Analysis of TCM Data Based on Partial Least Squares within Random Forest

Fang YU¹, Jian-qiang DU¹*, Bin NIE¹, Zhu-lin HAO², Qing-xia ZENG¹ and Ri-yue YU³

¹School of Computer, Jiangxi University of Traditional Chinese Medicine, Nanchang, 330004, China
²Shanghai Chongyu Intelligence Information Technology Company, Shanghai 200082, China
³College of Pharmacy, Jiangxi University of Traditional Chinese Medicine, Nanchang, 330004, China
*Corresponding author

Keywords: Random forest, Partial least square, Nonlinear, TCM data.

Abstract. Partial Least Square (PLS) seems hard to adapt to the characteristics of the nonlinear data due to its own linear feature. However, Random Forest Algorithm (RFA), which is assembled by several classifiers, is adaptive and suitable to nonlinear data. Based on this, a new method fusing RF into PLS is proposed, which build Random Forest through the principal components and the dependent variable extracted from PLS, and use the residual information to build Random Forest recursively until accuracy conditions are met. Using the data of the maxingshigan decoction of the monarch drug to treat the asthma or cough and some datasets in the UCI Machine Learning Repository, the results show that the improved algorithm has a certain degree of correctness and validity.

Introduction

TCM [¹] treating diseases, taking the dose-effect relationship of drugs as an example, chemical drug has clear and fixed structure. For that drug can only treat one diseases, the concepts, principles, methods and applications for dose-effect relationship has formed a relatively perfect system, such as Linear Regression (LR), Classification and Regression Tree (CART), Support Vector Machine (SVM), and Artificial Neural Network (ANN) [²].

However, Traditional Chinese medicine herbs present features of more compositions, more targets and more efficacies deciding that the Traditional Chinese medicine data present multiple variables, multiple dependent variables and nonlinear characteristics [³]. In paper [⁴], a feature extraction algorithm based on AutoEncoder is proposed, the results show it can get features well representative the image. Nevertheless, the algorithm requires large-scale training data, TCM data is usually small sample size owing to experimental factors, so it’s not appropriate to use the very method for TCM data. It results in linearly non-separable for high dimensional data in SVM [⁵] when the kernel function is inappropriate. As for Linear Regression (LR), although the model is simple and easy to explain, it is hard to express the nonlinear characteristic of TCM data.

Partial Least Squares (PLS) [⁶,⁷] can well explain the data with the characteristic of multi-independent variables and multiple dependent variables, but the principal components extract from PLS is linear combination of the independent variables’ column vector, apparently, the model is still linear relationship essentially when using such principal component and dependent variables for Multiply Linear Regression (MLR), therefore, PLS will not present good effects for raw TCM data directly.

In 2001, Leo Breiman [⁸,⁹] proposed Random Forest Regression (RFR), which is built by multiple learners. It randomly selected samples and features to balance noise and to further improve the generalization of the classifier and had strong self-adaptability for data. Compared to traditional classifier, there are many advantages for RF, such as stronger robust and solid ability, better effect in
modeling building and nearly non-overfitting and its own well ability to analyze and predict nonlinear data.

Based on this, a new method that fusing Random Forest into PLS has been introduced, through experimental analysis, the results seem to outperform than PLS and RF in the correctness and effectiveness of the algorithm.

Algorithm Introduction
Partial Least Squares
Partial Least Squares can not only do regression for data with multiply-independent variables and multi-dependent variables, but also can build model when even on small-scale data \(^{[10]}\). The main idea of Partial Least Squares is as followed:
Given the independent variables set \(X=(x_1, x_2, \ldots, x_p)\) and dependent variables \(Y=(y_1, y_2, \ldots, y_q)\). \(t\), \(u\) is the linear weight combination respectively, and the two need to meet the conditions below.

1. They both two must carry the variation information to the maximum extent.
2. The correlation between the two is also as large as possible

\(t\) and \(u\) is the first principal component extracted from \(X, Y\). we use \(t\) and \(u\) for multiply linear regression, and judge the residual information satisfying the pre-defined conditions or not , if meets requirements , then terminate the calculating, otherwise, continue to extract the second principal component \(t_2\), \(u_2\) from the residual information till satisfying result is obtained.

