Industrial Engineering and Innovation Management (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/ieim.2022.051205
ISSN 2522-6924 Vol. 5 Num. 12

Outlier Detection of Power Supplier Quotation Based on Characteristic Index

Bo Gao[1], Jiawei Pan[2], Xuwen Liu[1], Changyu Qian[2,*], Linjie Wang[1]

[1]*Jiangsu Power Exchange Center Co., Ltd., No. 62, Yunnan Road, Gulou District, Nanjing, China*
[2]*School of Electrical Engineering, Southeast University, No. 2, Sipailou, Xuanwu District, Nanjing, China*
[*]*Corresponding author*

Keywords: Outlier detection, Feature extraction, Principal component analysis, Local outlier factor

Abstract: In order to detect the abnormal behavior of unit quotation in power market, an outlier detection method based on characteristic index is proposed. According to the characteristics of abnormal behavior of unit quotation, the corresponding characteristic indexes are extracted, and the dimension is reduced by principal component analysis method. The local outlier factor algorithm is used to detect outliers, and the evaluation indexes are compared when different number of features are extracted. The experimental results show that the proposed method can detect abnormal quotation units to a certain extent, and the evaluation index is improved with the number of features extracted.

1. Introduction

With the development of the electricity spot market, the reliability and accuracy of electricity data in the process of trading and settlement are required to be higher. In the actual operation process of the metering system, due to the faults and other problems in the operation of the power meter on the user side, the quality of the power data is different, which leads to the loss and abnormality of the data. The identification and correction of abnormal data can ensure the accuracy of the user data in the electricity market, which is conducive to the safe and stable operation of the electricity market. The detection of abnormal quotations of units is also a part of maintaining the safe and stable operation of the electricity market.

For outlier detection, experts and scholars at home and abroad have proposed many methods. There are many traditional outlier detection methods, and the classical outlier detection methods are usually divided into four categories: statistical, cluster-based, taxonomy-based and proximity based 0.Outlier detection first appeared in the field of statistics, which is generally divided into parametric and non-parametric methods. Literature [2] has proposed more than 100 outlier detection methods for different data distributions. Typical non-parametric methods, such as histogram visualization methods, are used in intrusion detection [2-3] and defect detection 0 A wide range of applications. Outlier detection generally obtains data distribution features through unsupervised learning, so common clustering algorithms can be applied to outlier detection after modification. The method mainly

detects outliers by considering the relationship between objects and clusters. In literature [5], K-means partition was used to achieve outlier detection, while in literature [6], multiple reunion classes were used to achieve outlier detection. [5-6] For datasets with class labels, a classifier can be trained to distinguish between normal data and outliers 0. However, for the problem of outlier detection, the dataset is often highly biased. The common practice is to build a single classification model, that is, only use normal data to train the model, so that the data points that do not belong to the normal class are judged as outliers. The approach based on proximity 0.The core idea of the approach is to define the proximity measure between data and determine the outliers according to the value of this measure. Among them, the typical methods are distance-based methods and density-based methods. The former is based on distance 0. The former is based on distance and the latter on density 0 Embody proximity.

Compared with the outlier detection methods mentioned above, the anomaly detection algorithm based on unsupervised learning has a better application prospect. Because the unsupervised anomaly detection algorithm only needs to rely on the data without data label in the training, it can use the overall characteristics of the data to get more accurate rules for the classification of anomalies. Therefore, this kind of method is more suitable for the electricity market, where the label data is often difficult to obtain, or the acquisition cost is extremely high. Outlier detection is a method based on unsupervised learning. On the basis of previous studies, this method combines the characteristics of unit quotation dataset to extract features and reduce dimensionality, and then uses local outlier factor algorithm to detect outliers. The proposed outlier detection model includes feature extraction, principal component analysis, local outlier factor calculation and other modules. Firstly, a variety of feature quantities that can characterize the abnormal behavior are extracted from the unit quotation data set. Then, the dimensionality of the feature set is reduced by using principal component analysis (PCA). After dimensionality reduction, the first two principal components are mapped to the two-dimensional plane in the form of scatter points.

