Optimization Method of Fusing Model Tree into Partial Least Squares

Fang YU[a], Jian-Qiang DU[*], Bin NIE[b], Jing XIONG, Zhi-Peng ZHU, and Lei LIU

School of Computer, Jiangxi University of Traditional Chinese Medicine, Nanchang, 330004, China
[a]eshter_yuu1992@163.com, [b]460092757@qq.com
** Corresponding author: jianqiang_du@163.com*

Abstract: Partial Least Square (PLS) can't adapt to the characteristics of the data of many fields due to its own features multiple independent variables, multi-dependent variables and non-linear. However, Model Tree (MT) has a good adaptability to nonlinear function, which is made up of many multiple linear segments. Based on this, a new method combining PLS and MT to analysis and predict the data is proposed, which build MT through the main ingredient and the explanatory variables(the dependent variable) extracted from PLS, and extract residual information constantly to build Model Tree until well-pleased accuracy condition is satisfied. Using the data of the maxingshigan decoction of the monarch drug to treat the asthma or cough and two sample sets in the UCI Machine Learning Repository, the experimental results show that, the ability of explanation and predicting get improved in the new method.

1 Introduction

In real life, some actual processes are complex nonlinear processes, the nonlinear relationship reflects not only between independent variable and the independent variables, but also between independent variable and dependent variable[1]. Since the experiment and some objective or non-objective factors, it usually results in small data sample set, and some sample data size is even less than the sample dimension. PLS[2] was first proposed by Herman Wold, including principal component analysis (PCA),Canonical Correlation Analysis(CCA) and Multiple Linear Regression(MLR). It has a good explanatory power for the data with multiple independent variables[3], multiple dependent variables and a small sample size, however the nature of linear regression in PLS can not completely reflect the characteristic of the traditional Chinese medicine data.

In 1996, Qin S.J. [4] proposed a RBF neural network combined with PLS, which can establish a good nonlinear prediction model, while it's hard to explain the characteristic for its linear approximation to continuous function. Paper [5] proposed an algorithm with fuzzy neural network model embedded into the iterative PLS, and achieved good nonlinear mapping effect, but the results of the model is vulnerable to membership function. In 2013, paper[6] proposed a Kernel Partial Least Squares Method, which mapped the nonlinear data to high dimension linear space with the help of kernel function, to extract the relationship between dependent variables and independent variables as much as possible, the method can well reflect the nonlinear structure contained in sample data, however, choosing a good kernel function is extremely difficult.

Model Tree[7] is an algorithm proposed by Quinlan, which the leaf node adopt multiply linear function instead of the average processing method in traditional regression tree，it is constructed by several multiply linear pieces, and has a piecewise linear approximation to any unknown variable distribution trend, not only is the model structure simple, but also easy to explain the nonlinear data, with high efficiency and good robustness. Based on this, to make up the defect of the linear nature in PLS method within the model, it uses MT as the internal model in PLS to interpret the nonlinear characteristics in TCM data.

2 Partial Least Square(PLS)

Partial Least Square algorithm can not only build regression model for data with multi-independent variables, dependent variables, but also adapts to the situation when sample sizes are less than variables numbers[8].

The introduction of PLS is as follows:

To make the explanation easy, assume there are independent variables set $X = (x_1, x_2, \mathrm{L}\ , x_i, \mathrm{L}\ , x_p)$ and dependent variables set $Y = (y_1, y_2, \mathrm{L}\ , y_j, \mathrm{L}\ , y_q)$. t， u are the linear weight combination of independent variables and dependent variables, and they must satisfy the two conditions below: ①try to carry variance information of independent variables and dependent variables respectively as soon as possible; ②the correlation coefficient between the two is largest.

Extracted first principle component information from

X, Y as t_1, u_1, then make the t_1, u_1 do multiply linear regression, judge the residual information ,if satisfy the requirements, then terminate the process, else continue to extract principle component information from the residual information, the above procedure continues until satisfactory accuracy is achieved.

