Random Forest Regression Based on Partial Least Squares

Connect Partial Least Squares and Random Forest

Zhulin Hao[1], Jianqiang Du[1*], Bin Nie[1], Fang Yu[1], Riyue Yu[2] and Wangping Xiong[1]
[1]Computer School, Jiangxi University of Traditional Chinese Medicine, Nanchang, 330004, China
[2]College of Pharmacy, Jiangxi University of Traditional Chinese Medicine, Nanchang, 330004, China
*Corresponding author

***Abstract*—Partial Least Squares (PLS) Regression is lack of theoretical guidance of rules to achieve the nonlinear by the quasi linearization rule, and its accuracy declines in the face of the unknown variables distribution. Furthermore, the loss of information is easy to arise for the mean processing of the leaf in the Regression Tree of the traditional Random Forest Regression. On this basis, Partial Model Tree (PMT) is proposed combining Partial Least Squares Regression with Regression Tree, to achieve the nonlinear regression by constructing multiple linear fragments of Partial Least Squares to complete linear approximation of the unknown variables, and the information loss issue caused by that the leaf nodes are treated by direct mean processing is avoided, when PLS regression is used in the leaf nodes. It applies PMT to ensemble learning to build Partial Least Squares of Random Forests Regression (PLS-RFR), improving the generalization ability of PMT. The ability of explanation and predicting get improved in the experiment data of MaXingShiGan decoction of the monarch drug to treat the asthma or cough and five sample sets in the UCI Machine Learning Repository. Finally, it verifies that the PMT and RF-PLS possess a certain degree of validity and correctness.**

Keywords-partial least square; regression tree; random forest; linear approximation; TCM information

I. INTRODUCTION

The traditional Random Forest Regression (RFR) is a way of calculating the final predicted value. The procedure are as follows: firstly, using Bootstrap [1] to get the random sampling and extracting the training dataset in the original sample dataset; secondly, constructing various copies of the training dataset of distinct size; thirdly, creating distinct versions of the single basic regression tree in the every replica training dataset and calculating average value through the predicted value in all basic regression trees. But the basic regression tree, a kind of poor single learning machine, using the average value method to handle the leaf nodes directly, makes the system lose some valuable information when a leaf node has a suitable sample. Apparently, when needs to integrate of multiple versions of regression trees, the traditional Random Forest Regression has poor ability to explain the model. Meantime, the pruned regression tree is also very tedious and hard to explain, although regression tree provides the pruning method in the leaf node for the information loss. Basing on this, in the process of handling the leaf nodes, Quinlan [2] has provided a linear regression equation algorithm, which replace the average processing method in traditional regression tree, namely, M5 Model Tree. The Model Tree sets up a piecewise linear function by the leaf node, and divides the complex nonlinear system information into several multiple linear segments, has a piecewise linear approximation to any unknown variable distribution trend. Haijun Li et al [3] have proposed a mixed learning regression algorithm utilizing Naive Bayes in the leaf node of the Regression Tree, and the algorithm achieves good results in the UCI machine learning. Using LogitBoost to build Overlay regression in the leaf node of the decision tree, Kun Zhang et al [4] get a new LC Tree Model Tree algorithm which can analyze the feature of telecom users to predict the number of people offline. Besides, Partial Least Squares regression (PLSR) [5] solves the problem that traditional Multiple Regression can still be effective to construct regression model when the Gauss Markov assumptions invalidate. There are no strict restriction for the capacity of sample and the severity of independent variables multicollinearity and the variable number. Huiwen Wang et al [6] have detailedly discussed the application of Partial Least Squares in the multicollinearity elimination and auxiliary analysis and quasi-linearization aspects, verifying that Partial Least Squares has great benefits in system information recognition and modeling reliability comparing to traditional multiply regression. But Partial Least Squares nonlinear method is a quasi-linearization regression method that conducts the nonlinear preprocessing transformation of the data using quasi linear rules [7]. The transformation can't reflect the nonlinear features of the data when facing the unknown knowledge, and lack of theoretical guidance. At this moment, Partial Least Squares nonlinear model is not well in precision. Considering the insufficient of Partial Least Squares method in the nonlinear and the deficiency of the M5 model tree needing to satisfy Gauss Markov assumptions when using classical multiple regression to deal with leaf nodes, it adopts Partial Least Squares to process the leaf node of the M5 Model Tree. According to the model, Random Forest Regression of Partial Least Squares is constructed by multiply distinct versions of this modified M5 model. This method, achieving the nonlinear regression by constructing multiple linear fragments of Partial Least Squares to complete linear approximation for the unknown variables, making information loss issue gets avoided when the leaf nodes is directly for mean with appropriate samples, optimizes

the traditional Random Forest Regression and makes up the flaw of Partial Least Squares method in the nonlinear.

