Research Article

Research on Partial Least Squares Method Based on Deep Confidence Network in Traditional Chinese Medicine

Wang-ping Xiong,¹ Tian-ci Li,¹ Qing-xia Zeng,¹ Jian-qiang Du,¹ Bin Nie,¹ Chih-Cheng Chen ,² and Xian Zhou¹

¹School of Computer, Jiang Xi University of Traditional Chinese Medicine, Nanchang, Jiangxi, China

²School of Information and Engineering College, Jimei University, Xiamen, Fujian 361021, China

Correspondence should be addressed to Chih-Cheng Chen; 3343033397@qq.com

Received 21 February 2020; Accepted 4 May 2020; Published 6 June 2020

Guest Editor: Fuqiang Gu

Copyright © 2020 Wang-ping Xiong et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Partial least squares method has many advantages in multivariate linear regression modeling, but its internal cross-checking method will lead to a sharp reduction of the principal component, thereby reducing the accuracy of the regression equation, and the selection of principal components about the traditional Chinese medicine data is particularly sensitive. This paper proposes a kind of partial least squares method based on deep belief nets. This method mainly uses the deep learning model to extract the upper-level features of the original data, putting the extracted features into the partial least squares model for multiple linear regression and evading the problem that selects the number of principal components, continuously adjusting the model parameters until satisfied well-pleased accuracy condition. Using Dachengqitang experimental data and data sets in the UCI Machine Learning Repository, the experimental results show that the partial least squares analysis method based on deep belief nets has good adaptability to TCM data.

1. Introduction

The mechanism of treating diseases with traditional Chinese medicine is that multiple components of TCM correspond to multiple targets of diseases, so there are many related variables in TCM data. In the dose-effect relationship analysis of TCM prescriptions, the content of TCM prescriptions should be used to analyze the corresponding curative effects. To analyze the effects of independent variables on dependent variables, how to accurately select the number of independent variables is the key to establish a good dose-effect relationship model to achieve the goal of the better dose-effect analysis.

Partial least squares (PLS) [1, 2] is a multivariate statistical data analysis method, which integrates the basic functions of principal component analysis (PCA), canonical correlation analysis (CCA), and multiple linear regression (MLR) analysis. PLS simplifies the data structure by principal component extraction, selects the first several principal

components by truncation, and only uses these principle components to obtain a model with better prediction performance. If the subsequent principal components can no longer provide more meaningful information for the interpretation of the model, adopting too many principal components will instead reduce the precision of regression fitting. Therefore, the selection of the number of principal components is particularly important. However, there are many related variables in TCM data. The cross-validation method will drastically reduce the number of principal components and reduce the accuracy of the regression equation. Deep belief nets (DBN) [3] are a multilayer neural network that fit the input data as much as possible. They extract the upper-level features of the original data through deep learning of the existing model, avoid the problem of selecting the number of principal components, convert the original m -dimensional input data into n -dimensional output data through the deep network, and the output data are another deep nonlinear expression of the original input

data. In 2011, Zhou [4] proposed embedding the fuzzy neural network model into the iterative form of partial least squares, which avoids the problem of principal component selection while achieving the nonlinear mapping effect, but the model results are easily affected by membership functions. In 2013, Qin [5] proposed a kernel partial least squares method, which combines intrinsic dimension estimation with the kernel partial least squares method and maps data to the high-dimensional linear space by the kernel function. Although the algorithm can avoid the selection of principal components and the nonlinear structure contained in the reaction sample data, the selection of kernel function is extremely difficult. In 2017, Zhu and JianqiangDu [6] proposed a partial least squares method integrated with the restricted Boltzmann machine (RBM-PLS), which extracts the required upper-layer features through the neural network structure, avoiding the problem of principal component selection, but the model results are affected by the initial values.

Therefore, this paper integrates the deep belief nets into the partial least squares, extracts the upper-level features of the original data through the deep learning model, avoids the problem of selecting the number of principal components, thus carries out multiple linear regression, and finally restores to the regression equation of the original variables. Through the verification of the test data, the parameters of the deep learning model and the number of hidden layers are adjusted in turn, finally achieving the purpose of improving the accuracy of the regression equation. The algorithm not only retains PLS's characteristic of eliminating multiple linear correlations of traditional Chinese medicine data but also can establish a good regression model for data with a small sample size. While avoiding the problem of selecting the number of principal components, it makes up for the defect that traditional Chinese medicine data are particularly sensitive to the selection of principal components, thus establishing a regression model suitable for traditional Chinese medicine characteristics.

2. Partial Least Squares

Partial least squares [7] is a new multivariate statistical data analysis method. Different from traditional multivariate regression statistics, it mainly studies the regression modeling of multiple dependent variables to multiple independent variables [8]. Regression modeling can be carried out when the number of sample points is less than the number of variables or multiple correlations exist between variables, and the regression coefficient of each variable is easy to explain.