3 Model Tree(MT)

Model Tree(MT)[9] adopt multiply linear regression rather than the average processing method in leaf node like traditional classification and regression tree(CART). It divides the sample data into several discrete areas according to certain rules and choose a suitable model to construct a regression model for the areas. the model includes: tree building, searching for splitting attributes, handling internal nodes, pruning, smoothing, and prediction. The detailed introduction about how to build Model Tree browse paper[10-11].

4 MTree-PLS

4.1 The Algorithm Flow for MTree-PLS

MTree-PLS is formed by two modular: one is PLS, to extract principle component and to eliminate multiple variable correlation, the other is MTree, to establish the relationship between independent variables and dependent variables and to make the model nonlinear. MTree-PLS is built over the traditional partial Least Square(PLS), the external model still adopt the original method to extract principle component t, the internal model build Model Tree with the extracted principle components and dependent variables, then do multiply linear regression in leaf nodes of Model Tree, and get the predicted value, if the residual information satisfy the pre-defined condition, stop to build tree, else ,continue to build Model Tree by the residual information till satisfied accuracy obtained.

The process of the MTree-PLS algorithm is as follows.

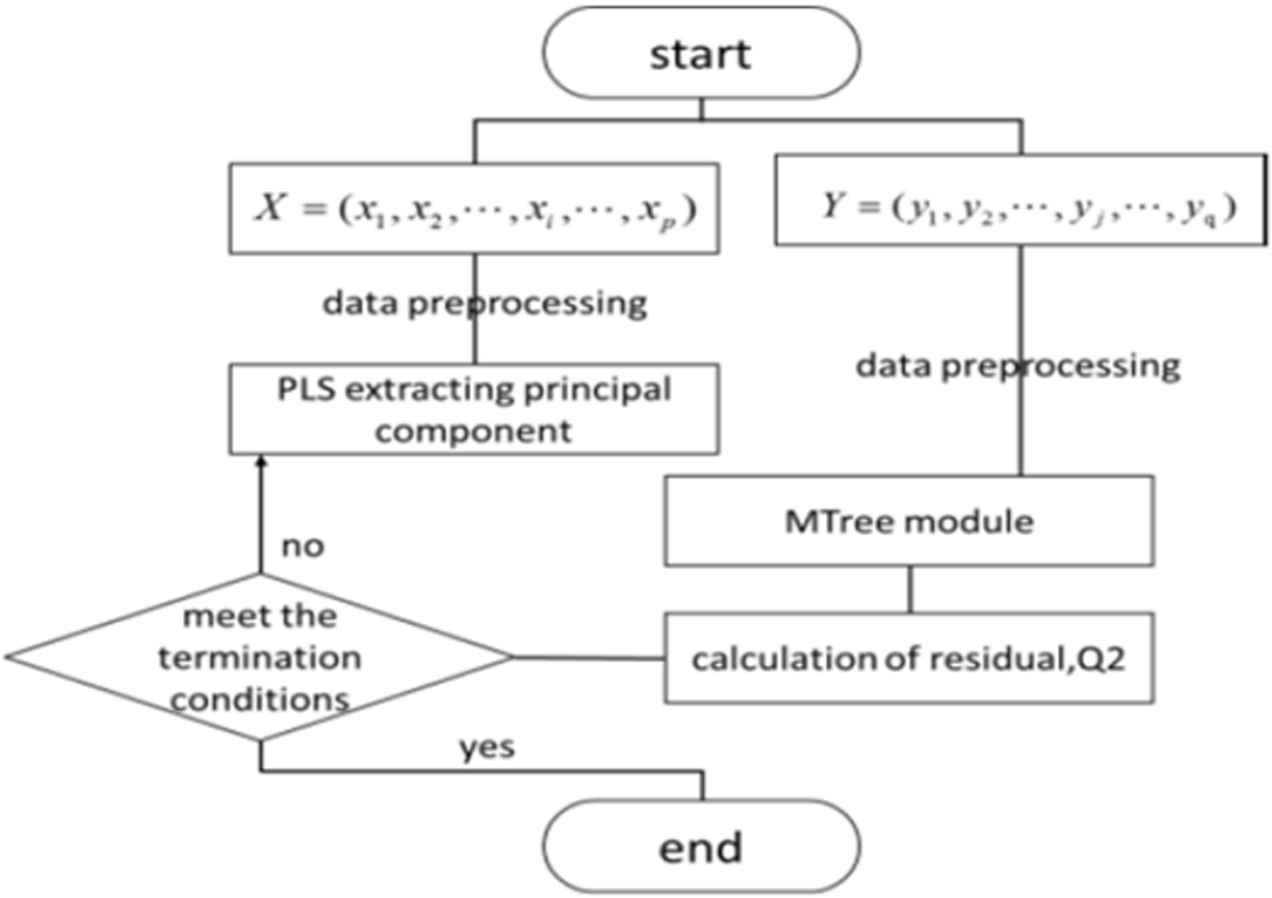

Figure 1. the process of the MTree-PLS algorithm

4.2 An Algorithm for Constructing Model Tree of Principal Components

The method for using Principal Components to construct Model Tree is to build Model Tree through Principal Components t_1 extracted from independent variable by PLS and the original dependent variable Y. Due to the t_1's linear characteristics, it can find the best split point by calculating the error of multiply linear regression between t_1 and Y, then split the t_1 into two subset according to the best split node, continue to split the subset in the manner above till the number of leaf nodes is less than the threshold predefined or the error fluctuation is not that obvious. The algorithm for using Principal Components to construct Model Tree shows in algorithm 1.

Algorithm 1: using Principal Components to build Model Tree

Input: principal component t, the attributeListY
Output: the built Model Tree RT
Step01 construct the basic regression tree RT.