II. PARTIAL LEAST SQUARES MODEL TREE

The Model Tree is obviously far superior to the conventional linear regression or the basic regression tree, adopting multiple linear regression model to nonlinear approximation of a continuous function. Not only is the Model Tree smaller and easier to be explained than the basic regression tree, but also much lower error generated in the sample data. The traditional Model Tree is built by mending the basic regression tree with using two important steps about pruning and smoothing, remedying the fault of the multiply regression. However, it can't be able to be integrated in the ensemble learning. One of necessary conditions for ensemble learning is a weak learner and the pruning significantly increases the computational time and space. Therefore, Partial Model Tree (PMT) is proposed, adopting Partial Least Squares regression to improve the traditional Model Tree.

Definition 1: Dividing the sample data into two parts by the principle of maximum change of variance, there are some conditions to determine whether it continues dividing or not, one of the conditions is the maximum allowable value of decreased space of leaf, the other is maximum allowable number of sample of leaf, if the node is the leaf then Partial Least Squares or the average method is handled according to the basic information of the sample. The regression tree is called as Partial Model Tree (PMT).

PMT doesn't construct a basic regression tree directly, after dividing the tree in a rule of the maximum fluctuation of total variance decline space, it selects average method or PLSR to build the linear model as leaf node directly. PLSR is especially suitable to solve the sample data where there is a feature of small sample size, multicollinearity and existing system noise. Therefore, PMT is better in those aspects than the basic regression tree and the traditional Model Tree for the same sample data. Similarly, it is easier for the pruning of PMT and increases the accuracy by the smoothing of Model Tree. The main algorithm in the PMT is as follows:

Algorithm 1: Partial Least Squares Model Tree

Dataset (D): The original sample data
AttributeList: Independent variable attribute list
PartialModelTree(Dataset, AttributeList)
Step1 Create root
Step2 Handle leaf
if sample size < MaxSampleSize or a column is exactly the same
Average mean for the dependent variable
else jump to Step 3
Step3 Partial Least Squares Regression [8]
Extract (X,Y) in the Dataset according to the AttributeList
The data standardization of (X,Y) is (E_0,F_0)
$i=1$
while the number of principal components i meets to the requirements
Singular value decomposition for $E_{i-1}^T F_{i-1}$
Get the feature vector axis (w_i,v_i) about (E_0,F_0)
Calculate principal component score $t_i = E_{i-1}w_i$ and $u_i = F_{i-1}v_i$
Loading vector is $p_i = X_{i-1}^T t_i / \|t_i\|^2$ and $r_i = F_{i-1}^T t_i / \|t_i\|^2$
Regression equation is $E_{i-1} = t_i p_i^T + E_i$ and $F_{i-1} = t_i r_i^T + F_i$
Residual information matrix is E_i and F_i
end
Integrate Partial Least Squares equation
Step 4 Dividable attribute judgment
Divide sample set according to each attribute of the AttributeList
Each attribute is divided into a sample set
Calculate the dividable mixed total variance after dividing the sample set
Calculate the mixed total variance before dividing the sample set
Step 5 Dividable judgment
if the total variance after partition – the total variance before partition > MaxDrop
Divide the sample data by (bestAttribute, bestAttributeValue)
Get the sample subset T1 and T2 after partition
else jump to Step 3
Step 6 Recursion
Recursion in the sample subset T1 and construct Partial Model Tree on (T1, AttributeList)
Recursion in the sample subset T2 and construct Partial Model Tree on (T2, AttributeList)
Step 7 end

