Vib. 299 933.

[9] Gao X D, You D Y and Katayama S 2012 Seam tracking monitoring based on adaptive Kalman filter embedded Elman neural network during high power fiber laser welding IEEE Trans. Ind. Electron. 59 4315–25

[10] Reddy R K and Ganguli R 2003 Structural damage detection in a helicopter rotor blade using radial basis function neural networks Smart Mater. Struct. 12 232-41