Step02 Search for splitting attributes.

search the splitting attribute of the subtree to internal nodes in RT, then make them form a set, which is called regression attributes.

Step03 Handles the internal nodes

select the current node data samples and part or all of its regression attributes to regression, traverse the current regression model, choose the regression model with the least error of the current sample data, as a regression model of the current node.

Step04 pruning

Traversing from bottom to top for RT, record all the leaf nodes

Build linear fitting regression equation f_{parent}, f_{leaf} respectively for each parent node and leaf node.

if $f_{parent}RMSE < f_{leaf}RMSE < f_{leaf}RMSE$

Pruning for the subtree

else preserve the leaf node

In particular, if the subtree 's parent node is the root, no pruning operation.

end

Step05 Smoothing

Traversing from bottom to up for the RT

Combine the fitting function of the child nodes and parent nodes into a new linear function,

$$f_{new} = \frac{n \cdot f_{child} + k \quad f_{parent}}{n+k} \qquad (1)$$

if $f_{new}RMSE - f_{child}RESE < Q.$, ($Q$ is the fixed threshold)

$f_{child} \rightarrow f_{new}.$

else no smoothing is performed

end

Step06 return RT

End of the algorithm

Introduction：RMSE is the Root Mean Square Error, expressed as follows

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (f_{obj,i} - T_{\mathrm{mod}el,i})^2}{n}} \qquad (2)$$

In the formula (1), n is the number of sample data of current father node, k[12] is a smoothing constant(defaults k=15), f_{child} and f_{parent} are a fitting function of current leaf nodes and current parent nodes respectively, f_{new} is a fitting function under smoothing.

4.3 Fusing Model Tree into Partial Least Squares

MTree-PLS get the principal components t_1 by using PLS, then adopt t_1 to do multiply linear regression with X, and meanwhile build MT with Y and t_1, it still employs the original PLS to get residual information of X, and the residual information of Y is the disparity between Y and the predicted $\hat{Y}$ which is the regression values by the leaf nodes of Model Tree. If not meet the accuracy requirements, then continue to extract the main component by using residual information and use the principal component and the residual information of Y to continue to build trees. Repeat the above procedure, terminating the algorithm until a nonlinear model with satisfactory precision is constructed.