The basic idea of partial least squares regression is that there is a set of independent variables $X = (x_1, x_2, x_3, \dots, x_n)$ and dependent variables $Y = (y_1, y_2, y_3, \dots, y_n)$, and the principal components (t_1 and u_1) extracted from independent variables and dependent variables are required to carry the variation information of their respective data tables to the maximum extent- $\text{var}(t_1) \rightarrow \max$, $\text{var}(u_1) \rightarrow \max$, they are required to have the greatest correlation degree $r(t_1, u_1) \rightarrow \max$, multiple linear

regression is then performed, and if the regression equation reaches satisfactory accuracy, the algorithm is terminated; otherwise, the second round of principal component extraction is continued from the residual information, thus reciprocating until satisfactory accuracy is reached.

3. Deep Belief Nets

Hinton and Salakhutdinov [9] proposed that restricted Boltzmann machines (RBMs) can be stacked greedily for training in 2016, thus forming deep belief nets (DBN). DBN are a probability generation model, which make the whole neural network generate training data according to the maximum probability through the weights between neurons in the training network structure, thus forming high-level abstract features [10].

DBN consist of two parts of neurons, namely, visible layer neurons and hidden layer neurons. The key component is RBM; feature extraction of input data can be realized by combining multiple layers of RBM. RBM is composed of two layers of neurons, visible layer and hidden layer, respectively. The specific structure diagram of RBM is shown in Figure 1.

Assuming that the RBM is n -layer, a DBN model that can extract features abstractly is constructed. v and h represent visible layer neurons and hidden layer neurons, respectively. The specific construction process is as follows: firstly, the weight and offset of the first trained RBM are fixed, the state of its hidden layer is taken as the input of the second RBM, the second RBM is trained and stacked on the first RBM, and so on to repeat the above process. The final model is shown in Figure 2.

Since DBN are formed by stacking multiple RBM, the main process of training DBN is the process of training RBM.

The number of neurons in the visible layer is defined as m , the number of neurons in the hidden layer is n , the state vector of the visible layer neurons is V , and the state vector of the hidden layer neurons is H . For a given set of state vectors, the energy of an RBM can be expressed by the following function [11]:

$$E(v, h | \theta) = - \sum_{i=1}^n a_i v_i - \sum_{j=1}^m b_j h_j - \sum_{i=1}^n \sum_{j=1}^m b_j w_{ij} v_i h_j. \quad (1)$$

Among them, the parameter of the model is $\theta = \{a_i, b_i, W_{ij}\}$, a_i is the offset of the visible layer neurons i , b_i is the offset of the hidden layer neurons j , and W_{ij} is the weight between the visible layer neurons and the hidden layer neurons j .

In the process of RBM training, the values of the parameter θ are obtained through continuous iteration (the maximum parameter is obtained by using the stochastic gradient descent method), and the end of the iteration is determined by the effect of the fitted training data.

For trained DBN, under the condition of known parameter $\theta = \{a_i, b_i, W_{ij}\}$, preprocess the original data to obtain $V = (v_1, v_2, \dots, v_n)$ and then put the data into the model, namely,

FIGURE 1: RBM structure diagram.

$$H = \text{sigmoid}(v \times w + a). \quad (2)$$

The original data are mapped into hidden layer data $H = (h_1, h_2, \dots, h_m)$ through sigmoid function. In this way, the original data can be transformed into another deep expression, which achieves the purpose of extracting upper-level features.

4. Partial Least Squares Construction Based on DBN

In partial least squares, principal component analysis (PCA) and canonical correlation analysis (CCA) are mainly used for principal component extraction. It is not necessary to construct all the components for regression modeling. Usually, we use the cross-validity method to determine the number of principal components. When $Q_h^2 \geq 0.0975$, the added component is effective, and the regression model can be improved; that is, when $\sqrt{\text{PRESS}_h} \leq 0.95\sqrt{\text{SS}_{h-1}}$, the added component is regarded as beneficial and otherwise invalid.

$$Q_h^2 = 1 - \frac{\sum_{j=1}^q \text{PRESS}_{h_j}}{\sum_{j=1}^q \text{SS}_{(h-1)_j}} = 1 - \frac{\text{PRESS}_h}{\text{SS}_{h-1}} \geq 0.0975, \quad (3)$$

where h represents the number of components, PRESS_h represents the sum of squared prediction errors of equation Y , and SS_h represents the sum of squared errors of Y fitted with h components. Obviously, the value of PRESS_h is greater than SS_h , but less than SS_{h-1} . In fact, the square roots of the three are not much different, making the condition of $\sqrt{\text{PRESS}_h} \leq 0.95\sqrt{\text{SS}_{h-1}}$ very easy to meet. This leads to a sharp decrease in the number of selected components, thereby reducing the accuracy of the regression equation.

