Recognition Method of Abnormal Data Based on Interval Estimation

Jingqiu Gu, Zhihai Wang and Huarui Wu

ABSTRACT

The emergence of substantial abnormal data in network data acquisition system is due to the cognitive level of users mostly. The abnormal data not only cause the waste of storage and the operation resources, but also seriously affect the results of statistical analysis. How to identify the massive abnormal data is the premise to eliminate the abnormal data. Accordingly a detection method of abnormal data based on confidence interval, which is aimed to filter the abnormal data from the group, was proposed. Compared with the traditional method of simple random sampling, the sample space is selected from the overall sample by the above-mentioned method has more credibility. And different from the commonly used detection method which is aimed at abnormal point, this method can batch test the unexpected data. The experimental results suggest that this method can effectively filter out the abnormal data particularly in the cases where the data size is huge and the data deviation is not obvious. And on the basis of above-mentioned method, the detection result has little difference with the actual situation and the accuracy is over 90%.

INTRODUCTION

With the development of 3G network, cloud computing and WEB applications, an era of new technology represented by information technology has arrived. The traditional way of collecting information by artificial investigation is not only a waste of manpower and material resources, but also difficult to achieve a result of comprehensive, timely and accurate, with the rapid development of information technology, a variety of Web application system of information collection and management has changed the traditional mode of information processing. Although it
is the prospect of PE researching, it need operators have much computer knowledge and high operating skills mastered only by professional training, the difference in computer operation ability leads to a series of problems such as numerical error caused by over-large number and the inconsistencies of data units and magnitude, which leads to the low reliability of the data and great waste of storage and computing resources and even cause a wrong result [1, 2].

The problem caused by abnormal data is very serious, in order to reduce the influence of abnormal data, the researchers put forward a series of algorithms for outlier detection, such as simple random sampling and cross validation [3]. But these methods do not ensure that the results based on the sample set are subject to minimal noise interference. There are also some outlier detection methods, such as the Pauta method, Grubbs method and Victor Shaw method [4], but the above methods are generally used to detect outliers.

In view of the above situation, an outlier detection method based on a credible sample and confidence interval is proposed, which is mainly aimed at the numerical data with a huge amount and no significant difference. Unlike traditional detection methods, the proposed method reduces the time consumption and improves the accuracy. The method can detect abnormal data and cannot be affected by the significant abnormal data. Take information management systems of basic level agricultural technology extension as an example, for the anomaly detection in the annual income of the primary agricultural personnel, the validity and feasibility of the method has been proved.

DETECTION SCHEME

The anomaly detection method is divided into three parts, which are the acquisition of the trusted sample space, the determination of the confidence interval and the processing method of the abnormal data.

For the huge amount of data of the research object, it is a great consumption calculated according to the overall sample, whether in time or space complexity. The consumption is a great problem that cannot be ignored for a WEB application system, so a method used a sample space instead of the matrix is proposed. Sampling methods can obtain reliable information by centralized limited resources and are widely used in research activities.

The confidence level is the probability that the population parameter value falls in a certain area. The confidence interval is defined as the error range between the sample statistic value and the total parameter value at a certain confidence level [5]. With the increase of the confidence level, the probability of the data to be judged as abnormal data will become smaller, that is, the probability of mixing abnormal data in the normal data will increase. Confidence level should be selected according to the actual situation, in general, the value is chosen to be 0.95. For example: for a given confidence level of 0.95 and obtain the confidence intervals for the [30000, 70000], so when testing 100 samples of numerical, there will be 95 fall within the range. Therefore, the confidence interval can be used to screen the sample data, which is outside the credible range, as the exception data processing. Figure 1 shows the confidence interval when the confidence level is 0.95:
In figure 1, the blank areas represent acceptable among areas, which falls in the area of the sample are credible data, on both sides of the black part is unacceptable area, which falls in the area of the sample are abnormal data.

**Obtain Reliable Sample Space**

In the sampling process, select the part of samples to replace the whole, which would inevitably produce error, the higher the consistency of the internal data matrix, the less the sampling error; with the increase of sample space, the sampling error will be smaller, but at the same time, the cost of research will increased. There is no mature theory or conclusion about the influence of the sample quantity on the prediction accuracy, and the amount of data extracted should be larger than credible sampling number 45 in mathematical statistics [6].

At the beginning of the test, given the range of acceptable data and determine the random and non-return sampling forms, after sampling, verify whether each sample point were within the range firstly, if there are some samples beyond the scope, the samples will be eliminated.

After the preliminary selection of the sample space, the experimenters should obtain a credible sample based on this and exclude the extreme values in the sample space, this article identify extreme values by the method of "standardization numerical (Z - score)". The normalized value reflects the relative distance, which is measured by the standard deviation, which indicates the relative distance between the i-th variable value and the sample mean and the formula is:

\[
Z_i = \frac{X_i - \bar{X}}{\sigma}
\]  

(1)

The standard of value 1 indicates that the numerical level of the i-th sample point is equal to that of the standard deviation, the standardized value 2 indicates that the numerical level of the i-th sample point is 2 times larger than that of the standard deviation, i.e., the sample space average value of relative distance is 2. According to the method of “Pauta criterion”, the exception data is the value that the distance
between the mean and the average is more than 2 times of the standard deviation. These values are called highly abnormal data that the distance from the average is more than 3 times of the standard deviation. In the process of data management, according to the specific circumstances to determine whether the abnormal value is proposed, but for those which are highly abnormal should be directly removed.

In order to enhance the randomness of the selected sample space and to ensure the quality of the sample space, it is necessary to select a plurality of trusted sample space by the above method, and then, compare and select the appropriate one.

The standard deviation is used to reflect the degree of dispersion of the data set. At this time, the sample space is regarded as a single parent:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}}$$  \hspace{1cm} (2)

The greater the standard deviation of the sample set, the greater the amount of information contained, for example, when calculating the average wage of the employees, if we only consider the income of the leadership, the standard deviation of the results will be smaller than that of all the staff, which means the distribution of each sample point is more compact, but the result is not reasonable. One-sided selection of the sample led to a relatively small standard deviation, therefore, we need to compare multiple sets of samples, and select the relatively large standard deviation of the sample collection as the final credible sample space.

**Determine Confidence Interval**

We can show by the central limit theorem, when the sample size is sufficiently large (usually the sample size is greater than or equal to 30), the sample function,

$$U = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$  \hspace{1cm} (3)

Which basically obey standard normal distribution, the average value of the sample space can be approximated by the normal distribution, which is expressed as

$$\bar{X} \sim N(\mu, \sigma^2 / n)$$  \hspace{1cm} (4)

So after setting the value of the confidence level, we can directly into the corresponding formula to calculate the confidence interval of ensemble:

$$\left[ \bar{X} - \frac{S}{\sqrt{n}} t_{a/2}(n-1), \bar{X} + \frac{S}{\sqrt{n}} t_{a/2}(n-1) \right], (\alpha = 1 - \mu)$$  \hspace{1cm} (5)

Among them, the average value of the sample space can be calculated by the formula below:
\[ \overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \]  \hspace{1cm} (6)

Which is for the standard deviation of the trusted sample and the formula is:

\[ S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{X})^2}{n - 1}} \]  \hspace{1cm} (7)

\( t_{a/2} (n-1) \) can be obtained by looking up the t distribution table.

According to the formula, the confidence interval of the matrix can be obtained. After getting the confidence interval, we can deal with the test data, for the data not in the confidence interval, it is needed to be regarded as abnormal data; for the data in the confidence interval, can be considered to meet the requirements for reliable data.

**Abnormal Data Processing**

The data detection is divided into two cases, one is screening of the stored data, and the other is screening of the non-storage data.

**STORED DATA PROCESSING**

After detect outliers in marked abnormal data to be processed [7-9]. When the overall amount of data is huge and the amount of abnormal data is small (abnormal data generally considered the proportion of 5% or less), excluding outliers direct way is very effective in both in terms of efficiency and reliability.

**NON STORAGE DATA PROCESSING**

When using the application software to enter data, due to the weakness of the artificial, mistakes are inevitable [10-12]. Therefore, testing the validity of data advance can ensure the accuracy of the database. When the data is inputting, according to the above method, if the data is falling out of the confidence interval, a warning will be given, if the user insists on the data entry, the system will mark the abnormal data which would be pretreatment for analysis.

**CASE ANALYSIS**

Based on the annual income in the information management systems of basic level agricultural technology as actual data the method is verified. The confidence level of the experimental set is 0.95, and the programming language is JAVA.
**Experimental Preparation**

In order to obtain a reliable sample, a total of 5 sets of random sampling were conducted, and the number of samples was 2000 in all groups. Table 1 shows the data from the database randomly selected samples, here to show some of the data.

Table I. Sample tables for the data to be tested.

<table>
<thead>
<tr>
<th>No.</th>
<th>True data</th>
<th>No.</th>
<th>True data</th>
<th>No.</th>
<th>True data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>41000</td>
<td>11</td>
<td>40739</td>
<td>21</td>
<td>20172</td>
</tr>
<tr>
<td>2</td>
<td>56956</td>
<td>12</td>
<td>9012</td>
<td>22</td>
<td>55000</td>
</tr>
<tr>
<td>3</td>
<td>60252</td>
<td>13</td>
<td>100000</td>
<td>23</td>
<td>20980</td>
</tr>
<tr>
<td>4</td>
<td>40356</td>
<td>14</td>
<td>57660</td>
<td>24</td>
<td>50775</td>
</tr>
<tr>
<td>5</td>
<td>52804</td>
<td>15</td>
<td>82567</td>
<td>25</td>
<td>50000</td>
</tr>
<tr>
<td>6</td>
<td>48385</td>
<td>16</td>
<td>88416</td>
<td>26</td>
<td>38404</td>
</tr>
<tr>
<td>7</td>
<td>62000</td>
<td>17</td>
<td>39848</td>
<td>27</td>
<td>48100</td>
</tr>
<tr>
<td>8</td>
<td>49200</td>
<td>18</td>
<td>54000</td>
<td>28</td>
<td>60154</td>
</tr>
<tr>
<td>9</td>
<td>69200</td>
<td>19</td>
<td>58852</td>
<td>29</td>
<td>92988</td>
</tr>
<tr>
<td>10</td>
<td>2400</td>
<td>20</td>
<td>85300</td>
<td>30</td>
<td>66536</td>
</tr>
</tbody>
</table>

The most widely accepted range of the data set is 1 to 10 million yuan, as the threshold for the preliminary screening of the data.

**Experimental Results**

The detection method is used to screen out 5 groups of sample space variance respectively: 13594, 14949, 10328, 12458 and 10217, through the comparison of mean variance, select the second samples space as the credible sample space. According to the given confidence level 0.95, the confidence interval of the confidence sample space is obtained, then the abnormal value is screened out by using the obtained confidence interval, and compared with the abnormal data obtained by artificial test. Test results are shown in table II.

Table II. Table of test results.

<table>
<thead>
<tr>
<th>No.</th>
<th>Number</th>
<th>Confidence Interval</th>
<th>Abnormal points detected / actual abnormal points</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2000</td>
<td>[36810,70612]</td>
<td>92.68%</td>
</tr>
</tbody>
</table>

According to table II, we can find that there is little difference between the anomaly data and the actual situation by using this detection method, and the accuracy is more than 90%. The test strategy has been successful applied to data cleaning and verification process of the information management systems of basic level agricultural technology.
CONCLUSIONS

Aiming at the data group of large amounts and no significant deviation a new data rectification method based on the confidence interval is proposed. The anomaly detection method has good confidence in the sample space and can outlier detection for Batches. The practical application proves that the method has a great improvement on both efficiency and accuracy.

ACKNOWLEDGEMENT

This work was supported by Special Fund for Agro-scientific Research in the Public Interest (201303107).

REFERENCE