The research of BP Neural Network based on One-Hot Encoding and Principle Component Analysis in determining the therapeutic effect of diabetes mellitus

Yuchen Qiao, Xu Yang*, Enhong Wu

School of Automation, Wuhan University of Technology, Wuhan, Hubei, China
*yx_auto@whut.edu.cn

Abstract. The BP neural network is a hybrid algorithm based on One-Hot Encoding and Principle Component Analysis (PCA). In order to make the distance calculation between variables more reasonable, the hybrid algorithm first reduces the dimension of input variables by means of PCA, and then processes the principal component variables with One-Hot Encoding. Thereafter, the BP neural network is established with all the specially processed data. The evaluation system of the diabetes therapeutic effect is established on this hybrid algorithm and the system can reduce the burden of doctors and enhance the treatment efficiency. Compared with the evaluation system established by directly using BP neural network, the BP neural network evaluation system based on One-Hot Encoding and PCA is faster and more accurate in the evaluation of the therapeutic programs of diabetes. It also provides a technological base for the evaluation of therapeutic programs of other diseases.

1. Introduction

It is important to make the best treatment plan for different type of diabetics. How to quickly and accurately determine the effect of different plans on the treatment of diabetes is very important. In view of the current situation of too many patients and too few doctors, it is of great significance to design an intelligent algorithm to quickly and accurately evaluate the treatment plan. Only in this way can the burden of doctors be reduced and the treatment efficiency improved.

One-Hot Encoding uses N-bit state registers to encode N states, each of which has its own register bits, and at any time, only one of them is valid. This method can solve the problem that the input variable in the diabetes treatment plan is discrete drugs data.

Principle Component Analysis can transform multiple input variables into a linear combination of a few unrelated comprehensive variables. This linear combination can eliminate the correlation between the principal components, make the results more reasonable, and achieve the purpose of dimensionality reduction. The variance of multiple variables is explained through the linear combination of a few principal components. We derive a few principal components while retaining the information of the original variables as completely as possible and keeping them unrelated to each other. In this way the data are simplified. Therefore, PCA can reduce the number of neurons in the input layer of BP neural network and simplify the function of neural network structure.

In this paper, BP neural network based on One-Hot Encoding and PCA is used to establish a BP neural network evaluation model to determine the therapeutic effect of diabetics. Compared with the traditional diagnosis and treatment process, this method not only improves the accuracy of the treatment scheme, but also improves the treatment efficiency.
2. Experimental data and methods

2.1 Experimental data
The experimental data of diabetics’ treatment come from 130 hospitals in the United States. In order to make the distance calculation between features more reasonable, the researcher digitizes the data, reduces the dimension by means of PCA, and corrects the encoded variables of principal components with One-Hot Encoding method.

2.2 Analysis method

2.2.1 Principle Component Analysis. Principal Component Analysis (PCA) uses the idea of dimensionality reduction, which is a method to recombine many original indicators with certain correlation into a new set of a few unrelated comprehensive indicators \[5\]. That is:

\[ F_p = a_{i1} * Z_{X1} + a_{i2} * Z_{X2} + \ldots + a_{ip} * Z_{Xp}, \quad i = 1, \ldots, p \quad (1) \]

The \( a_{i1}, \ldots, a_{ip} \) (\( i = 1, \ldots, p \)) is the eigenvectors corresponding to the eigenvalue of \( \Sigma \) -- the covariance matrix of \( X \). \( Z_{X1}, Z_{X2}, \ldots, Z_{Xp} \) is the standardized value of the original variable. In order to simplify the structure of BP neural network, PCA is used to reduce the dimensionality of 32 input variables affecting the treatment results of diabetes. The calculation steps of PCA include: standardize the original data with SPSS software; determine the correlation between indexes and the number of principal components \( m \); and Work out the new Fi expression of the principal component.

2.2.2 One-Hot Encoding. One-Hot Encoding is a process of converting category variables into a form readily available to machine learning algorithms. We reconstruct the encoding dimension of the original network packet in the data set by using the single-heat coding to form two-dimensional data \[1\], so as to make the data more reasonable.

2.2.3 BP neural network. BP Neural Network (Back Propagation Neural Network) is a kind of multi-layer feed-forward Neural Network trained according to error back-propagation algorithm. The advantage of BP neural network lies in its strong nonlinear mapping ability and flexible network structure \[7\]. Since BP neural network is more complex and more flexible than linear statistical model, it is able to model strong nonlinear dependence. The second advantage of BP neural network is that it can learn priori unknown relations directly from training data \[8\]. Therefore, we plan to use BP neural network to establish an experimental analysis of diabetes treatment program evaluation system, in order to achieve the analysis of the existing principal components to obtain the final evaluation results.