The detailed steps are as follows:

（1） Data preprocessing

normalizing X,Y, to obtain E_0 and F_0;

（2） Extracting principal component

t_1 is the first principal component extracted from E_0, existing $t_1 = E_0 w_1.$, $u_1 = F_0 v_1.$, w_1, v_1 is respectively the weight of E_0 and F_0, and $\| w_1 \| = 1$, $\| v_1 \| = 1.$, make:

$$\begin{aligned} &\mathrm{var}(t_1) \rightarrow \max \\ &\mathrm{var}(u_1) \rightarrow \max \\ &r(t_1, u_1) \rightarrow \max \end{aligned} \qquad (3)$$

According to the $\mathrm{cov}(t_1,u_1) = r(t_1,u_1)\sqrt{\mathrm{var}(t_1)\,\mathrm{var}(u_1)}$ $\Rightarrow \mathrm{cov}(t_1,u_1) \rightarrow \max.$,

and $\mathrm{cov}(t_1,u_1) = \frac{1}{n} < t_1, u_1 > .$, plug $t_1 = E_0 w_1.$, $u_1 = F_0 v_1.$ in this，then：

$$\max < E_0 w_1, F_0 v_1 > = (E_0 w_1)^T (F_0 v_1) \quad \max \qquad (4)$$

According to the principle of Lagrange multiplier, respectively calculated w_1、v_1 are the eigenvectors corresponding to the largest eigenvalue of $X^T Y Y^T X$ and $Y^T X X^T Y$. Thus, it is easy to calculate the corresponding t_1.

(3) Regression of the principal component model tree

Extracted principal components t_1 from E_0 ,implements the linear regression about E_0 to t_1, $E_0 = t_1 p_1' + E_1$, whereby $p_1 = E_0^T t_0 / \|t_1\|^2$。Due to t_1 carry the information of dependent and independent variables, and will make t_1 and $F_0(j)$, (j=1, 2, 3, . . . q) build model tree *tree*(1j) separately, and calculates the corresponding prediction coefficient $predict(t_{1j})$ of the model and residual information matrix $E_1 = E_0 - t_1 p_1'$, $F_1 = F_0 - \sum_{j=1}^{q} t_1 * predict(t_{1j})$.

(4) Judge the condition, stop the cycle operation

It can be judged whether the model satisfied the precision requirement according to the explanation of the model(R^2) or the Sum of Residual Squares of training set (SSETrain). if satisfy the condition, stop counting, if not,

then using residual messages E_1、 F_1,follow the steps(2), until meet the conditions.

(5) Integrating the equation of MTree-PLS

$$F = \sum_{j=1}^{q} t_1.predict(t_{1j}) + \sum_{j=1}^{q} t_2.predict(t_{2j}) + \mathrm{L} + \sum_{j=1}^{q} t_m.predict(t_{mj}) \quad (5)$$

The standardized for coefficients and reduction multiple regression equation of Y to X.

The algorithm of Fusing Model Tree into Partial Least Squares as follows:

Algorithm 2: the algorithm of Fusing Model Tree into Partial Least Squares

Input: The original sample data(Dataset(D))，the property list of independent variable: attributeList X_{n*p}，dimension: p，the property list of dependent variable: attributeList Y_{n*q}，dimension: p.