The feature extraction of the DBN network is mainly divided into two aspects: pretraining and parameter tuning. First, pretrain the network. RBM is the basic structure of the DBN network, so training is performed layer by layer in RBM. The standardized data set is used as the input of the underlying RBM, and the first layer feature h_1 of the data set is obtained through the previously trained parameter θ . Taking feature h_1 as the input data of the next layer of RBM, we can get feature h_2 which is more abstract than feature h_1 . Through iterative abstraction, the final feature h_n is obtained. At this time, h_n integrates the different features abstracted by the underlying RBM, making the features more representative of the variables in the original data.

At this time, no matter whether there is a strong correlation between the original variables, different features can be extracted through the DBN network. Furthermore,

according to the different functions of variables, they can be combined into more abstract and upper-level features that can represent different types of variables. In the end, the contribution of each independent variable to the experiment of the traditional Chinese medicine is brought into full play. Therefore, the DBN-PLS method proposed in this paper is to use DBN to replace the main component extraction part of PLS. The upper-level features of the original data extracted by DBN are used as the principal components, then the PLS external model is used for regression modeling, and finally the original variable regression equation is restored. The verification of the test data, in turn, adjusts the parameters of the deep learning model and the number of hidden layers. The goal of improving the accuracy of the regression equation is finally achieved.

Let the independent variable set $X = (x_1, x_2, x_3, \dots, x_p)$ and dependent variable set $Y = (y_1, y_2, y_3, \dots, y_q)$; X are matrices of $n \times p$, and Y are matrices of $n \times q$.

The specific construction process is as follows:

- (1) Data preprocessing: standardized pretreatment of X and Y , respectively, to obtain E_0 and F_0 .
- (2) Independent $E_0 = (E_1, E_2, E_3, \dots, E_p)$ and dependent $F_0 = (F_1, F_2, F_3, \dots, F_q)$ variables are, respectively, put into the DBN model for training, and the processing steps are as follows:
 - (1) According to the number of characteristic attributes of the independent variable $E_0 = (E_1, E_2, E_3, \dots, E_p)$, the number of visible layer neurons in the DBN is determined, p . Since the main purpose is to reduce the dimension of characteristics, the number of hidden layers is generally smaller than the number of neurons in the visual layer. Here, the number is p_1 ($p_1 < p$).
 - (2) Random initialization parameter $\theta = \{a_i, b_j, w_{ij}\}$. The offset vectors of the visible layer and the hidden layer are $a = \{a_1, a_2, \dots, a_p\}$ and $b = \{b_1, b_2, \dots, b_{p_1}\}$, respectively, and the weight matrix is $W = \{W_{ij} | 0 \leq i \leq p, 0 \leq j \leq p_1\}$.
 - (3) Take independent variables $E = \{E_1, E_2, \dots, E_p\}$ as the input of the visual layer, train each RBM layer by layer, the hidden layer of the previous RBM is the visual layer of the next RBM and the output of the prethe input of the next RBM. In the training process, it is necessary to fully train the RBM of the previous layer before training the RBM of the current layer until the last layer.
 - (3) Take the hidden layer of the last layer of DBN $T = \{t_1, t_2, \dots, t_k\}$ and $U = \{u_1, u_2, \dots, u_l\}$ (k and l are the number of principal components extracted for independent variables and dependent variables, respectively) which are taken as principal components in PLS.
 - (4) After calculating the principal components in step 3, we put the principal components into the partial least squares external model for multiple linear regression analysis. Finally, the coefficient is reduced to a multiple linear regression equation of Y with respect to X . According to two evaluation indexes of

FIGURE 2: DBN structure diagram.

root mean square error (RMSE) and redetermination coefficient (R^2), it is judged whether the model is satisfied with the precision requirement at this time. If the requirement is met, the algorithm is stopped. If not, the superparameters of the model are continuously adjusted until the requirement is met.

As shown in Algorithm 1, the DBN-PLS model constructed in this paper is mainly composed of two parts: feature extraction and nonlinear PLS regression. Among them, DBN use the characteristics of their interlayer transmission to automatically extract features from the original data, and PLS is responsible for the regression of the extracted features. The model retains the characteristic that PLS can solve the problem of multiple correlations of traditional Chinese medicine data, while also avoiding the defect of cross-validity of PLS. It first performs canonical correlation analysis and multiple linear regression and then restores the regression equation of the original variables. Through the verification with the test data, the deep learning layer and parameter settings are adjusted in turn. Finally, the purpose of improving the accuracy of the final regression equation is achieved. The overall structural model is shown in Figure 3.

5. Experiment Results and Analysis

The experimental data in this paper are mainly from Dachengqitang experimental data (DCQT) of the Key Laboratory of Jiangxi University of Traditional Chinese Medicine and Housing, AirQuality, and CBM [12] on UCI data sets.