3. Results and analysis

3.1 Results of principal component analysis
The original data contained 32 input variables that were likely to influence the outcome of treatment, and we applied MATLAB programming to conduct principal component analysis. The eigenvalues and variance contribution rates of the correlation coefficient matrix \( R \) in the data set are shown in table 1.

| The principal components | The eigenvalue | Variance contribution rate | Cumulative contribution rate |
|--------------------------|---------------|----------------------------|-----------------------------|
| g1                       | 28.2883       | 0.471                      | 0.471                       |
| g2                       | 16.5389       | 0.2753                     | 0.7463                      |
| g3                       | 8.6441        | 0.1439                     | 0.8902                      |
| g4                       | 2.5989        | 0.0433                     | 0.9335                      |
| g5                       | 2.0478        | 0.0341                     | 0.9676                      |

Table 1. The eigenvalues and variance contribution rates of R.
In practical applications, the corresponding components whose cumulative contribution rate is above 85% are generally selected as the retention components [6] so that we can make full use of the original information. As can be seen from table 1, the contribution rate of the first five principal components has reached 96.76%, indicating that the first five principal components basically contain all the information of the features. Therefore, we use the first five principal components and the original data to form a new sample set, and we name the new variables A1, A2, A3, A4 and A5.

3.2 Results of one-hot Encoding

After the principal component analysis of the original input variables, the obtained variables including A1, A2, A3, A4, and A5 were adjusted by adding one-hot coding. Eight of the original data (metformin, repaglinide, glimepiride, glipizide, glyburide, pioglitazone, rosiglitazone, insulin) had four states: DOWN, NO, STEADY, and UP, coded as 0, 1, 2, and 3. In order to better calculate the distance between various variables, we use One-Hot Encoding method to encode them as (00, 01, 10, 11), and divide the 8 features into 16 features (X1, X2, ..., X16), and add these variables into the pivot variables for data correction.

3.3 Design and result analysis of BP neural network evaluation system

(1) Establishment of BP neural network evaluation system based on One-Hot Encoding and Principal Component Analysis

Since the 3-layer BP neural network with a single hidden layer is sufficient to perform any complex function mapping, we adopt the 3-layer BP neural network with a hidden layer. The number of neurons in the input layer is 21, corresponding to A1, ..., A5 and X1, ..., X16 variables; Its neurons in the output layer are 3, which corresponds to the assessment degree of diabetes treatment plan, respectively good, medium and bad. After repeated calculation and verification, the number of neurons in the hidden layer was set to 12. Tansig transfer function is used from input layer to hidden layer. The transfer function from hidden layer to output layer is purelin; the training algorithm adopts the nonlinear damped least square optimization algorithm, and the function is trainlm. The maximum cycle number is 1000. When the RMS error of network training is less than e^{-20}, the training is completed, and the learning rate is e^{-8}.

(2) Training and simulation of BP neural network evaluation system based on One-Hot Encoding and Principal Component Analysis

We randomly divided 101,767 sets of data, including input variables and output variables, of which 100,000 sets of data were used for the training and simulation of the evaluation system, and 20,000 sets of data were used for the verification of the evaluation system. In the training process, the error convergence curve of the BP neural network evaluation system is shown in figure 1. The evaluation system rapidly converges in the training process, which only takes 5 seconds and reaches the requirements at the 56th iteration. After the training of the evaluation system, 20,000 sets of data were used for simulation.
(3) Verification of the BP neural network evaluation system based on One-Hot Encoding and Principal Component Analysis.

We input 20,000 sets of reclassified data and use the established BP neural network evaluation system for simulation, and compare the simulation results with the target output results to verify its ability to evaluate various treatment schemes.

The 20,000 input data were brought into the BP neural network evaluation system, and the correct rate of the final output result was 71.39%. The main reason for misjudgment was that the correlation between input variables was not very high, and the other part was due to the structural weight of the BP neural network evaluation system itself.

The regression curve was obtained by MATLAB programming, as shown in figure 2, and its performance was described by the regression equation. The BP neural network constructed by us has better performance.

(4) Performance comparison between the BP neural network evaluation system and the BP neural network evaluation system based on One-Hot Encoding and PCA.

We compare the BP neural network evaluation system based on One-Hot Encoding and PCA with the evaluation system established by directly using BP neural network, both of them using the same settings, so as to better prove the advantages of the BP neural network evaluation system based on one-hot Encoding and Principal Component Analysis.