Output: equation of MTree-PLS

Step01: Extracting the attributeListX，attributeListY as (X,Y) from the original data，and Standardized the (X,Y) as (E_0,F_0)

Step02: partial least squares regression(PLSR)

$i=1$

while judge the number i of the principal component satisfying the requirements or not

Based on the Lagrange principle to get weight coefficient w_i， v_i

Calculate the maximum eigenvector w_i、v_i corresponding to the maximum eigenvalue of matrix $F_{i-1}^{'}E_{i-1}E_{i-1}^{'}F_{i-1}$ and $F_{i-1}^{'}E_{i-1}E_{i-1}^{'}F_{i-1}$

Calculate the score vector $t_i = E_{i-1}w_i$ and build tree $tree$(i)

Get the corresponding predicting coefficient $predict(t_{ij}),(j=1,2\mathrm{L}\ q)$

Regression equation $E_{i-1} = t_i p_i^T + E_i$ and $F_{i-1} = F_i + \sum_{j=1}^{q} t_i.predict(t_{i,j})$

Load vector $p_i = X_{i-1}^T t_i \big/ \|t_i\|^2$

Get the residual information matrix E_i and F_i

i=i+1

end

Step03

Integration the equation of MTree-PLS

$$F = \sum_{j=1}^{q} t_1.predict(t_{1j}) + \sum_{j=1}^{q} t_2.predict(t_{2j}) + \mathrm{L} + \sum_{j=1}^{q} t_m.predict(t_{mj}) \quad (6)$$

Anti-standardized to the coefficient of the equation, and get the equation of the Y and X.

Step04 end

5 Experimental Analysis

The experimental data is from the key laboratory of Modern Preparation of TCM, Ministry of Education in Jiangxi University of Traditional Chinese medicine, which supports us with the precious data of maxingshigan decoction of the monarch drug to treat the asthma or cough, the paper still choose another two sample sets in the UCI Machine Learning Repository, namely yacht_hydrodynamics[13] and CCPP_Folds5x2_pp[14] to testify the improved algorithm.

5.1 The Explaining about the Experimental Data

The part of data about maxingshigan decoction of the monarch drug to treat the asthma(mxsgpc) showed in Table 1, has a total of 46 samples. it is about the impact of pharmacological indicators about the blood medicine composition in rats under 10 distinct dosage of herbal ephedra respectively. There are five compositions about the blood medicine composition in rats, There are two pharmacological indicators namely, incubation period (Unit: s) and cough duration (Unit: min). The first five compositions is the independent variable, the rest two is dependent variable.

Table 1. the data of maxingshigan decoction of the monarch drug to treat the asthma

ephedrine	pseudoephedrine	methyl ephedrine	wild black cherry glycosides	licorice glycosides	incubation period(s)	cough duration（min）
0.93	0.52	0.14	0.00	0.51	79	8
0.97	0.48	0.16	0.34	0.53	51	18
0.95	0.53	0.17	1.67	0.48	44	22
0.92	0.59	0.39	0.00	0.57	66	9
1.09	0.43	0.41	0.00	0.42	71	19
…	…		…	…	…	…

The part of data about maxingshigan decoction of the monarch drug to treat the cough(mxsgzk) showed in Table 2, has a total of 62 samples. it is about the impact of pharmacological indicators about the blood medicine composition in rats under 10 distinct dosage of almond respectively. There are five compositions about the blood medicine composition in rats There are one pharmacological indicators namely, cough times. The first five compositions is the independent variable, the rest one is dependent variable.

Table 2. the data of maxingshigan decoction of the monarch drug to treat the cough

ephedrine	pseudoephedrine	methyl ephedrine	amygdalin	wild black cherry glycosides	cough times
402.00	369.93	48.46	0.79	1.87	25
491.00	385.79	47.32	0.00	0.00	50
412.00	314.74	41.28	0.00	0.00	35
519.00	316.81	39.50	0.61	1.42	37
387.09	290.05	15.29	0.81	3.17	40
…	…	…	…	…	…

The description of yacht_of hydrodynamics(yacht) and CCPP_Folds5x2_pp(CCPP) shows in http://archive.ics.uci.edu/ml/.

5.2 Analysis of the Procedure and Result of the Experimental

To validate the effect of the new model, it adopt the traditional PLS and Random Forests Regression (RFR) as a contrast. Based on this, it divides the original data randomly in the proportion of 7:3, 70% of them are training set and the rest are test set and adopts the experimental data of maxingshigan decoction of the monarch drug to treat the asthma or cough and two sample sets in the UCI Machine Learning Repository to compare.