Random Forest Regression Algorithm (RFRA)
Random Forest regression \(^{[11]}\) is adaptive and suitable for linear and nonlinear regression. It achieves excellent performance by an ensemble of multiple weak learning (simple regression tree). Each of the regression tree is built using a bootstrap sample of the data (the data not sampled is called OOB, as testing data), and in terms of attribute division, we randomly choose \(r_{-n}\) \((r_{-n}=\sqrt{p}\ or\ int(1+\log_2p)\) attributes instead of selecting all of the original feature attributes to build a tree.

The main idea of how to build a Random Forest has been described briefly in paper \(^{[12]}\).

Fusing Random Forest into Partial Least Squares Method
An Algorithm for Constructing Random Forest of Principal Components
A method of constructing Random Forest of Principal Components uses the PLS to extract principal components \(t\), and utilize Bootstrap to resample randomly from raw data to generate several sub-sample set \(d_i\) \((i=1,2,3\ldots n_{-C})\) (has the same size of \(t\)). Then build a regression tree by \(d_i\) and dependent variable \(Y\). Because \(d_i\) is continuous, we adopt the sum of square error minimum criterion to find the best splitting point, according to which divided \(d_i\) into two subsets.

Under the aforementioned pattern, continue to recursively divide these two subsets until the number of leaf nodes is less than predetermined threshold or the error is reduced un-apparently.

The process of the method of constructing Random Forest of Principal Components(CRFPC) is as follows:

<table>
<thead>
<tr>
<th>Algorithm 1</th>
<th>An Algorithm for Constructing Random Forest based on Principal Components</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input</strong>: principal components: (t); sampling times: (r_{-m}); Total number of learning machine: (n_{-C}), dependent variable attributes set: (attributeListY)</td>
<td></td>
</tr>
<tr>
<td><strong>Output</strong>: the built Learner: (RTree) for (i=1) to (n_{-C})</td>
<td></td>
</tr>
</tbody>
</table>
Step01 using Bootstrap to resample for \( r \times m \) times from \( t_i \), to get the sample set \( d_i \), (Given we divide \( t_i \) into \( K \) cells \((R_1, R_2, \cdots, R_K)\), and each of the units there is a fixed output value as \( c_i \) )

Step02 building the basic CART \((RTree(x) = \sum_{i=1}^{K} c_i I(x \in R_i))\) by \( d_i \) and \( attributeListY \)

Step03 looking for the splitting attributes
adopt the sum of the squared errors minimum criterion to search for the splitting attribute node of the subtree (including the current node) for the internal node in \( RTree_i \), then merge all of the splitting attributes. The merged set is called regression attributes

Step04 handling for the internal node
Choose the part of or all regression attributes of the sample data in the current node to execute regression operation, and traversal all the regression model further, we pick the model that owns the sum of the squared errors least as the regression model of current node

end

return \( RTree = \sum_{i=1}^{n} RTess_i / n \times C \)

---

**Fusing Random Forest into Partial Least Squares (RFR-PLS)**

The main steps of RF-PLS are as follows:

Extract first principle component information from \( X, Y \) as \( t_1, u_1 \), then, we make the \( X \) and \( t_1, Y \) and \( u_1 \) execute Random Forest regression, and judge the residual information further, if the residual accuracy satisfy the meets, then terminate the process, else continue to extract principle component information from the residual information of regression result of \( X \) and \( t_1, Y \) and \( u_1 \), the above procedure continues until satisfactory accuracy is achieved. Finally, we extract the \( m \) principal component \( \{t_1, \cdots, t_m\} \) from \( X \), and utilize the method of Random Forest Embedded into Partial Least Squares Regression to conduct regression of \( y_k \) for \( t_1, \cdots, t_m \), furthermore, we get the expression of \( y_k \) for raw independent variables \( \{x_1, \cdots, x_m\} \) (where \( k \in \{1, 2, \cdots, q\} \)