1) After we use One-hot Encoding and PCA algorithm, there are only 21 input variables, while the original data has 32 variables that may affect the final result, so the number of network input layer nodes is 32, and the number of output layer nodes is still 3. Finally, the number of neurons in the hidden layer was set as 18.

2) After training and simulation with the same data, it can be seen from figure 3 that the BP neural network evaluation system meets the requirements at the 27th iteration, but its running time is 19 seconds, so the training time of the BP neural network evaluation system based on One-Hot Encoding and PCA is shorter. Finally, the verification data is brought into the BP neural network evaluation system, and the regression performance curve obtained is shown in figure 4. The regression equation is: Out = 0.024T + 0.28, and the correlation coefficient R is 0.15403. The comparison of the two systems shows that the BP neural network evaluation system based on One-Hot Encoding and PCA is more close to the reality.
Figure 3. The convergence curve of BP neural network before improvement.

Figure 4. The regression curve of simulation data in evaluation system before improvement.

It can be seen from the comparison that the BP neural network evaluation system based on One-Hot Encoding and PCA has better performance in determining the therapeutic effect of diabetes.

4. Conclusion

This paper discusses the application of BP neural network based on One-Hot Encoding and Principal Component Analysis in determining the therapeutic effect of diabetes, and draws the following conclusions.

(1) We used MATLAB programming to conduct principal component analysis on 32 input variables in the original data. Since the cumulative contribution rate of the first five principal components has reached 96.76%, we constructed five new input variables to reduce the dimension of the original data.

(2) On the basis of principal component analysis of the original data, we conducted One-Hot Encoding for the five principal components, and used One-Hot Encoding for these discrete features, adding X1, X2 and... X16 to correct the input variables to make the calculation of the distance between these characteristic variables more reasonable. This has played a great role in improving the accuracy of the evaluation system.

(3) After processing the original data, we used BP neural network to construct a diabetes treatment effect evaluation system. Compared with the BP neural network evaluation system established by directly extracting data, it was found that the operation time of the improved evaluation system is reduced to 5 seconds, and the number of neurons in the hidden layer is 12, which is less than 18 in the original evaluation system. Therefore, the network structure is simplified. After the simulation of 20,000 sets of data, it was found that the regression equation of the improved evaluation system was closer to reality. For 20,000 sets of data, the accuracy was increased from 69.58% to 71.39%.

(4) Therefore, the BP neural network evaluation system is faster and more accurate in evaluating the treatment programs of diabetes. It also provides a technical basis for the evaluation of the treatment programs of other diseases.

Acknowledgements

This paper is supported by National Students’ Platform for Innovation and Entrepreneurship Training Program in 2018(ID:201810497222)

References

[1] Liang Jie, Chen Jiahao, Zhang Xueqin, Zhou Yue and Lin JiaJun. 2018 J. Anomaly detection based on singular heat coding and convolutional neural network. Journal of tsinghua university (natural science edition). 25 1–7

[2] Zhang Luyao, Ma Zulu, Qin xingming. 2018 J. Application of principal component analysis and
factor analysis in the study of hydrochemical characteristics of karst water. *Journal of guangxi academy of science.* **34**(4) 1–7

[3] Li Junmei, Hu Yihua and Tao Xiaohong. 2005 *J.* Identification method based on principal component analysis and BP neural network. *Infrared and laser engineering.* **06** 719–723

[4] He Yong, Li Xiaoli and Shao Yongni. 2006 *J.* Identification of apple varieties by near infrared spectroscopy based on principal component analysis and neural network. *Spectroscopy and spectral analysis.* **05** 850–853

[5] Zhang Shaoze, Zhang Chunying, Meng Qingyang. 2013 *J.* Comprehensive evaluation of regional technological innovation capacity in China based on principal component analysis. *Economic journal.* **Z2** 90–91

[6] Tian Yajuan, Ma Wei, Liu Ye, Xiao Zhi-hong and Cheng Guojian. 2014 *J.* Prediction of water flooding effect in oil Wells based on pca-fnn. *Journal of xi’an petroleum university (natural science edition).* **29**(03) 83–86+10–11

[7] Gu Hengming and Hu Liangping. 2008 *J.* Multiple linear regression analysis based on machine learning statistics. *Sichuan mental health.* **31**(01) 15–18

[8] Kukkonen J, et al. 2003 *J.* Extensive evaluation of neural network models for the prediction of NO2 and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki. *Atmos. Environ.* **37** 4539–4550