5.1. Experimental Data Description. The experimental data of Dachengqitang consist of 9 samples, which are the effects of active ingredients on pharmacological indexes in rat plasma at 9 different doses. Independent variables are the

main active components in rat plasma, namely, emodin, rhein, chrysophanol, aloë-emodin, emodin methyl ether, magnolol, honokiol, hesperidin, and hesperetin; the dependent variables are the pharmacological indexes examined, namely, first defecation time, motilin, and vasoactive intestinal peptide. Some experimental data are shown in Table 1.

For UCI data, Housing, AirQuality, and CBM data sets are selected, with sample sizes of 506, 9,357, and 11,934, respectively. See <http://archive.ics.uci.edu/ml/> for a detailed description.

5.2. Data Standardization. Data normalization (normalization) processing is a necessary task when to start data mining. Different data have different dimensions and dimension units, and this situation will affect the results of data analysis. To eliminate the dimensional impact between data and the data itself, data standardization is needed to resolve the comparability of data metrics. In order to comprehensively compare and evaluate, the above statistical data are separately standardized. After the original data are standardized, the indicators reach the same level, which is suitable for comparative evaluation [4]. The standardized method used in this paper is Z-score standardization, also called standard deviation standardization. The processed data conform to the standard normal distribution, that is, the mean is 0, and the variance is 1. Its conversion function is

$$x^* = \frac{x - \mu}{\sigma}, \quad (4)$$

where $\mu = (1/n) \sum_{i=1}^n x_i$ represents the mean of each feature of data X and $\sigma = \sqrt{(1/n-1) \sum_{i=1}^n (x_i - \mu)^2}$ represents the standard deviation of each feature of data X .

Taking Dachengqitang data as an example, the data set is subjected to data standardization processing. The results are shown in Table 2.

Input: original sample data set Dataset (D);

Output: DBN-PLS equation.

Step 1 preprocesses the data to obtain (E_0, F_0)

Step 2 deep belief nets (DBN)

Initialize model parameters $\theta = \{a_i, b_i, W_{ij}\}$

Layersize = 1

hidden_layers_sizes = [4, 4, 4]

While Whether the number of RBM layer size reaches the precision condition

while Whether the number layersize of each layer of neurons reaches the accuracy condition

for z in layersize

Calculating the probability that hidden layer neurons are activated $P_z(h_i|v_i)$

Take Gibbs sampling to extract a sample: $v_{i+1} \sim P_z(v_{i+1}|h_i)$

Reconstruct the visual layer with h_i is used to calculate the probability of activated hidden layer neurons

$P_z(v_{i+1}|h_i)$

Take Gibbs sampling to extract a sample: $v_{i+1} \sim P_z(v_{i+1}|h_i)$

V_{i+1} is used to calculate the probability of activated hidden layer neurons $P_z(h_{i+1}|v_{i+1})$

Update weight

$b_i \leftarrow b_i + \lambda(v_i - v_{i+1})$

Step 3 The characteristics extracted by independent variables and dependent variables are calculated by the DBN model, respectively.

$T = \text{sigmoid}(E_o \times W^t + a^t)$, $U = \text{sigmoid}(F_o \times W^u + a^u)$, put the PLS outer model into multiple linear regression, and find the DBN-PLS equation.

Denormalized reduction of coefficients Y multiple linear regression equations for X

Step 4 End

ALGORITHM 1: Partial least squares algorithm based on DBN.

FIGURE 3: DBN-PLS model framework.

TABLE 1: Experimental data set of Dachengqitang (DCQT).

Emodin	Rhein	Chrysophanol	Aloe-emodin	Emodin methyl ether	Magnolol	Honokiol	Hesperidin	Hesperetin	First defecation time	Motilin	Vasoactive intestinal peptide
1.19	31.35	2.82	2	0.67	0.5	0.18	4.91	0.99	6.91	0.15	3.46
2.6	54	3.95	1.89	0.68	0.26	0.13	4.61	1.17	6.55	0.18	3.27
2.82	26.55	4.85	1.49	0.76	0.15	0.28	6.91	2.49	5.78	0.19	3.88
3.08	131.55	16.5	2.3	2.07	0.72	0.79	7.91	0.3	6	0.19	3.88
4.08	66.75	26	1.77	1.5	0.69	0.44	8.15	2.33	5.88	0.18	3.61
...	...	...	...	...	...	...	...	...	...	...	...

TABLE 2: Standardization processing results (DCQT).

Emodin	Rhein	Chrysophanol	Aloe-emodin	Emodin methyl ether	Magnolol	Honokiol	Hesperidin	Hesperetin	First defecation time	Motilin	Vasoactive intestinal peptide
-0.689	-0.741	-1.071	-0.479	-1.027	-0.603	-0.822	-1.153	-0.638	0.224	-1.492	-1.021
-0.563	-0.599	-1.014	-0.506	-1.022	-1.114	-0.876	-1.218	-0.550	-0.139	0.426	-1.335
-0.544	-0.771	-0.969	-0.606	-0.974	-1.348	-0.715	-0.721	0.093	-0.914	1.066	-0.327
-0.521	-0.115	-0.383	-0.403	-0.200	-0.135	-0.167	-0.504	-0.975	-0.693	1.066	-0.327
-0.432	-0.520	0.095	-0.536	-0.537	-0.199	-0.543	-0.452	0.015	-0.814	0.426	-0.773
...	...	...	...	...	...	...	...	...	...	...	...

5.3. *Auxiliary Analysis.* The T^2 ellipse-assisted analysis technique finds the singular points in the sample by analyzing the contribution rate of the sample points to the components. We define the contribution rate T_{hi}^2 of the i -th sample point to the h -th component t_h to be $T_{hi}^2 = t_{hi}^2 / (n-1)s_h^2$, where t_{hi} is the projection value of the i -th sample point on the h -th main axis a_h and s_h^2 is the variance of the component t_h . Then, the cumulative contribution rate T_i^2 of the i -th sample point pair component $t_1, t_2, \dots, t_m$ is $T_i^2 = \sum_{h=1}^m t_{hi}^2 / (n-1)s_h^2$. When T_i^2 satisfies $T_i^2 \geq (m(n^2-1)/n^2(n-m))F_{0.05}(m, n-m)$, it is considered that, at the 95% test level, the cumulative contribution rate of the i -th sample point to the component $t_1, t_2, \dots, t_m$ is too large, and the sample point i is called a singular point.

In particular, when $m=2$,

$$T_i^2 = \left(\frac{t_{1i}^2}{(n-1)s_1^2} + \frac{t_{2i}^2}{(n-1)s_2^2} \right) \geq \frac{2(n^2-1)}{n^2(n-2)} F_{0.05}(2, n-2)$$

$$\frac{t_{1i}^2}{s_1^2} + \frac{t_{2i}^2}{s_2^2} \geq \frac{2(n-1)(n^2-1)}{n^2(n-2)} F_{0.05}(2, n-2).$$
(5)

Let

$$c = \frac{2(n-1)(n^2-1)}{n^2(n-2)} F_{0.05}(2, n-2),$$
(6)

that is,

$$\frac{t_{1i}^2}{s_1^2} + \frac{t_{2i}^2}{s_2^2} \geq c.$$
(7)

Equation (7) represents an area outside the ellipse $(t_{1i}^2/s_1^2) + (t_{2i}^2/s_2^2) = c$ range. In other words, the singular point is distributed outside the ellipse range; then, on the t_1/t_2 -plane coordinate system, the T^2 ellipse and the scatter

plot of the data set can be drawn. Those points that are outside the T^2 ellipse can be called singular points.

From the analysis of the four sets of data, from equation (7), when the extracted component m is 2, the right side of the inequality is an ellipse, and we can make an ellipse on the t_1/t_2 planar graph. Thus, the existence of singular points can be visually reflected in the graph. Therefore, in this paper, two principal components are extracted separately for the four models, and the singular points in the sample points in the model are discriminated according to discriminant conditional formula (7). Four T^2 ellipses were created using SIMCA-P software to show the presence of sample points, as shown in Figures 4–7.

From Figures 4–7, it can be concluded that there are no singular points in the DCQT data set, and there are 22, 544, and 645 singular points in the Housing, AirQuality, and CBM data sets, accounting for 4.3%, 5.8%, and 5.4% of the total samples, respectively. The sample points concentrated outside the ellipse take values away from the average level of the sample points, the so-called singular points. Find out these singular points and eliminate them, and then use the processed data to optimize the experimental analysis.

There are also some common methods in the singularity elimination phase of partial least squares. The direct decomposition algorithm (TDDA) [13] directly decomposes the multidimensional data set according to all its dimensions to obtain several one-dimensional data sets. Therefore, the multidimensional data are mapped to the one-dimensional space, and the purpose of dimensionality reduction is realized. The multidimensional scaling method (TMSM) [13] is a graphical method for visually representing a research object in a low-dimensional space based on the similarity between the research objects and performing clustering or dimensional inclusion analysis. The T^2 ellipsoid method [14] is a modification of the T^2 ellipse analysis technique, extending the two-dimensional space T^2 elliptical plan to the three-dimensional T^2 ellipsoid and then using the T^2 ellipsoid to

FIGURE 4: Singular point identification of DCQT.

FIGURE 7: Singular point identification of CBM.

FIGURE 5: Singular point identification of Housing.