Table 3. Description of the experimental data

data name	sample number	independent variables number	dependent variables number	train sample number	test sample number
maxingshigan decoction cure asthma	46	5	2	32	14
maxingshigan decoction cure cough	62	5	1	43	19
yacht_hydrodynamics	308	7	1	216	92
CCPP_Folds5x2_pp	9568	4	1	6700	2868

Table 4. The result among PLS、RFR and MPTree – PLS

	PLS		MPTree-PLS		RFR	
	SSETrain	SSETest	SSETrain	SSETest	SSETrain	SSETest
maxingshigan decoction cure asthma	20580.6507	30434.7553	7485.3271	14618.4831	2184.2919	21114.3382
maxingshigan decoction cure cough	3841.8357	1761.1864	1568.5878	1653.9989	3558.0025	1777.0677
yacht_hydrodynamics	15990.5122	8151.6082	455.5790	1400.0808	8706.8571	4900.1754
CCPP_Folds5x2_pp	161316.0535	70113.7642	128136.5498	59418.3421	145551.7916	64731.8109

Review the Sum of Squares for Error of Training Set(SSETrain), Sum of Squares for Error of Testing Set(SSETest) separately. The result is as follow. It can be seen clearly in the figure below:

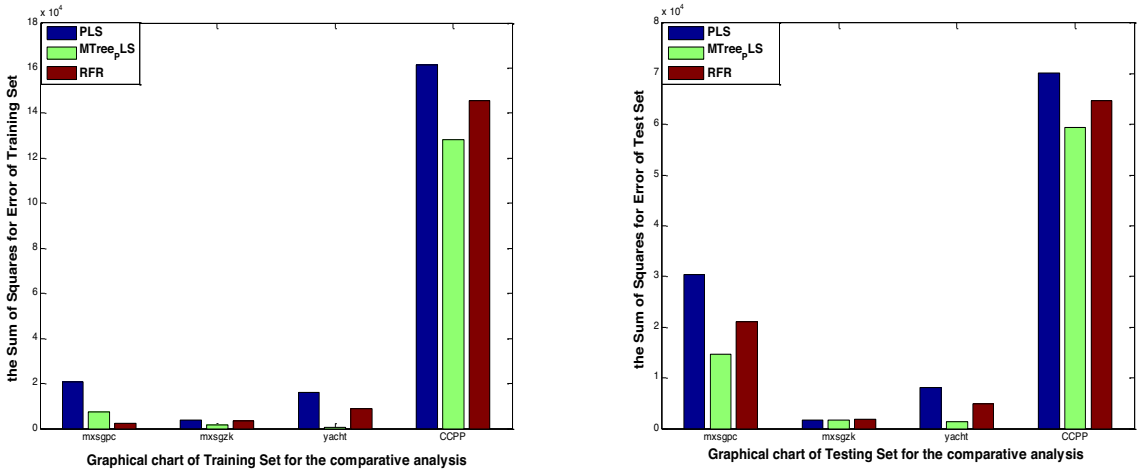

Figure 2. Graphical chart for the comparative analysis

From Table 4 and Fig 2, we can see the four points below:

Firstly，PLS has a poor ability of explanation and prediction to nonlinear data and shows obvious inadaptability than the improved PLS.

Secondly, No matter SSETrain or SSETest in the data of maxingshigan decoction cure asthma or cough, yacht_hydrodynamics and CCPP_Folds5x2_pp, compared with PLS and RFR algorithm, the improved PLS has different level of good effect.

Thirdly, although the SSETrain of the improved algorithm is not so well than the RFR', but the prediction ability of RFR is poor in evidence.

Last, the improved PLS method not only has good adapt ability to TCM data, but also has good adapt ability to UCI'S nonlinear data with middle or huge level sample.