The Schematic diagram of RF-PLS is as shown in Figure 1. In the figure, \( X \) denotes the independent variable data matrix, \( Y \) is the observation variable data matrix, \( t_{(i)} \) (the value i begins with 1) denotes the main components of \( X, X_i, Y \) (the value i between 0 and n) are the residual matrix of \( X \) and \( Y \) respectively. \( \hat{Y}_{(i)} \) is the predictive value (the value i begins with 0). The evaluation criteria, which adopts the residual sum of squares \( (R_{RF}^2) \) of predictive value in OOB data, is as followed

\[
MSE_{OOB} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_{i,OOB})^2
\]

\[
R_{RF}^2 = 1 - \frac{MSE_{OOB}}{\sigma_Y^2}
\]

\( Y_i \) is the real value of dependent variables in OOB, \( \hat{Y}_{i,OOB} \) is the predictive value of OOB data with RF method, \( \sigma_Y^2 \) is the variance of OOB data with RF method.
The main algorithm is below:

**Algorithm 2. Fusing Random Forest into Partial Least Squares**

**Input**: raw data set: Dataset(D); independent variables attributes list: attributeListX, dimension: p; dependent variables attributes list: attributeListY, dimension: q; total number of Learner: n_C; resample times: r_m; size of random input vector: r_n

**Output**: the predictive value R(X)

Step01 Extract the attributeListX, attributeListY as (X, Y) from the original data, and normalize the (X, Y) as (E_X, E_Y)

Step02 partial least squares regression(PLSR)

- For 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 \( F_i E_i E_i E_i F_i \) and \( E_i E_i E_i E_i F_i \)
    - Calculate the score vector \( t_i = E_i w_i \) and get the RF output values
    - \( E_e = RF (t_i, n_C, r_m, r_n, E_i) \)
    - \( F_e = RF (t_i, n_C, r_m, r_n, F_i) \)
    - Regression equation \( E_{i+1} = E_i + E_e \) and \( F_{i+1} = F_i + F_e \)
    - Get the residual information matrix \( E_i \) and \( F_i \)
  - i = i + 1

Step03

Integration the equation of RF-PLS

\[ F_e(t) = \sum_{i=1}^n \sum_{j=1}^{n_C} RTree_i / n_C, t = \{1, 2, \ldots, n\} \]

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

Step04 end

The **RF (X, n_C, r_m, r_n, Y)** mentioned above is:

**Algorithm3: Random_Forest**

**RF (X, n_C, r_m, r_n, Y)**

for i=1 to n_C
  - using Bootstrap to resample from D for r_m times to get a sub-training set d_i
  - randomly select input vectors of size r_n without replacement in the independent attributes set of d_i, to form a new subset d_i
  - build a learner RTree_i by dataset(d_i, Y)
end

return \( R(x) = \arg\max \sum_{i=1}^{n_C} \text{RTtree}_i \)
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. It still choose another two sample sets in the UCI Machine Learning Repository, namely Slump and CCPP_Folds5x2_pp[14] to testify the improved algorithm.

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 and two pharmacological indicators namely, incubation period (Unit: s) and cough duration (Unit: min). The first five compositions are the independent variable, the rest two is dependent variable.

As for the data of maxingshigan decoction of the monarch drug to treat the cough(MXSGZK), it has the same independent variables with MXSGPC, the difference is that the dependent variable is only one target - cough duration. That’s why we didn’t show the data in this paper.

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

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

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

Analysis of the Procedure and Result of the Experimental

In order to verify the effectiveness of the improved algorithm, we adopt several traditional methods to compare in four datasets showed in 3.1. In this paper, it randomly divided the raw data with the ratio of 7: 3, the 70% as the training dataset, the remained as the test sample.

For the data of the maxingshigan decoction of the monarch drug to treat the asthma (MXSGPC), there were 46 samples, 5 independent variables, 2 dependent variables, 32 training samples and 14 test sets. As for the data of maxingshigan decoction of the monarch drug to treat the cough(MXSGZK), there were 63 samples, 5 independent variables, 1 dependent variables, 44 training samples and 19 test sets.

For Slump, it has 103 samples, 7 independent variables, 3 dependent variables, 72 training samples and 31 test sets. CCPP_Folds5x2_pp has a total of 9568 samples, 4 independent variables, 1 dependent variables, 6698 training samples and2870 test sets.

We adopt the (sum of Squares for Error of Train (SSETrain) and sum of Squares for Error of Test (SSETest) as the observation target, the results are as show below.
We compared our improved algorithm (RFR-PLS) to the other three methods, which is the classic Partial Least Squares (PLS), Random Forest (RF) and the method of Fusing Model Tree into Partial Least Squares (MT-PLS). The results in the Figure 2 show: the improved algorithm has obvious applicability on the four datasets.