2. Principle of Outlier Detection Algorithm Based on Feature Index

2.1. Feature Extraction

The data set contains the quotation of N unit at T, and the quotation pattern of the unit is represented by its hourly average quotation. Then the quotation sequence of each unit can be expressed as a T dimensional vector $x_n = \left\{ x_n^{(t)}, t = 1, 2, \ldots\ldots, T \right\}$, and all units can be expressed as a dataset $X = \left\{ x_n, n = 1, 2, \ldots\ldots, N \right\}$.On the basis of the dataset X, the characteristic quantity of unit quotation pattern can be further extracted.

According to the actual situation of electricity market operation, the variability index, trend index, volatility index and other indicators of abnormal behavior of unit quotation are proposed 0.

2.1.1. Indicators of Variability

The variability indicator refers to the head-end difference measure of the unit quotation pattern. Include:

1) The difference between the average quotation at the previous h time and the following h time for each unit $davg = \sum_{i=1}^{h} x_n^{(i)} / h - \sum_{i=1}^{h} x_n^{(T-i)} / h$.

2) The slope of linear fitting of quotation data at T times for each unit.

2.1.2. Trend Indicator

The steps to calculate a trend indicator are as follows:

1) Enter the unit hourly average quote dataset at X.

2) Calculate a series of n point simple moving averages F for each unit quote time series A.

3) Statistical series A and the relative size of the series F at each time point, if A has u segment below F, the number of points $a_1, a_2, \cdots\cdots, a_u$ contained in each segment is respectively, and A has v segment above F, the number of points $b_1, b_2, \cdots\cdots, b_v$ contained in each segment is respectively.

4) Calculate the upward trend index tra and downward trend index trb.

$$tra = \sqrt{\sum_{i=1}^{u} \left(a_i\right)^2} / u \tag{1}$$

$$trb = \sqrt{\sum_{i=1}^{v} \left(b_i\right)^2} / v \tag{2}$$

2.1.3. Volatility Indicators

1) Standard deviation of the quote series at T moment for each unit sd.

2) The standard deviation bsd of the quotation series at the previous h time.

3) Standard deviation esd of the quote series at the following h moment.

2.1.4. Other Indicators

1) Ratio of the average quote at the last h moment to the average quote at all moments.

2) Correlation coefficient between the quotation series of each unit and the median quotation series of all units at each moment.

2.2. Principal Component Analysis

Due to the large number of extracted features and different features may contain overlapping information, in order to visually display the quotation patterns of each unit in the low-dimensional plane and efficiently mine the abnormal unit, it is necessary to reduce the dimensionality of the dataset, that is, dimensionality reduction processing. The so-called dimension reduction is to transform the data set to represent as much information as possible in the original data set by a small number of new attributes. Principal component analysis (PCA) is a representative dimensionality reduction method.

Principal component analysis (PCA) is a data dimensionality reduction method. The basic idea of principal component analysis is to recombine the original correlated indicators into a small number of uncorrelated comprehensive indicators. $x_1, x_2, \ldots\ldots, x_p$ The comprehensive indicators should reflect the information represented by the original variables to the greatest extent, and ensure that the new indicators remain independent of each other.

If $y_1, y_2, \ldots\ldots, y_m \left(m \le p\right)$ is used to represent the m principal components of the original variable $x_1, x_2, \ldots\ldots, x_p$,i.e.

$$\begin{cases} y_1 = l_{11}x_1 + l_{12}x_2 + \ldots\ldots l_{1p}x_p \\ y_2 = l_{21}x_1 + l_{22}x_2 + \ldots\ldots l_{2p}x_p \\ \ldots\ldots \\ y_m = l_{m1}x_1 + l_{m2}x_2 + \ldots\ldots l_{mp}x_p \end{cases} \tag{3}$$

2.3. Local Outlier Factor

The Local Outlier Factor (LOF) algorithm is a density-based local outlier detection algorithm proposed by Breunig in 2000. This method is applicable to the data with different density dispersion of different clusters.