The Algorithm 1 mainly involves two important parameters. First is the maximum allowable number of sample (MaxSampleSize) in the leaf nodes. It is usually fixed that the number of layer in building tree is not too high, according to the sample size, 10%-20% of the sample is suggested. Second is the maximum allowable drop in space of the leaf node (MaxDrop). In the process of building tree, it usually sets the maximum allowable sample in the leaf node as first priority, and the MaxDrop needs to deploy precisely by examining and debugging in the detailed drawing of the tree. Considering Partial Least Squares method is applied to the leaf nodes and it can deal with multi-independent and multi-dependent variable problem, PMT can be stretched directly as a new way of supporting multivariate data analysis with multi-independent and multi-dependent variable. When dividing the origin sample data set, not only should the total variance of multi-dependent variable take into account about the entirety, but the mixed total variance produced after dividing the tree should also be considered about the entirety. The detailed flow chart of specific algorithm of PMT is shown in Figure I. Based on the principle of maximum fluctuation drop in space of the total variance in Model Tree [9], PMT looks for the division attribute and chooses the attribute with the maximum reduction of the expected error as the best division attribute in the current node. To measure the consistence of numeric data, it sets the sum of squares of subtraction between the average and the independent variable of every data as measuring rule, namely, the reduction of the total variance, the formula is as follows:

$$SDR = sd(T) - \sum_{i=1}^{2} \frac{|T_i|}{T} sd(T_i) \quad (1)$$

Here, T_1 and T_2 is the data subset of two child nodes, and they are divided from parent node T by division attribute.

For the number of the principal component in PLSR, it will not extract all the principal component to build the model. The existing literature already shows [10] that the number of the principal component is m when extracted information can represents the most of the origin sample data or there is little disturbance residual information, it can stop extracting, particularly, the two is equal to each other. Therefore, normally, in the analysis of Partial Least Squares, when it is up to 80% for the percentage rate of the extracting information to stop extract information from the origin regression data set. In the later calculation, it applies this rule uniformly to judge the extraction number of the principle component.

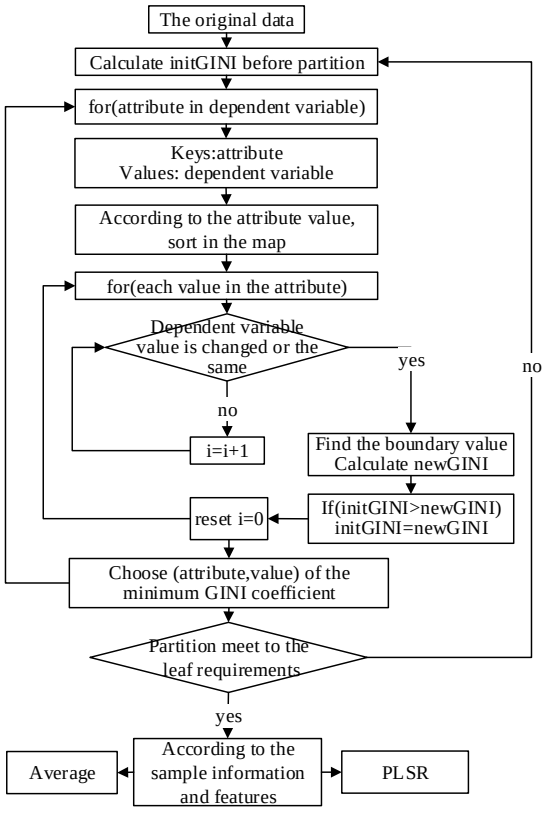

FIGURE I. THE ALGORITHM FLOW GRAPH OF PARTIAL LEAST SQUARES MODEL TREE.

III. RANDOM FOREST OF PARTIAL LEAST SQUARES

One branch of Random Forest is numerical regression, called Random Forest Regression algorithm. Random Forest Regression is an ensemble learning method to solve the same problem by lots of distinct versions of the individual learner, thus heavily improving the learning ability and generalization of system [11]. Because PMT is a weak individual learner like regression tree, it has a power of integrated into ensemble learning and enhances the ability of generalization for the unknown data.

Definition 2: Using the bootstrap random sampling on the sample data, called as the training set in the bag (ITB), building PMT model on many distinct versions of the training set, the predicted result of the final model is made decisions by the mean in many PMTs. This model is called as Random Forest Regression of Partial Least Squares (PLS-RFR).

