Integration of ANN and Statistical Method for Outlier Detection in Complex System

Weixiang Zhao, Lide Wu
Computer Science Department,
Fudan University,
Shanghai 200433, P. R. China
Email: weixiang_zhao@263.net

Abstract
In this paper, an outlier detection method based on radial basis functions–principal component analysis (RBF-PCA) approach and Prescott method, a statistical detection approach, is proposed to detect the outlier in the complex system without clear mechanisms. Making full use of the capacity of neural networks on nonlinear mapping and the effect of Prescott method on outlier detection in linear model, the integration of two approaches makes the outlier detection in complex nonlinear system more convenient, reliable, and precise. The experiments show us the satisfactory effects of the proposed method and its superiority over some distances based methods. Furthermore, some rules were discussed for the wide use of the proposed integrated method.

Keywords: outlier, radial basis function networks, principal component analysis, statistics

1. Introduction
The measurement of the process variables often contain outliers, which have a large negative impact on the accuracy and reliability of the models, so how to detect the outliers in the sampling data has been given more and more consideration. Among the statistical detection methods, which have been widely applied due to the relatively mature mathematical theories [1][2][3], an approximate test for outlier [1], Prescott method, showed its conciseness and effectiveness on detection. But the successes of these mentioned methods must have two prerequisites that 1) the analyzed object could be depicted by linear model and 2) the model structure should be known beforehand. However, in many cases, the object is nonlinear and even has no clear interior mechanism, so some methods based on the distances between patterns have been proposed to detect the outlier in the complex system [4][5]. Because distance is not the best criterion for outlier verification, in this paper an integrated outlier detection method suitable for the complex system without clear mechanisms is proposed to overcome the drawbacks of some distances based methods and to extend the application fields of Prescott method.

2. Distances based detection method
In this section, a typical distances based method [4] is shown as below.

1) Let the pattern be $s_i$ ($i=1,2,...N$). Calculate the mean minimal distance (MMD) of all patterns.

$$d_m = \frac{1}{N} \sum_{i=1}^{N} \min_{j \neq i} \left[ \sum_{q=1}^{d} \left( \frac{s_{iq} - s_{jq}}{\nu_{eq}} \right)^2 \right]^\frac{1}{2}$$

......(1)

where $N$ is the patterns number, $d$ is the dimensions number, and $\nu_{eq}$ is the $q$th diagonal element of the covariance matrix of all patterns.

2) Calculate the closeness degree of each pattern.

$$w_{gi} = \frac{d_m}{\text{Dist}_i}$$

......(2)

where $w_{gi}$ is the closeness degree of the $i$th pattern, and $\text{Dist}_i$ is the minimal distance between the $i$th pattern and any other patterns.

3) Make judgement according to the closeness degrees and some rules.
In general, the pattern with smaller closeness degree is more likely to be considered as outlier. But, it is obvious that this is not widely feasible because in some cases distances can not represent the essence of system.
3. An approximate test for outlier in linear model

Prescott method has demonstrated its convenience and effectiveness on outlier detection because the test procedure is applicable to any linear model and does not require a re-analysis with the suspected outlier omitted or treated as a missing value. The kernel part of Prescott method is briefly shown as below [1].

The usual form of a linear model is

\[ y = X\beta + \epsilon \]  

.....(3)

where \( y \) is the \( n \times 1 \) vector of observations, \( X \) an \( n \times q \) matrix of known constants, \( \beta \) a \( q \times 1 \) vector of parameters and \( \epsilon \) an \( n \times 1 \) vector of normally distributed errors. The least squares estimate of \( \beta \) is given by

\[ \hat{\beta} = (X^TX)^{-1}X^Ty \] 

and the vector of residual is

\[ \epsilon = y - X\hat{\beta} = (I - X(X^TX)^{-1}X^T)\epsilon \]

Assuming that \( E(\epsilon) = 0 \) and \( Var(\epsilon) = \sigma^2I \), it follows that \( E(\epsilon) = 0 \) and

\[ Var(\epsilon) = (I - X(X^TX)^{-1}X^T)\sigma^2 \]

If \( \sigma^2 \) is estimated using \( u^2 = \epsilon^T\epsilon/(n-q) \), then the estimated variance-covariance matrix of the residuals is

\[ \hat{Var}(\epsilon) = (I - X(X^TX)^{-1}X^T)u^2 \]

The estimated standard deviation of the \( i \)th residual \( e_i \) is \( s_i \) where \( s_i^2 \) is the \( i \)th diagonal element of \( \hat{Var}(\epsilon) \).

It is suggested that if the variances of the residuals vary a lot, it is reasonable to use \( e_i/s_i \) as the “normal deviate form” of the residuals. Compared with \( e_i \), \( e_i/s_i \) can be considered as standardized residual. Computation of the standard residual is an easy matter and it would be a simple precaution to examine \( e_i/s_i \) rather than \( e_i \) as a matter of routine particularly when considering the possibility of discarding outliers. Finally Prescott proposed an approximate test for the outlier in the linear regression, in which the \( i \)th pattern can be considered as outlier, if \( |e_i/s_i| \) larger than the critical value. Except in the extreme cases, the adequate critical values [1] for max \( |e_i/s_i| \) could be given by.

\[ d_c = [(n - q)F/(n - q - 1 + F)]^{1/2} \]

where \( F \) is the \( 100(1 - \alpha/n) \) percentage point of the \( F \) distribution with 1 and \( n-q-1 \) degrees of freedom, and \( \alpha \) is the significance level. For convenience, the tables for the critical value \( d_c (\alpha, n, q) \) were presented [6].

The above approximate test method with the critical value tables has been applied successfully to detect the outlier in the linear model for the relation between phosphorus content of the corn and the inorganic and organic phosphorus contents in the soil [6].

In fact, there are a lot of highly nonlinear systems without clear mechanisms. Although the presented approximate test for outlier detection is concise and effective, it is only suitable for linear model, so in next section an integrated outlier detection method would be proposed to extend the uses range of Prescott method to the system without known mechanisms.

4. Outlier detection based on nonlinear transformation and Prescott method

It is obvious that Prescott method for outlier is effective in linear model, so if the relation between the input and output data could become linear after some transformation, Prescott method could be easily applied into nonlinear systems.

As we know, radial basis functions networks (RBFN) have strong function mapping ability. The hidden layer of RBFN performs a fixed nonlinear transformation with no adjustable parameters and it
maps the input space to a transition space, while the output layer implements a linear combiner on the transition space and the only adjustable parameters are the weights of this linear combiner [7][8][9]. The widely used method to acquire the weights of the linear combiner is linear regression, therefore RBFN could be used to detect the outlier in the system without clear mechanism if Prescott method is applied in the step of linear regression of RBFN. The flowchart of this new detection procedure is shown as Fig. 1.

Because the conventional RBFN based on K-Means algorithm has the following drawbacks: (1) the calculation of radial basis vectors easily falls into local optima; (2) the ill-conditioned regression matrix makes it difficult to calculate the weights of output layer when all patterns, some of which are very close to each other, must be used as radial basis vectors, this paper apply the principal component analysis (PCA) based structure determination strategy to make the training of RBFN more robust.

In RBF-PCA approach, each pattern was considered as a radial basis vector, thus the hidden layer (radial basis layer) of RBFN contains all information of the patterns, but this might result in overfitting. So PCA which has good capacity on data compression, feature extraction, and multicollinerity elimination [10][11][12] is used to form comprehensive radial bases. Compared with the radial basis in the hidden layer, the comprehensive one can be called principal radial basis (PRB), the corresponding layer can be called principal radial basis layer (PRBL), and this kind of networks can be called PRBFN. The structure of PRBFN is shown in Fig. 2.

Suppose the input and output data are \{x_i\} and \{y_i\} (i=1,2,...,N, N is the patterns number), thus RBF-PCA approach can be shown as below.

1). Consider the pattern \(x_i\) as the \(i\)th radial basis vector, and calculate the output matrix of the hidden layer \(A(N\times N)\), each element of which can be expressed by

\[
a_{ji} = \exp\left(-\frac{||x_i - x_j||^2}{\sigma^2}ight) \quad \cdots \quad (4)
\]

where \(\sigma\) is the width of \(i\)th center \((i, j=1,2,\ldots, N)\).

2). Apply PCA to form \(k\) PRB \(\{p_1, p_2, \ldots, p_k\}\), and the output matrix of PRBL is \(B=AP, P=\{p_1, p_2, \ldots, p_k\}\).

3) Calculate the weights of output layer through linear regression, i.e.,

\[
y=Bv+e
\]

where, \(v\) is the weights of output layer. Note: here let \(\sigma\) in eq. (4) be 1.

Now if Prescott method is used in the above linear regression, the outlier can be detected. The proposed integrated detection procedure can be
shown as below.

1). Make nonlinear transformation on the input matrix $X$ according to eq. (4).

2). Perform PCA on the output matrix $A$ to form the output matrix of PRBL, $B$.

3). Detect the outlier with Prescott method during the linear regression between $B$ and $y$.

The above is the proposed outlier detection method. Since RBF-PCA approach, in which PCA makes the RBFN training procedure more robust, is well suited to model the system with complex or unclear mechanism and Prescott method is convenient and effective on detecting the outlier in linear model, the proposed detection method based on these two techniques would have the satisfactory ability to detect the outlier in the system without clear mechanism.

5. Illustrative examples

In this section, a simulation experiment and a real example were performed to demonstrate the satisfactory effects of the proposed outlier detection method.

5.1 Simulation experiment

Let us consider the simulated system

$$z = xy + 2 \cdot \sin(x y)$$

Generated 20 pair of random real numbers in the interval $[0, 2]$ for $x$ and $y$, calculated the corresponding values of $z$, multiplied the last $z$ by 0.6 and considered it as an outlier.

First, the typical distances based method in section 2 was used to detect the outlier in these 20 patterns. The closeness degree of each pattern was listed in Table 1.

However, the closeness degrees of No. 2, 10, 11, 13 are smaller than that of outlier. Obviously, it will bring troubles to the further judgement based on these closeness degrees. This experiment showed us the drawbacks of some distances based methods.

Then the proposed integrated method was used to verify its ability to detect outlier. Determined the significance level $\alpha$ as 0.01, and the PRB number as 6. The results of the experiment were listed in Table 2, and the useful part of Lund’s table [6] was listed in Table 3.

<table>
<thead>
<tr>
<th>No.</th>
<th>$x$</th>
<th>$y$</th>
<th>$z$</th>
<th>$S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.6762</td>
<td>0.9931</td>
<td>3.6559</td>
<td>0.8164</td>
</tr>
<tr>
<td>2</td>
<td>0.0393</td>
<td>1.7995</td>
<td>0.2119</td>
<td>-1.5399</td>
</tr>
<tr>
<td>3</td>
<td>1.3626</td>
<td>1.6433</td>
<td>3.8089</td>
<td>0.5536</td>
</tr>
<tr>
<td>4</td>
<td>0.7590</td>
<td>1.2898</td>
<td>2.6387</td>
<td>0.1905</td>
</tr>
<tr>
<td>5</td>
<td>1.6636</td>
<td>1.6359</td>
<td>3.5371</td>
<td>-0.2649</td>
</tr>
<tr>
<td>6</td>
<td>1.0056</td>
<td>3.2052</td>
<td>3.2692</td>
<td>0.3033</td>
</tr>
<tr>
<td>7</td>
<td>1.4189</td>
<td>0.6839</td>
<td>2.6208</td>
<td>-0.1491</td>
</tr>
<tr>
<td>8</td>
<td>0.8578</td>
<td>0.5795</td>
<td>1.4507</td>
<td>-0.0079</td>
</tr>
<tr>
<td>9</td>
<td>0.6092</td>
<td>0.6824</td>
<td>1.2235</td>
<td>0.3376</td>
</tr>
<tr>
<td>10</td>
<td>0.3793</td>
<td>0.0682</td>
<td>1.1935</td>
<td>-0.1375</td>
</tr>
<tr>
<td>11</td>
<td>0.3869</td>
<td>1.4542</td>
<td>1.6293</td>
<td>0.0496</td>
</tr>
<tr>
<td>12</td>
<td>1.3644</td>
<td>0.6186</td>
<td>2.3387</td>
<td>-0.3777</td>
</tr>
<tr>
<td>13</td>
<td>0.6055</td>
<td>1.6770</td>
<td>1.1619</td>
<td>1.4898</td>
</tr>
<tr>
<td>14</td>
<td>1.0833</td>
<td>1.1361</td>
<td>3.1164</td>
<td>0.1055</td>
</tr>
<tr>
<td>15</td>
<td>0.3017</td>
<td>0.7408</td>
<td>0.6609</td>
<td>0.5248</td>
</tr>
<tr>
<td>16</td>
<td>0.3958</td>
<td>1.4055</td>
<td>3.8108</td>
<td>0.3185</td>
</tr>
<tr>
<td>17</td>
<td>0.7567</td>
<td>1.9391</td>
<td>2.2994</td>
<td>0.1122</td>
</tr>
<tr>
<td>18</td>
<td>1.7200</td>
<td>0.8898</td>
<td>3.5288</td>
<td>0.9097</td>
</tr>
<tr>
<td>19</td>
<td>1.7073</td>
<td>1.3891</td>
<td>3.7638</td>
<td>0.2617</td>
</tr>
<tr>
<td>20 (o)</td>
<td>1.1871</td>
<td>1.2426</td>
<td>2.0760</td>
<td>-3.3869</td>
</tr>
</tbody>
</table>

$x, y, z$: two inputs and the stipulated output respectively. $S$: the standardized residual.