LOF algorithm the basic idea is, according to the data points around the dense situation, first calculate a local to each data point density, then further by local of density of each data point is calculated from a group of factors, the factor that identifies a data point from the group of stray, factor value, the greater the said the higher the degree of outliers, the smaller the factor value, Denotes a lower degree of outliers. Finally, the top (n) points with the largest degree of outliers are output. The accurate description of the local outlier factor is based on the following definitions.

Definition 1 Distance from point to point:

$d(p,o)$, the distance from data point p to data point o.

Definition 2 k distance

The first k distance of a data point p from $d_k(p)$, defined as: $d_k(p) = d(p,o)$, satisfies

a) there are at least k points in the set o' that are not included p, such that $d(p,o') \le d(p,o)$;

b) there are at most $k-1$ excluded points in the set o', such that $d(p,o') < d(p,o)$.

Definition 3 k distance neighborhood

The set of p points within the k distance neighborhood $N_k(p)$ of data point p, the k distance of a pointer, including the point on the k distance.

Easy to know, have $|N_k(p)| \ge k$.

Definition 4 k reachable distance

$$reach\_dist_k(o,p) = \max\{d_k(o), d(o,p)\} \tag{4}$$

Definition 5 Locally achievable density:

$$lrd_k(p) = 1/\left(\frac{\sum_{o \in N_k(p)} reach\_dist_k(o,p)}{|N_k(p)|}\right) \tag{5}$$

The k local reachable density of a data point p is the reciprocal of the average k reachable distance of all points in the neighborhood from point p to point p. It represents the density of point p. The higher the density of point p and its surrounding points, the more likely the reachable distance of each point is the smaller k distance, and the larger is the value of lrd.The lower the density of point p and surrounding points, the more likely the reachable distance of each point is the actual distance between the larger two points, and the smaller is the value of lrd.

Definition 6 Local outlier factor:

$$LOF_k(p)=\frac{\sum_{o\in N_k(p)}\frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|}=\frac{\sum_{o\in N_k(p)}lrd_k(o)}{|N_k(p)|}/lrd_k(p) \tag{6}$$

The k local outlier factor of data point p means that the average local accessible density of all points in the $N_k(p)$ neighborhood of point p is compared with the local accessible density of point p. The more the ratio is greater than 1, the less the density of point p is than the density of its surrounding points, and the more likely it is that point p is an outlier. The less this ratio is, the more the density of point p is greater than the density of the surrounding points, and the more likely point p is to be normal.

3. Algorithm Flow

3.1. Sample Matrix

The process of outlier detection algorithm based on feature index is shown in Figure 1. Firstly, the unit quotation data are collected and preprocessed, and then the abnormal quotation features are extracted to obtain the sample matrix $x=(x_1,x_2,\ldots\ldots,x_p)$, which is normalized to obtain the standardized sample matrix $\tilde{x}=(\tilde{x}_1,\tilde{x}_2,\ldots\ldots,\tilde{x}_p)$. The standardized data is denoted as $\tilde{x}_{ij}=\frac{x_{ij}-\bar{x}_j}{s_j},(i=1,2,\ldots\ldots,n;j=1,2,\ldots\ldots,p)$, where $\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$ is the sample mean of the j index. $s_j=\frac{1}{n\text{-}1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$ is the sample standard deviation of the j index.