Given the basic regression tree is bigger, complex and difficult to explain, in PLS-RFR, the basic regression tree isn't as individual learner any longer but PMT instead. The main idea is as followed: first, adopting the Bootstrap sampling to copy the learning training set, generating lots of distinct versions of the sub-learning training set; then, building lots of distinct versions of the sub-learner to go along with the sub-learning training set, namely, Partial Least Squares Model Tree; last, deciding the final predicted result by the mean. Because of the weak learning ability of PMT, the ensemble learning can enhance the generalization of PLS-RFR. Hence, PMT is a kind of localized nonlinear modeling method for complex data. The main algorithm is as follows:

Algorithm 2 : Random Forest of Partial Least Squares

Dataset (D): The original sample data, the sample size is m
AttributeList: Independent variable attribute list, the number is *mall*
nTree: The total number of individual learner
E: The self-help sampling frequency
F: The size of the random input vector
NIPALSRFRegress(Dataset, AttributeList, *nTree*, *E*, *F*)
for i=1 to *nTree*
  Bootstrap random sampling E times
  Get a self-help training set d_i
  Construct the random input vector of F on the AttributeList
  Build PMT on the d_i and F
end

$$R(x) = \arg mean \sum_{i=1}^{nTree} PMT_i$$

The algorithm 2 get improved based on PMT and RFR. Considering the features of RFR, besides the two parameters of PMT, there are another three parameters involved and parameters recommendation is the same as the traditional RFR [12]: one is the random input vector F, F generally sets to $\sqrt{q}$ or $\text{int}(1+\log_2 M)$, here, int represents the rounded down among them. Second is the sampling frequency of Bootstrap E, it is generally recommended as the capacity of origin sample data set m. Third, it is the number of individual learner *ntree*, which is decided by the change of residual by increasing the number of the random tree. With the increasing of the size of the tree, the residual will be stable. At the beginning, *ntree* sets 100 in the training process of modeling. The three parameters still take example by the recommended value of RFR. Since the PLS-RFR is based on the PMT, after stretching PMT to analyze the data with multi-independent and multi-dependent variable, PLS-RFR also supports multi-independent and multi-dependent variable analysis. The flow graph of PLS-RFR is as followed:

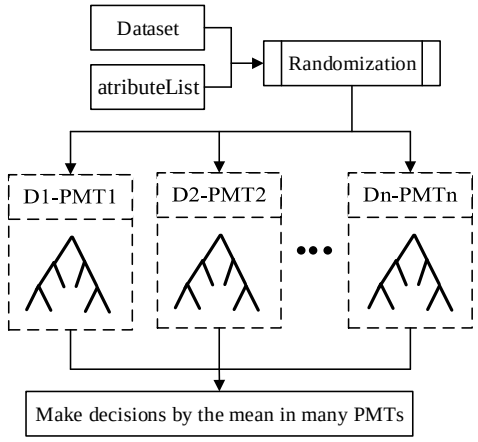

FIGURE II. THE ALGORITHM FLOW GRAPH ABOUT PARTIAL LEAST SQUARES OF RANDOM FORESTS REGRESSION

In Figure II, after constructing ITB and the random input vector on the original sample data by using the bootstrap random sampling, lots of distinct versions of PMT are built, the final predicted result is decided by the average.

IV. EXPERIMENTAL ANALYSIS

To observe the regression model effect about PMT and PLS-RFR, it adopt the traditional Multiple Regression (MR), the traditional PLSR and the traditional Random Forests Regression (RFR) as a contrast. Based on this, it divides the original data randomly into the proportion of 7:3, the part of 70% is the training set to build the model, and the remaining is the testing set to validate the model. There are five indicators for investigation, namely, Sum of Squares for Error of Train (SSETrain), Sum of Squares for Error of Test (SSETest), coefficient of determination (R-Square) and time consuming (Time, Unit: ms). In order to facilitate the experiment, verifying the effect of the model and establishing enough testing set for contrast, it defines that the original sample data obtained here is greater than or close to 50. Here, it adopts the experimental data of MaXingShiGan decoction of the monarch drug to treat the asthma or cough and five sample sets in the UCI Machine Learning Repository [13] to compare.

A. The Experiment of MaXingShiGan Decoction of the Monarch Drug to Treat the Asthma

TABLE I shows the part of experiment data of MaXingShiGan decoction of the monarch drug to treat the asthma for rats in the key laboratory of Modern Preparation of TCM, Ministry of Education in Jiangxi University of Traditional Chinese medicine, 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, namely, ephedrine, pseudoephedrine, methyl ephedrine, wild black cherry glycosides and licorice glycosides. There are two pharmacological indicators namely, incubation period (Unit: s) and cough duration (Unit: min). Based on this, there are five independent variables and two dependent variables, it divides the original data randomly into the proportion of 7:3, the part of 70% is the training set to build the model, and the remaining is the testing set to validate the model. Then, it compares the model result between PMT, PLS, RFR, MR and PLS-RFR by JAVA programming. The experiment results are shown in TABLE II.

TABLE I. THE EXPERIMENTAL DATA OF THE MAXINGSHIGAN DECOCTION OF THE MONARCH DRUG ABOUT THE TREATMENT OF THE ASTHMA CHANGING THE EPHEDRA DOSE

ephe-drine	pseudo-phedrine	methyl ephedrine	wild black cherry glycol-sides	licorice glycol-sides	incubation period (Unit: s)	cough duration (Unit: min)
1.04	0.47	0.19	0	0.47	68	28
0.95	0.53	0.17	1.67	0.48	44	22
1.99	1.92	0.17	0	0.18	108	10
1.09	0.43	0.41	0	0.42	71	19
2.31	1.89	0.08	0.121	0.21	150	9
2.96	2.96	0.353	0.216	0.83	49	24
1.11	0.89	0.92	0	0.87	54	9
3.41	2.85	0.293	0	0.9	80	7
5.59	4.92	0.4363	0	0.89	52	17
5.08	5.22	0.4063	0.977	0.88	80	12
……	……	……	……	……	……	……

TABLE II. COMPARISON OF FIVE METHODS IN THE EXPERIMENTAL DATA OF THE MAXINGSHIGAN DECOCTION OF THE MONARCH DRUG ABOUT THE TREATMENT OF THE ASTHMA

	Time	**R-Square**	**SSETrain**	**SSETest**
MR	2	0.2198	20193.6746	9848.1760
PLSR	32	0.1483	20537.1629	9910.0989
PMT	8	0.4262	14851.2187	9535.6027
RFR	140	0.3474	16892.1152	8921.9143
RFR-PLS	142	0.5159	12529.7715	8714.3140

In TABLE II, for the Maxingshigan Decoction of the monarch drug about the treatment of the asthma of 2 independent variables, PLSR don't show a good expected effect, the explanation level reached only 14.83%, RFR is only 34.74%, and PMT is 42.62%. However, the improved PLS-RFR is up to 51.59% in the ensemble learning, Sum of Squares for Error has a certain degree of decline in the training set and testing set.

B. The Experiment of MaXingShiGan Decoction of the Monarch Drug to Treat the Cough

TABLE III shows the part of experiment data of MaXingShiGan decoction of the monarch drug to treat the cough for rats in the key laboratory of Modern Preparation of TCM, Ministry of Education in Jiangxi University of Traditional Chinese Medicine, a total of 63 sample data, it is about the impact of pharmacological indicators about the blood medicine composition in rats under 10 distinct dosage of almonds respectively. There are five compositions about the blood medicine composition in rats, namely, ephedrine, pseudoephedrine, methyl ephedrine, wild black cherry glycosides and laetrile. There are one pharmacological

indicators namely, cough times (Unit: times). Based on this, there are five independent variables and one dependent variables, it divides the original data randomly into the proportion of 7:3, the part of 70% is the training set to build the model, and the remaining is the testing set to validate the model. Then, it compares the model result between PMT, PLS, RFR, MR and PLS-RFR by JAVA programming. The experiment results are shown in TABLE IV.

TABLE III. THE EXPERIMENTAL DATA OF THE MAXINGSHIGAN DECOCTION OF THE MONARCH DRUG ABOUT THE TREATMENT OF THE COUGH CHANGING THE ALMOND DOSE

ephedrine	pseudo-ephedrine	methyl ephedrine	laetrile	wild black cherry glycosides	cough times (Unit: times)
3.74	3.78	0.52	33.2	24.4	41
373.59	217.94	16.38	0.629	2.24	58
402	369.93	48.46	0.785	1.87	25
9	9.35	0.54	16.6	16.6	39
18.3083	21.5	1.08	4.81	40.1	22
42.532	42.6	2.8383	2.3	8.32	47
2.25	2.975	0.493	40	40	42
20.1733	19.95	0.98	4.77	10.2	23
43.296	37.6	2.5483	2.852	9.45	45
21.7417	24.36	1.17	4.62	26	38
……	……	……	……	……	……