To make the data fluctuation clear, we centralizer the SSETrain and SSTest to the same level,

Table 2. The different algorithm’s SSTrain and SSTest under each data set.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MXSGPC</th>
<th>MXSGZK</th>
<th>CCPP_Folds5x2_pp</th>
<th>Slump</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS</td>
<td>20580.6507</td>
<td>3841.8357</td>
<td>161316.0535</td>
<td>15994.1733</td>
</tr>
<tr>
<td>RFR-PLS</td>
<td>7265.4268</td>
<td>781.8543</td>
<td>141386.5010</td>
<td>8785.6886</td>
</tr>
<tr>
<td>RFR</td>
<td>2184.2919</td>
<td>2184.2919</td>
<td>145551.7916</td>
<td>20649.4744</td>
</tr>
<tr>
<td>MT-PLS</td>
<td>7485.3271</td>
<td>1568.5878</td>
<td>128136.5498</td>
<td>8888.8669</td>
</tr>
<tr>
<td>SSETrain</td>
<td>30434.7553</td>
<td>1761.1864</td>
<td>70113.7642</td>
<td>7729.8852</td>
</tr>
<tr>
<td>SSETest</td>
<td>10855.0917</td>
<td>1544.3966</td>
<td>21114.3382</td>
<td>11250.2986</td>
</tr>
<tr>
<td></td>
<td>14618.4831</td>
<td>1653.9989</td>
<td>14618.4831</td>
<td>9629.1733</td>
</tr>
</tbody>
</table>

From the table 2 and Figure 2, compared to PLS and RFR, the SSTrain and SSTest of the improved algorithm has different degree of decreasing. Briefly to say:

(1) For MXSGPC, it manifests a distinct non-linear characteristic, that’s why has good performance in RFR, MT-PLS, and RFR-PLS. However, compared to the improved algorithm (RFR-PLS), the generalization ability of RFR and MT-PLS seems poor.

(2) For MXSGZK, RFR-PLS seems to show a clear advantage with respect to the explainable and generalization of the model than the other three prediction models.

(3) For CCPP_Folds5x2_pp and Slump, it is obvious that PLS can’t well explain the data. RF has a good result for CCPP_Folds5x2_pp, but the effectiveness in Slump is disappointing. However, when we combine this two algorithms, the SSTrain and SSTest has different degree of declining, as well as the MT-PLS. But the generalization of MT-PLS is a little poor than RFR-PLS on CCPP_Folds5x2_pp, which validates Random Forest’ generalization ability is stronger than Model Tree.

All in all, for different data from different domains or with different size, the improved algorithm outperforms than traditional PLS, RF and even MT-PLS in terms of the sum of Squares for Error of Train and the sum of Squares for Error of Train.
The Algorithm’s Time Complexity

For the PLS, the time complexity is mainly in the process of 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 Random Forest, the time complexity is mainly reflected in the part of the tree built and the number of trees, given the number of trees is $k$, Random Forest’s time complexity is $O(kn^2)$. For the improved algorithm, it is assumed that the number of the principal components extracted is $m$, and one tree is built when extracting a principal component, so the time complexity of the improved algorithm is $O(kmn^2)$ $k, m$ are the constant.

Conclusion

Focus on the unfitting of PLS for the nonlinear data, a new method fusing Random Forest into PLS is proposed, which can ensemble multi-learning machine and make full use of the random selecting of feature and sample, the new algorithm can well explain the nonlinear data and has a strong adaptive ability. For experiments on several TCM data and two regression data sets in the UCI Machine Learning Repository, the result show: the improved algorithm has better generalization accuracy and stronger classification prediction ability, and has a certain practical significance for the guidance of clinical medication. However, the number of the weak Learners directly decide the robustness of the model, what we do next is to find suitable learner’s number.

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). Jiangxi Province key research and development program key projects (20171ACE50021). Jiangxi Province undergraduate colleges and universities in the young teachers’ visiting fund.

References


http://archive.ics.uci.edu/ml/datasets/Concrete+Slump+Test