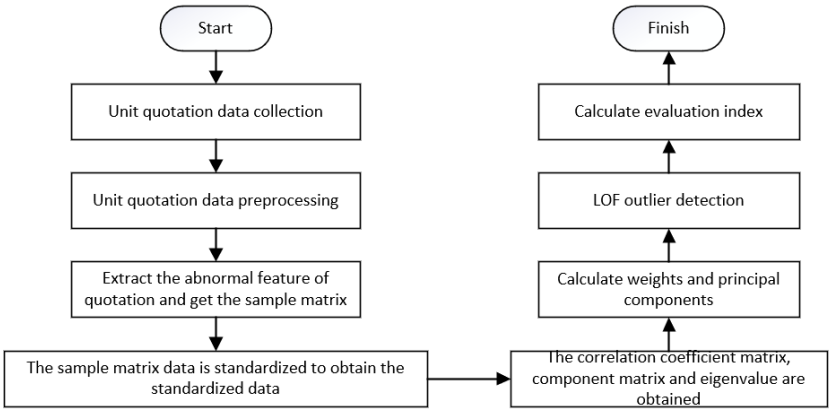

Figure 1: Outlier detection algorithm flow based on feature index

3.2. Calculation of Principal Components

First, compute the correlation coefficient matrix R, where the elements $r_{ij}=\frac{\sum_{k=1}^{n}\tilde{x}_{ki}\tilde{x}_{kj}}{n-1}(i,j=1,2,\ldots\ldots,p)$, where $r_{ii}=1,r_{ij}=r_{ji}$, r_{ij} are the correlation coefficients of the i

index and the j index. Then, the eigenvalue $\lambda_1 \geq \lambda_2 \geq \ldots\ldots \geq \lambda_m \geq 0$ of the correlation coefficient army array R and the corresponding eigenvector $u_1, u_2, \ldots\ldots, u_m$ are calculated. The eigenvector $u = \frac{A}{\sqrt{\lambda}}$ can be calculated from the component matrix A and eigenvalue λ .Finally, the principal components $y = u\tilde{x}$ are calculated, and the variance contribution rate $b_j = \frac{\lambda_j}{\sum_{k=1}^{m} \lambda_k} (j = 1, 2, \ldots\ldots, m)$ and the cumulative variance contribution rate $\alpha_t = \frac{\sum_{k=1}^{t} \lambda_k}{\sum_{k=1}^{m} \lambda_k}$ are calculated. When $\alpha_t \geq 0.85$, the former t indicator variable $y_1, y_2, \ldots\ldots, y_t$ is selected as the t principal component.

3.3. Outlier Detection

LOF is used for outlier detection. When evaluating the function of outlier algorithm, accuracy P (precision), recall R (recall) and F-Measure are used in this paper. Let L_1 be the set of detected outliers, and L_2 be the set of all outliers. The calculation formula is:

$$P = \frac{|L_1 \cap L_2|}{|L_1|} \times 100\% \tag{7}$$

$$R = \frac{|L_1 \cap L_2|}{|L_2|} \times 100\% \tag{8}$$

$$F = \frac{(\beta^2 + 1) PR}{\beta^2 P + R} \tag{9}$$

In Equation (0-1), β is the relatively important weight of response accuracy and recall. In order to find as many abnormal quotation units as possible, β will be set as 1.

4. Example Analysis

4.1. Principal Component Analysis

4.1.1. Feature Extraction

The quotation data of 117 units at each time of 24 hours a day are selected, and the characteristics are extracted according to the variability index, trend index, volatility index and other indicators of abnormal behavior of unit quotation. The 13 features extracted from unit quotation series include: Rising trend index C1, C2 downward trend indicators, before and after 1, 4, 8 times offer mean the difference between the C3, C4 and C5, offer all time sequence of each pump unit standard deviation C6, before and after 8 times offer sequence the standard deviation of the C7, C8, after 4, 8, 12 times the average offer with all the moments mean ratio of C9, C10, C11, The slope of linear fitting of quotation series C12, and the correlation coefficient C13 between the quotation series of each unit and the median quotation series of all units at each time. After the feature extraction is completed, the

sample matrix is obtained and standardized.