TABLE IV. THE RESULT COMPARISON OF FIVE METHODS IN THE EXPERIMENTAL DATA OF THE MAXINGSHIGAN DECOCTION OF THE MONARCH DRUG ABOUT THE TREATMENT OF THE COUGH

	Time	**R-Square**	**SSETrain**	**SSETest**
MR	2	0.1203	3468.3287	1622.4532
PLSR	31	0.0587	3710.8791	1533.7792
PMT	16	0.6671	1312.3766	1524.6328
RFR	109	0.0975	3558.0023	1777.0677
RFR-PLS	124	0.3433	2588.8972	1553.9246

In TABLE IV, for the Maxingshigan Decoction of the monarch drug about the treatment of the cough of 1 independent variables, PMT and PLS-RFR is nearly up to 66.71% and 34.33% in the explanation level. PMT drop to 1312.3766 and 1524.6328 about Sum of Squares for Error on the training set and testing set. Simultaneously, PLS-RFR falls to 2588.8972 and 1553.9246.

C. UCI Machine Learning Repository

To validate the feasibility and effectiveness of PMT and PLSR, it adopts five sample sets in the UCI Machine Learning Repository. TABLE V shows the basic information about five machine learning data set.

TABLE V. THE SAMPLE SETS IN THE UCI MACHINE LEARNING REPOSITORY

Dataset	Abbreviation	The number of independent variable	The number of dependent variable	The training set	The testing set
Concrete Compressive Strength	Concrete	8	1	721	309
Slump	Slump	7	3	72	31
Yacht Hydrodynamics	Yacht	6	1	216	92
Housing	Housing	13	1	354	152
Airfoil Self-Noise	Airfoil	5	1	1052	451

In the process of experiment, making the model to be optimal by adjusting the parameter of model, it compares the model result under the condition of same parameters. First, it divides the original data randomly into the proportion of 7:3, the part of 70% is the training set to build the model, and the remaining is the testing set to validate the model. Then, it compares the quality and effect under the same learning training set in several ways. Last, it uses JAVA to accomplish this process. The experimental results are shown in TABLE VI-X.

TABLE VI. THE RESULT COMPARISON OF FIVE METHODS IN THE CONCRETE COMPRESSIVE STRENGTH DATASET

	Time	R-Square	SSETrain	SSETest
MR	7	0.5824	78433.8274	33045.0562
PLSR	50	0.5739	80026.3124	33655.5497
PMT	281	0.905	17835.0102	14541.2386
RFR	338	0.3478	122501.7651	64488.188
RFR-PLS	385	0.7496	47029.7116	20739.7566

TABLE VII. THE RESULT COMPARISON OF FIVE METHODS IN THE SLUMP DATASET

	Time	R-Square	SSETrain	SSETest
MR	2	0.4656	15432.9726	7212.6384
PLSR	47	0.5891	15594.1733	7729.8852
PMT	15	0.7925	6918.391	7554.5867
RFR	171	0.285	20649.4744	11250.2986
RFR-PLS	344	0.6173	11052.1156	6536.1835

TABLE VIII. THE RESULT COMPARISON OF FIVE METHODS IN THE YACHT HYDRODYNAMICS DATASET

	Time	R-Square	SSETrain	SSETest
MR	2	0.6587	16076.1336	8159.7909
PLSR	41	0.6586	16076.4914	8151.2763
PMT	14	0.9967	155.7321	266.439
RFR	107	0.8151	8706.8571	4900.1754
RFR-PLS	168	0.987	614.3444	560.7723

TABLE IX. THE RESULT COMPARISON OF FIVE METHODS IN THE HOUSING DATASET

	Time	R-Square	SSETrain	SSETest
MR	1	0.8733	3184.3654	83131.7422
PLSR	48	0.8733	3184.4137	81262.7214
PMT	198	0.9469	1333.597	70160.8003
RFR	439	0.6859	7892.3088	15598.435
RFR-PLS	550	0.9115	2224.4969	8247.0806

TABLE X. THE RESULT COMPARISON OF FIVE METHODS IN THE AIRFOIL SELF-NOISE DATASET

	Time	R-Square	SSETrain	SSETest
MR	2	0.5026	24374.287	10311.6879
PLSR	50	0.5024	24381.2663	10289.1888
PMT	108	0.8979	5003.6441	3877.8539
RFR	283	0.2014	39131.9915	18286.096
RFR-PLS	249	0.6429	17497.5751	7568.6188

In TABLE VI-X. it will have a good visual effect shown by chart. For MR, PLSR and RFR, the explanation level gets improved about five sample sets of the UCI Machine Learning Repository by PMT and PLS-RFR. Simultaneously, SSETrain and SSETest have a certain degree of decline. PLS-RFR gets obviously than PMT about the generalization. It shows that PMT and PLS-RFR have a certain degree of validity and feasibility.

V. CONCLUDING REMARKS

Through the above analysis, we can obtain the following conclusions:

Firstly, For the problem that Partial Least Squares Regression is lack of theoretical guidance of rules by the quasi linearization rule to achieve the nonlinear, and the accuracy declines in face of the unknown variables distribution, and the information is easy to loss about the mean treatment of the leaf in the Regression Tree and the Random Forest Regression, by adopting the PLSR at the leaf nodes, the information loss issue gets avoided when the leaf nodes is directly for mean with appropriate samples. Simultaneously, by applying the dividable principle of the attribute in Regression Tree, it makes several linear segments of Partial Least Squares approach unknown variable distribution to achieve the nonlinear regression.

Secondly, by Bootstrap sampling the origin data set, choosing randomly features in all attributes by the number, constructing various copies of the training dataset of distinct size, building distinct versions of PMT for ensemble learning, PLS-RFR is constructed, enhancing the generalization of PMT.

Thirdly, for the experiment data of Maxingshigan Decoction of the monarch drug about the treatment of the asthma or cough and five sample sets in the UCI Machine Learning Repository, the explanation level get increased after adopting the improved model, it shows that PMT and PLS-RFR have a certain degree of validity and feasibility.

Fourthly, it is able to pull PLS, PMT, PLS-RFR into the data of Traditional Chinese Medicine to provide better technical support.

ACKNOWLEDGMENT

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).

REFERENCES

[1] Z. Q. Jia, J. Y. Cai, Y. Y. Liu. Real-time performance reliability evaluation method of small-sample based on improved Boostrap and Bayesian Bootstrap. Journal of Computer Applications. Application Research of Computers. vol. 26, no. 08, pp. 2851-2854, 2009.

[2] J. R. Quinlan. Learning with continuous classes. Proceedings of the 5th Australian joint Conference on Artificial Intelligence. Singapore, 1992.

[3] H. J. Li, Z. X. Wang, L. M. Wang, et al. Bayesian-network based Regression Tree Learning Algorithm. Chinese Journal of Scientific Instrument. vol. z3, pp. 387-388, 2004.

[4] K. Zhang, Z. C. Mu, X. H. Chang, et al. On A New Model Tree Algorithm and Its Application . Control Engineering of China. vol. 15, no. 01, pp. 103-106, 2008.

[5] H. Abdi, L Williams. Partial Least Squares Methods: Partial Least Squares Correlation and Partial Least Square Regression. Reisfeld B, Mayeno A N, Humana Press, pp. 930, 549-579, 2013.

[6] H. W. Wang, Z. B. Wu, H. Meng, et al. Linear and nonlinear method of partial least squares regression. Beijing: National Defend In-dustry Press, 2006.

[7] Y. Liu. Response Surface Modeling by Local Kernel Partial Least Squares. Tsinghua University, 2013.

[8] Z. L. Hao, J. Q. Du, G. L. Wang, et al. Analysis of partial least squares method combining with SBM. Computer engineering and design. pp. 2896-2900, 2014.

[9] F. H. Shang, W. Zhang. Study on inferring interwell connectivity of injection-production system based on decision tree. Application Research of Computers. vol. 30, no. 07, pp. 2051-2054, 2013.

[10] J. X. Guo. Study on Improved High-Dimension and Nonlinear Partial Least-Squares Regression Method and Applications. Tianjin University, 2010.

[11] W. X. Xi, X. Y. Tang. Pedestrian detection based on random forest and support vector machine. Journal of Computer Applications. vol. 34, no. S2, pp. 283-285, 2014.

[12] Z. F. Cao. Study on optimization of random forests algorithm. Capital University of Economics and Business, 2014.

[13] UCI machine learning repository. 2014.