FIGURE 6: Singular point identification of AirQuality.

identify specific points. Taking the Housing data set as an example, the above method is used to identify the specific point of the data set. Figure 8 shows the graphical results of the T^2 ellipsoid method. The results of all methods are shown in Table 3.

Obviously, by analyzing the comparison results, it can be found that the Housing data set after T^2 ellipse cleaning has the highest fitting value of PLS. TDDA may cause omissions when detecting the specific points of multidimensional data, so some specific points will be ignored; when using TMSM, the effect of each dimension on the Euclidean distance is equal, but in practice, the fluctuation range of each dimension is not the same, and the TCM data are highly correlated, so it is not too suitable for this; the T^2 ellipsoid

FIGURE 8: Results of the T^2 ellipsoid method for processing the Housing data set.

method detects some normal data as a specific point, which leads to a decrease of redetermination coefficient (R^2). From there, it can be judged that the T^2 ellipse analysis method has a better performance in the recognition of specific points.

5.4. Experimental Process and Result Analysis. In order to verify the feasibility and effectiveness of the DBN-PLS method, three deep belief net models are, respectively, set to train the number of hidden layers. The number of hidden layers in the first DBN model is set to 3. The number of hidden layers in the second DBN model is set to 4. The number of hidden layers in the third DBN model is set to 5. The number of neurons in the hidden layer of the above three DBN models depends on the specific situation. Each DBN model is combined with PLS, and the above three situations are analyzed and compared with the original partial least squares (PLS) and the partial least squares (RBM-PLS) integrated into the restricted Boltzmann machine in the literature [12]. Four groups of experimental data are used for comparison and verification. The specific description of the data set is shown in Table 4.

In the specific process of the experiment, the model is optimized by adjusting the model parameters, and the effects

TABLE 3: Results of the specific point identification method of the Housing data set.

Data name	Original number of records	Number of specific points	Cleaned data	R^2
T^2 ellipse method	0.9687	0.6934	0.2797	0.3479
TDDA	40.9841	41.2391	33.4516	40.1487
TMSM	74.5369	76.1351	17.5692	18.3622
T^2 ellipsoid method	0.0634	0.0165	0.0075	0.0076

TABLE 4: Data set description.

Data name	Number of samples	Number of independent variables	Number of dependent variables
DCQT	9	9	3
Housing	506	11	3
AirQuality	9357	10	3
CBM	11934	16	2

of several methods are compared at the same level of the learning and training set. Root mean square error (RMSE) and redetermination coefficient (R^2) were investigated, respectively. The experimental results are shown in Table 5.

According to the experimental results in Table 5, RMSE and R^2 of DBN-PLS are better than PLS and RBM-PLS. When the number of hidden layers (i.e., the number of RBM overlays) of the DBN-PLS method is 3, the effect is the best, the number of hidden layers increases continuously, and the effect also decreases gradually. The specific analysis process is as follows.

In the experimental data of Dachengqitang, the RMSE of PLS is much larger than that of RBM-PLS and DBN-PLS, and the RMSE of RBM-PLS is larger than that of DBN-PLS. When the number of hidden layers of the DBN-PLS method increases continuously, the RMSE increases continuously, that is, the effect decreases continuously. The RMSE of the five methods is 0.9687, 0.6934, 0.2797, 0.3420, and 0.3479, respectively. The R^2 value of DBN-PLS is greater than PLS and RBM-PLS, and the R^2 value of RBM-PLS is greater than PLS. When the number of hidden layers of the DBN-PLS method increases continuously, the R^2 value decreases continuously, that is, the effect decreases continuously. The R^2 values of the five methods are 0.6942, 0.7962, 0.9421, 0.9135, and 0.9105, respectively. When the number of RBM superposed by DBN-PLS is 3, 4, and 5, respectively, the values of RMSE and R^2 are 0.2797 and 0.9421, 0.3420 and 0.9135, and 0.3479 and 0.9105, respectively. Based on the above two evaluation indexes, it can be concluded that DBN-PLS is better than RBM-PLS and PLS for the experimental data of Dachengqitang, and RBM-PLS is better than PLS. When the number of RBM superposed by DBN-PLS is 3, the effect is most significant.

In the UCI standard data set, Housing is a medium data sample, while AirQuality and CBM are large data samples. Among all kinds of data, DBN-PLS has the best effect. In Housing and AirQuality data sets, RBM-PLS has the same effect as PLS, even slightly worse than PLS. For example, in the Housing data set, RMSE of RBM-PLS is larger than PLS, 41.2391 and 40.9841, respectively. Similarly, the R^2 value of RBM-PLS is less than PLS, 0.1813 and 0.1987, respectively. In the CBM data set, RMSE of RBM-PLS is less than PLS,

0.0165 and 0.0634, respectively. In contrast, the R^2 value of RBM-PLS is less than PLS, 0.5010 and 0.6935, respectively. The above effect is produced because the results of the RBM-PLS model are affected by the initial values. Different initial values have different effects, which will produce slightly worse effects than PLS. To sum up, DBN-PLS has the most significant effect on the above two evaluation indexes, whether it is the experimental data set of Dachengqi decoction or the UCI standard data set. Especially, when the number of RBM in the model is 3, the effect is the best. On the contrary, the superposition number increases, and the effect decreases. This shows that when DBN-PLS is used for feature extraction, the superposition number of RBM is 3, and the effect is the best.

To display the experimental results more intuitively, graphs are drawn to reflect the fluctuation of root mean square error (RMSE) and redetermination coefficient (R^2). Due to the different orders of magnitude of RMSE and R^2 of each data set, to compare the fluctuation of RMSE and R^2 of each data set in different methods conveniently, the experimental result data are centrally processed and mapped to [0,1], Figures 9 and 10 are drawn.

From Figures 9 and 10, it can be seen more intuitively that DBN-PLS has significantly improved effect in various indexes, and the effect is better than RBM-PLS and PLS. When the number of hidden layers of the DBN-PLS method is 3, the effect is the best. The number of hidden layers increases continuously, and the effect of each index shows a slow downward trend. In the DCQT data set, RBM-PLS has better performance than PLS, while in the UCI standard data set, RBM-PLS has slightly worse performance than PLS. This is because the results of the RBM-PLS model are affected by the initial values. The initial values selected are different, and the effects are different.

To sum up, among the above four groups of experiments, DBN-PLS has the best effect of multiple linear regression, which indicates that the fitting effect is better. As the number of hidden layers in DBN-PLS increases, the effect weakens, indicating that when DBN-PLS is used for feature extraction, the best effect is when the number of RBM overlays is 3. In different data sets, the effects of RBM-

TABLE 5: Comparison of experimental results.

Data name	PLS		RBM-PLS		DBN(3)-PLS		DBN(4)-PLS		DBN(5)-PLS	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
DCQT	0.9687	0.6942	0.6934	0.7962	0.2797	0.9421	0.3420	0.9135	0.3479	0.9105
Housing	40.9841	0.1987	41.2391	0.1813	33.4516	0.3680	37.2034	0.3017	40.1487	0.2330
AirQuality	74.5369	0.9247	76.1351	0.6209	17.5692	0.9487	18.0087	0.9460	18.3622	0.9439
CBM	0.0634	0.6935	0.0165	0.5010	0.0075	0.7940	0.0075	0.7919	0.0076	0.7879

FIGURE 9: RMSE comparison of different methods under four groups of experimental data.

FIGURE 10: Comparison of methods R^2 under four groups of experimental data.

PLS and PLS are slightly different because the results of the RBM-PLS model are affected by initial values. To sum up, the partial least squares method based on DBN has good adaptability to TCM data.

6. Conclusions

The TCM data are modeled based on the partial least squares method. In addition to the same process as the modeling methods in other fields, there are also features such as large data volume, multiple correlations, nonlinearity, and random distribution of data. The main performance is ① Modeling samples selected for massive data affect the accuracy of the PLS model. ② Partial least squares method has

limited effect in identifying and eliminating abnormal observation data records. ③ Limitations of principal component extraction methods.

Therefore, the work of this paper is improved based on the above partial least squares method:

- (1) The experimental data of the traditional Chinese medicine: there are often many variables related to the dependent variable, and even the observation data of some independent variables are costly, but they can now be collected by leading-edge equipment. If such variables are included in the regression model, not only will the calculation amount be increased and the application cost of the model be increased but also the sample data are multidimensional, that is, a sample is characterized by multiple features. The dimensions and numerical values of these features are different, making the model unstable. Therefore, the work of this paper is firstly through the Z-score standardization method for data standardization processing compared with the min-max standardization method [15]; it is not necessary to redefine the values of max and min so that different features have the same scale. Thus, when learning parameters, different features have the same degree of influence on the parameters.
- (2) High-quality data from drug trials are the basis for building analytical models. The experimental data are not only affected by the statistical distribution characteristics of the system parameters but also affected by the operating habits of the experimental personnel so that the number of occurrences of various dispensing groups is extremely uneven, and the observation data cannot contain large errors; otherwise, the result is not reliable. Therefore, the T^2 ellipse is used to observe the distribution of the sample points and the similarity structure on the t_1/t_2 plan and to find those singular points whose values are far from the sample points combined with the average level. Achieve the purpose of eliminating abnormal data. Compared to TDDA, TMSM, and T^2 ellipsoid method mentioned by Sun Shuwei [16] and Cui Lizhen [17], the T^2 ellipse method can better eliminate noise and improve the value of the re-determination coefficient of PLS.
- (3) In the partial least squares algorithm, if the number of principal components selected to participate in the regression is too small, some useful information in the original independent variable matrix is ignored,

which affects the fitting ability and prediction accuracy of the model; if the number of selected principal components is too large, although the fitting accuracy of the model can be improved, overfitting may occur due to introduction of irrelevant noise, resulting in a decrease in prediction accuracy. The most common method for extracting the number of components by partial least squares is to use the cross-validation (CV) [18] method to investigate the change of model prediction ability after adding new components, but the existing CV is insufficient. If the number of verification sets is too small, it is easy to overfit; the criteria for determining the optimal number of verification sets need to be given.