In summary, the Model Tree shows a strong analytical and predictive effect for multidimensional nonlinear data. whether it is for small or large sample data, in the degree of interpretation of the model, or the analysis and prediction of data, The improved algorithm is superior to the PLS and the RFR.

5.3 The analysis of the Algorithm's Time Complexity

For the PLS, the time complexity is mainly expressed in the principal component extraction. Since the eigenvalues and eigenvectors can be solved by the singular value matrix, only the covariance matrix exists in the time complexity, and the time complexity is $O(n^2)$. For the model tree, the time complexity is mainly reflected in the part of the tree built, as $O(n^2)$. For the improved algorithm, it is assumed that the number of the principal components extracted is m, and each time a tree is built when extracting a principal component, so the time complexity of the improved algorithm is $O(mn^2)$.

Summary

To deal with Partial Least Squares can't well explain nonlinear data, the thesis put forward Fusing Model Tree into Partial Least Squares. The Partial Least Square make full use of the nonlinear characteristic that the regression model constructed by model tree is formed by many multiple linear segments and after a series of experiments, the conclusion shows the improved algorithm can well explain the level of the model and have more accurate prediction ability. However, the number of leaf node directly decide the calculate result of model. Thus, what we should study further is to choose appropriate leaf node.

Acknowledgement

This work is supported by the Key Laboratory of modern preparation of Traditional Chinese Medicine (TCM), Ministry of Education and two National Natural Science Foundations (61363042 & 61562045). This research also is supported by a major project of Jiangxi Natural Science Foundation (20152ACB20007) and the Postgraduate innovation fund of Jiangxi University of Traditional Chinese Medicine(JZYC16S05).

References

1. Zhang Boli, Wang Yongyan. Components compatibility theory and practice of development of modern Chinese medicine prescriptions, the key scientific issues of basic research [M]. Shenyang: liaoning science and technology press, 2010.
2. Wold H. Nonlinear estimate by iterative least squares procedures. Research Papers in Statistics, 1966, Wiley, New York.
3. Abdi H, Williams L. Partial Least Squares Methods: Partial Least Squares Correlation and Partial Least Square Regression [M]. Reisfeld B, Mayeno A N, Humana Press, 2013: 930, 549-579.
4. Qin S J, McAoy T J. Nonlinear FIR modeling via a neural net PLS approach. Computer & Chemical Engineering, 1996, 20(2):147-159.
5. Zhou Lin. Feature extraction method based on nonlinear PLS research [D]. Nanjing university of

science and technology, 2011.
6. Liu Yu. The modeling and simulation of the response surface based on local kernel PLS method [D]. Tsinghua university, 2013.
7. Quinlan J R. Learning with continuous classes[C]. Proceedings of the 5th Australian joint Conference on Artificial Intelligence. Singapore, 1992.
8. Abdi H, Williams L.Partial Least Squares Methods:Partial Least Squares Correlation and Partial Least Square Regression[M].Reisfeld B,Mayeno A N, Humana Press,2013: 930, 549-579.
9. Gu Yayun, Hu Linxian. The application of M5 model tree in thermal power plant load optimization [J]. Energy saving technology. 2013, 31(5): 426-429.
10. Frank E, Wang Y, Inglis S, et al. Using Model Trees for Classification[J]. Machine Learning, 1998, 32(1):63-76.
11. Wang Y, Witten I H. Induction of model trees for predicting continuous classes. Working paper 96/23[C] 1997.
12. Malerba D, Appice A, Bellino A et al. Stepwise Induction of Model Trees.[C] Congress of the Italian Association for Artificial Intelligence on Advances in Artificial Intelligence. Springer-Verlag, 2001:20-32.
13. UCI Machine Learning Repository[Z],[2013-01-03]. http://archive.ics.uci.edu/ml/datasets/Yacht+Hydrodynamics.
14. UCI Machine Learning Repository[Z],[2014-03-26]. http://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant.