Data Processing of Deception Detection Based on PCA

ZiLong Chen[1], Dong-Liu[2], ZhiWei Gong[3], YunSheng Liu[4], and RuYu Li[5]

[1]Electrical Engineering College of Chongqing University, Chongqing, China
[2]Hong Shen College of Chongqing University, Chongqing, China
[3]Mechanical Engineering College of Chongqing University, Chongqing, China
[4]Hong Shen College of Chongqing University, Chongqing, China
[5]Civil Engineering College of Chongqing University, Chongqing, China

Keywords: Data processing; Principal component analysis; Deception detection

Abstract. Based on the analysis of the difficulty in data processing parameters caused by too many parameters in traditional deception detection, it is presented in this paper to use the method of principal component analysis (PCA) to reduce the number of parameters and the dimension of variables, thus facilitating data processing. Though specific experiments, it is found that the accuracy rate is 90.5% after using PCA, which is not quite different from that obtained before using this method. Moreover, the parameters highly correlated can be well eliminated with PCA, and the redundant data among variables can also be reduced. On the basis of retaining most data, the problems in original data can be effectively solved by using fewer variables, thus fulfilling the purpose of simplifying data and facilitating the analysis of problems. For deception detection experiments with large amounts of data, this is undoubtedly a feasible approach.

Introduction

Deception detection technology has been more and more widely used in the industries such as criminal investigation. With the development of sensor and computer technologies, the accuracy of deception detection has been increasingly high, and the methods become more and more advanced. The fourth-generation polygraph made by Dektor Counterintelligence and Security Cooperation of the United States allows simultaneous measurement of multiple signals, so as to provide accurate data for deception detection.

With the improvement of polygraph, there are more and more parameters for deception detection, and some polygraphs even have as many as 15 parameters. Though the increase of parameters helps improve the accuracy of deception detection, it will accordingly waste a lot of resources, which is also not conducive to the rapid processing of experimental data. In order to solve this problem, from the perspective of relevance, PCA is employed and fewer parameters are used to interpret most of the variation in data, thus cutting down the number of variables. This will facilitate the subsequent data processing, so as to improve the efficiency and accuracy of deception detection.

Major parameters determined through PCA

The idea of PCA is to turn more correlated variables into few independent or irrelevant parameters that are then used to reflect the experimental data. This method can be seen as dimension degradation of large amounts of data. The basic idea and method are as follows:

Determine index variables. Suppose there are q parameters for deception detection $x_1, x_2, \cdots x_q$, of which the values in No. i experiment are:

$$a_1, a_2, \cdots a_q, \quad i = 1,2,\cdots,n$$

As the units of the variables may be different, it may lead to largely different results in PCA. Therefore, in order to avoid the influence of dimension, the experimental data should be standardized. The standardization of data is shown as follows:

$$\hat{a}_{ij} = \frac{a_{ij} - u_j}{s_j}, \quad i = 1,2,\cdots,n; j = 1,2,\cdots,m$$

In the equation, sample mean value is $u_j = \frac{1}{n}, \ j = 1,2,\cdots,m,$ sample standard deviation is $s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_{ij} - u_j)^2}, \ j = 1,2,\cdots,m,$ and the index variable of standardization is defined as:

$$\hat{x}_j = \frac{x_j - u_j}{s_j}, \quad j = 1,2,\cdots,m \qquad ①$$

Definition of correlation coefficient matrix R. The correlation coefficient matrix correlation coefficient matrix $R = (r_{ij})_{m\times m}$ is defined. It mainly reflects the correlation degree between different parameters. The larger the value in a position of the correlation coefficient matrix is, the stronger the correlation degree between the two parameters that correspond to the position will be, and the more closely they will be related to each other.

$$r_{ij} = \frac{\sum_{k=1}^{n} \hat{a}_{ki} \cdot \hat{a}_{kj}}{n-1}, \quad i, j = 1,2,\cdots,m, \qquad ②$$

In the equation, $r_{ii} = 1, r_{ij} = r_{ji}$, r_{ij} is the correlation coefficient between indicators i and j.

Calculation of characteristic value and characteristic vector. Calculate the characteristic values $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_m \geq 0,$ and characteristic vectors of correlation coefficient matrix

$\mathbf{u}_1, \mathbf{u}_1, \cdots, \mathbf{u}_{\mathrm{m}}$, $\mathbf{u}_1 = \left[u_{1\mathrm{j}}, u_{2j}, \cdots, u_{\mathrm{mj}}\right]^T$ and thus new variables composed of characteristic vectors are obtained:

$$\begin{aligned} \mathrm{y}_1 &= u_{11}\widehat{x}_1 + u_{21}\widehat{x}_2 + \cdots + u_{\mathrm{m1}}\widehat{x}_m \\ \mathrm{y}_2 &= u_{12}\widehat{x}_1 + u_{22}\widehat{x}_2 + \cdots + u_{\mathrm{m2}}\widehat{x}_m \\ &\vdots \\ \mathrm{y}_m &= u_{1\mathrm{m}}\widehat{x}_1 + u_{2m}\widehat{x}_2 + \cdots + u_{\mathrm{mm}}\widehat{x}_m \end{aligned} \qquad ③$$

In the equation, y_1 is the first new parameter generated, y_m is the No. m new parameter,

Selection of new principal component and calculation of comprehensive evaluation value. Calculate the data contribution rate λ_i and cumulative contribution rate:

The data contribution rate is:

$$\mathrm{b}_j = \frac{\lambda_{\mathrm{j}}}{\sum_{k=1}^{m}\lambda_k}, \quad j = 1,2,\cdots,m, \qquad ④$$

The cumulative contribution rate is as follows:

$$\alpha_{\mathrm{p}} = \frac{\sum_{k=1}^{p}\lambda_k}{\sum_{k=1}^{m}\lambda_k}, \qquad ⑤$$

Through the cumulative contribution rate, we can decide what parameters can be used as new variables. Under normal conditions, when α_p is close to 1, the first p parameters can be selected to replace the original variables, basically with no information loss, which can be well used as principal component to make comprehensive analysis.

Carry out comprehensive analysis on the principle component:

$$Z = \sum_{j=1}^{p} b_j y_j, \qquad ⑥$$

In the equation, b_j is the information contribution rate of No. j principal component, and y_j is No. j principal component. Finally, through the size analysis of Z, it can be determined whether the subject is lying. When Z is larger than the specified threshold value, it means the subject is lying, and when Z is smaller than the specified threshold value, it means the subject is not lying.

Specific experimental analysis

Theoretical analysis of experimental parameters. Select heartbeat, breath, skin resistance, body temperature and pupils as the original polygraph variables. It can be seen that there are not few parameters for deception detection, and there may be redundant information among different variables. The sampling of parameters is carried out by using the self-made multi-purpose psychological polygraph. The error of polygraph accuracy is within the permitted range, so in the following analysis, the impact of polygraph errors can be ignored.

After the PCA of collected data is carried out by using Mallab, the characteristic root is obtained, contribution rate and cumulative contribution rate of different polygraph parameters, as shown in the following table:

Table 1 The contribution rates of different polygraph parameters

	Characteristic root	Contribution rate	Cumulative contribution rate
1	7.3033	34.0090	34.0090
2	4.4614	33.4409	67.4499
3	3.1435	31.3349	98.7848
4	2.1945	1.1001	99.8849
5	1.1475	0.1100	99.9949

It can be seen from the cumulative contribution rates in the Table, the sum of the cumulative contribution rates of the first three principal components is about 98.7848, which can well reflect the information in the data, so on the premise of guaranteeing accuracy, the first three principal components can be used as new parameters.

The characteristic vectors of the first three principal components are:

Table 2 The characteristic vectors of the first three principal components

	Heartbeat	Breath	Skin resistance	Body temperature	Pupils
1	0.331	0.343	0.329	0.102	0.013
2	0.216	0.199	0.201	0.303	0.204
3	0.314	0.041	0.113	0.141	0.406

It can be found from the data in the table above, the first principal component mainly reflects the indicators of heartbeat, breath and skin resistance; the second primarily reflects the temperature indicator; the third mainly reflects the indicator of pupils. The three principal components can comprehensively reflect the information of original parameters, and can be used as new parameters in the polygraph experiment.

The contribution rates of the three principal components above are used as weight, the comprehensive evaluation relation of deception detection is obtained as follows:

$$Z = 34.009y_1 + 33.4409y_2 + 31.3349y_3 \qquad ⑦$$

In this relation, y_i represents different principal components, respectively $i = 1,2,3.$

When the value of Z is larger than 0.9, the subject is regarded to be lying; otherwise, the subject is regarded to tell the truth.

Analysis of experimental accuracy. After the standard polygraph experiment on 150 subjects, the theoretical predicted results are obtained through the comprehensive evaluation formula, and then by interviewing the subjects, the real condition whether they have lied is revealed. If the predicted result reveals the real condition, it indicates successful deception detection; on the contrary, the deception detection fails. The final experimental results were as follows:

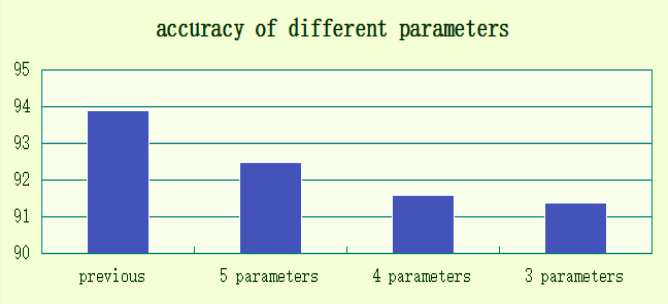

Fig. 1 The accuracy when different parameters are selected

It can be seen from the figure above that there is no large difference between the use of the three principal components and the use of the original parameters in terms of polygraph accuracy. However, it reduces the number of polygraph parameters from 5 to 3, and helps to reduce computational amount and the difficulty and complexity of data processing, thus facilitating the fast and effective polygraph experiments in practical application. This suggests that it is feasible and effective to use PCA in polygraph experiments with multiple experimental parameters.

References

[1] Si shoukui,Sun Xiqing, Application of mathematical model, National Defence industry press, Beijing, 2001.8.

[2] Fu Genyue,Chen Changkai,The development of traditional lie detection," Advances in Psychological Scinence, vol. 11(1) . 108-115, 2003.

[3] Li Yongxin,Li Yiming,Li Xinwang,"The synopsis of lie detection in china," Journal of Railway Police College, vol. 13(53),2003 .

[4] Hu Xiaoqing,Fu Gengyue,"Event-related Potentials in Deception Detection," China Journal of Clinical Psychology, vol. 17,No.2,2009 .

[5] Li Qiang,Fu Gengyue,“Deception Detection in Criminal Investigation,” Journal of Hunan Public Security College, vol. 21(1),No.2,2009 .

[6] Bai Liandong,“Deception Detection of Psychology in Criminal Investigation,” North China Coal Medical College, vol. 8(5),2006 .

[7] Lin Wenjuan,“The Overview of advanced psychology,” Advances in Psychological Science, vol. 16(3),2008 .

[8] You Jian,“An Analysis on Practical Application of Interrogation Applied to Investigation and Trial after Psychology Test,” Journal of Fujian Public Security College, vol. 4,2008 .