Table 3. Lund Critical Value ($\alpha = 0.01$)

<table>
<thead>
<tr>
<th>n</th>
<th>q</th>
<th>2</th>
<th>3</th>
<th>6</th>
<th>10</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>3.06</td>
<td>3.04</td>
<td>2.95</td>
<td>2.76</td>
<td>2.20</td>
<td></td>
</tr>
</tbody>
</table>

n: the patterns number,
q: the independent variables number.

It could be seen from Tables 2 and 3 that the absolute standardized residual of the last pattern was 3.3869 much larger than the critical value, so the last pattern can be considered as outlier with the probability no less than 0.99. The detection result was identical with reality. Because the simulation system was nonlinear and the modeling procedure did not refer to any prior mechanisms or empirical models,
the result showed the satisfactory effect of the proposed method on the outlier detection in the system without known mechanisms. The reasonable integration of RBF-PCA approach and Prescott method extends the application fields of Prescott method from linear system to nonlinear system without clear mechanisms.

Besides, the number of PRB is a key to ensure the success of the proposed method. Table 4 shows us the different detection results of various PRB numbers.

<table>
<thead>
<tr>
<th>q</th>
<th>2</th>
<th>3</th>
<th>6</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.2290</td>
<td>0.2023</td>
<td>0.1107</td>
<td>0.0418</td>
</tr>
</tbody>
</table>

q: PRB number  A: average relative fitting error of normal patterns  
S: standardized residual of outlier

It can be seen from Table 4 that when the PRB number equals 2, the absolute standardized residual of the doubtful pattern is smaller than the critical value 3.06, which makes the method lose efficacy, but while the PRB number increases, the absolute standardized residual of doubtful pattern becomes larger than the corresponding critical value and the average relative fitting error of the 19 normal patterns decreases. This means that the larger PRB number can not only make verification of outlier but also increase the approximation effect on the normal patterns. However, to the contrary, the too larger PRB number will also lead the detection into fiasco. Suppose if the PRB number is 20, thus all patterns including outlier will be fitted without error, i.e. overfitting. Obviously, it is impossible to detect outlier in this condition. So only the proper PRB number, which can be determined by accumulated variance ratio, can guarantee the effect of the proposed detection method.

### 5.2 Real example

In this section, some data for studying quantitative structure-activity relationship (QSAR) of N-phenylacetamides[13] were collected to demonstrate the ability of the proposed method to detect the outlier in the complex nonlinear system. QSAR is a extremely complex map from the seven structure physico-chemical parameter, such as lipophilicity, molar refractivity and polar constant etc., to the biological activity. Unlike the simulation experiment in which we added the random noise on the dependent variable, we add a random noise on the independent variable, molar refractivity, of the first pattern and consider it as outlier (There are 20 patterns all together.).

Through the proposed integrated method, the standardized residual of the surprising pattern was 3.6607 when PRB number equaled 6. Compared with the critical value, the surprising pattern can be considered as an outlier with the probability no less than 0.99, which is identical with the real condition. The success of this experiment not only shows the capacity of the proposed method on the outlier detection in the real system without clear mechanisms, but also demonstrate the proposed method can detect the outlier no matter it lies in dependent or independent variables.

### 6. Conclusion

In this paper, a detection method integrating RBF-PCA approach with Prescott method was proposed to detect the outlier in the system without clear mechanisms. In the proposed method, PCA makes the RBFN training more robust, which gives full support to the further detection. The satisfying results of the simulation and real examples not only show us the superiority of the proposed method over some distances based methods, but also demonstrate that the integration of these two approach makes full use of the ability of neural networks to process complex system and successfully extend the application fields of Prescott method from linear system to the nonlinear system as well. In a word, the proposed integrated method affords us a concise and effective way to detect the outlier in the system without clear mechanism.

### References