This paper proposes a partial least squares method based on DBN, which makes full use of the joint probability density distribution in the DBN generation model to learn the upper-layer features of the data and avoids the selection of the number of principal components. Combined with the characteristics of PLS, in the case of small sample size, it is still possible to regression modeling and maximize the relationship between independent variables and dependent variables and give full play to the advantages of the algorithms themselves. Experiments on Chinese medicine data and UCI standard data sets show that DBN-PLS significantly improves the accuracy of the regression equation and the expression of nonlinear structures. Compared with the conventional partial least squares method, it is more in line with the theoretical value and better solves the problem of determining the number of principal components in the partial least squares method.

Data Availability

The data in this article can be shared and used free of charge, but the data not listed in this article come from the Key Laboratory of traditional Chinese medicine preparation of Jiangxi University of Traditional Chinese Medicine, which involve patient privacy and business secrets and cannot be shared and used.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (nos. 61762051 and 61562045), the Jiangxi Province Science and Technology Plan Projects (nos. 20171BBG70108 and 20171ACE50021), and the Project of Jiangxi Health Planning Commission (nos. 20165537, 2017A284, 2017A313, and 20185516).

References

- [1] Z. G. Shang, Y. H. Dong, and M. M. Li, "Robust feature selection and classification algorithm based on partial least square regression," *Computer Application Research*, vol. 37, no. 3, pp. 871–875, 2017.
- [2] H. Abdi and L. Williams, B. Reisfeld and A. N. Mayeno, *Partial Least Squares Methods: Partial Least Squares Correlation and Partial Least Square Regression*, vol. 930, pp. 549–579, Humana Press, Totowa, NJ, USA, 2013.
- [3] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
- [4] L. Zhou, *Research on Feature Extraction Based on Nonlinear Partial Least Square*, Nanjing University of Technology, Nanjing, China, 2011.
- [5] J. C. Qin, "An optimal band selection method for hyperspectral images based on kernel partial least square method," *Journal of Photogrammetry*, vol. 30, no. 2, pp. 171–176, 2013.
- [6] Z. P. Zhu and J. Q. JianqiangDu, "Partial least square optimization method integrating restricted Boltzmann machine," *Calculating Machine Works*, vol. 7, pp. 193–202, 2017.
- [7] J. Magnanensi, F. Bertrand, M. Maumy-Bertrand, and N. Meyer, "A new universal resample-stable bootstrap-based stopping criterion for PLS component construction," *Statistics and Computing*, vol. 27, no. 3, pp. 757–774, 2017.
- [8] G. A. F. Seber and A. J. Lee, *Linear Regression Analysis*, Wiley, Hoboken, NJ, USA, 2012.
- [9] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [10] F. Y. Liu et al., "Review of deep confidence network models and applications," *Computer Engineering and Applications*, vol. 54, no. 1, pp. 11–19, 2018.
- [11] X. S. Yang, *Nature Inspired Meta-Heuristic Algorithms*, Luniver Press, Frome, UK, 2010.
- [12] UCI Machine Learning Repository, <http://archive.ics.uci.edu/ml/datasets/Air+Qu>.
- [13] G. R. Zhang, *On the Detection Method Multidimensional Outliers*, Shenyang Sport University, Shenyang, China, 2011.
- [14] J. X. Guo, *Study on Improved High-Dimension and Nonlinear Partial Least-Squares Regression Method and Applications*, Tianjin University, Tianjin, China, 2010.
- [15] X. J. Ji, B. GuiS, and D. WangG, "Using min-max normalization to measure the differences of regional economic growth—a case study of yulin area, shanxi province," *Economy And Management*, vol. 30, no. 3, pp. 54–56, 2016.
- [16] S. W. Sun, *Reconstruction Model of Support Vector Machine and Research on SMO Algorithm*, Xidian University of Electronic Technology, Xi'an, China, 2009.
- [17] L. Z. Cui, "Multidimensional scaling positioning algorithm based on RSSI in underground coal mine," *Safety in Coal Mines*, vol. 48, no. 6, pp. 108–111, 2017.
- [18] C. Y. Sun and M. Y. Chang, "The study of beijing in water demand forecast model," *Journal of Applied Statistics and Management*, vol. 36, no. 6, pp. 1049–1058, 2017.