4.1.2. Principal Component Calculation

The correlation matrix is a square matrix of the correlation coefficients between two variables. Generally speaking, most variables with higher correlation coefficients will be classified into the same principal component, and when the correlation coefficients of most variables in the original data are greater than 0.3, the application of principal component analysis will achieve satisfactory results. As can be seen from Table1, Most of the values of the elements in the correlation matrix are greater than 0.3, so the effect of PCA is better.

Table 1: Correlation matrix

Correlation	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
C1	1.000	-0.045	-0.210	-0.210	0.808	-0.807	0.210	-0.818	0.849	-0.839	-0.845	-0.802	0.107
C2	-0.045	1.000	0.022	0.022	-0.151	0.151	-0.022	0.152	-0.143	0.142	0.143	0.150	0.125
C3	-0.210	0.022	1.000	1.000	-0.023	0.017	-1.000	0.083	-0.105	0.039	0.076	-0.009	0.126
C4	-0.210	0.022	1.000	1.000	-0.023	0.017	-1.000	0.083	-0.105	0.039	0.076	-0.009	0.126
C5	0.808	-0.151	-0.023	-0.023	1.000	-1.000	0.023	-0.998	0.994	-0.998	-0.996	-0.999	-0.390
C6	-0.807	0.151	0.017	0.017	-1.000	1.000	-0.017	0.998	-0.994	0.998	0.996	1.000	0.389
C7	0.210	-0.022	-1.000	-1.000	0.023	-0.017	1.000	-0.083	0.105	-0.039	-0.076	0.009	-0.126
C8	-0.818	0.152	0.083	0.083	-0.998	0.998	-0.083	1.000	-0.998	0.997	0.998	0.996	0.396
C9	0.849	-0.143	-0.105	-0.105	0.994	-0.994	0.105	-0.098	1.000	-0.998	-1.000	-0.991	-0.353
C10	-0.839	0.142	0.039	0.039	-0.998	0.998	-0.039	0.997	-0.998	1.000	0.999	0.997	0.346
C11	-0.845	0.143	0.076	0.076	-0.996	0.996	-0.076	0.998	-1.000	0.999	1.000	0.994	0.350
C12	-0.802	0.150	-0.009	-0.009	-0.999	1.000	0.009	0.996	-0.991	0.997	0.994	1.000	0.386
C13	0.107	0.125	0.126	0.126	-0.390	0.389	-0.126	0.396	-0.353	0.346	0.350	0.386	1.000

Table 2: Principal components and contribution rates

The principal components	The eigenvalue λ	Variance contribution rate b (%)	Cumulative variance contribution rate α (%)
Principal component 1	7.861	60.469	60.469
Principal component 2	3.023	25.252	85.721

The variance contribution rate b and cumulative variance contribution rate α are mainly used to judge how many principal components are appropriate to extract. Table 2 shows that the cumulative contribution rate of the first principal component and the second principal component is 85.721%, which means that the two new variables can be used to replace the original 13 variables to achieve the purpose of dimension reduction.

Through $y_1 = u_1\tilde{x} = \frac{A_1}{\sqrt{\lambda_1}}\tilde{x}, y_2 = u_2\tilde{x} = \frac{A_2}{\sqrt{\lambda_2}}\tilde{x}$, calculate the principal components.

4.2. Outlier Detection

The outlier detection results based on LOF algorithm are shown in Figure 2 to Figure 5, where the abscissa is the first principal component, the ordinate is the second principal component, and the output result is the point with the highest degree of $top(n)$ outliers. The black point is the normal quotation unit, the red point is the unit with a certain degree of outliers, and the size of the red circle is the size of the outlier degree. The larger the radius of the red circle, the higher the outlier degree of the unit.

In order to test the effect of the extracted quotation abnormal feature indexes, 6 features (Figure 2), 8 features (Figure 3), 10 features (Figure 4) and 13 features (Figure 5) were extracted respectively. With the increase of the number of extracted feature indexes, the number of detected outliers also increased.

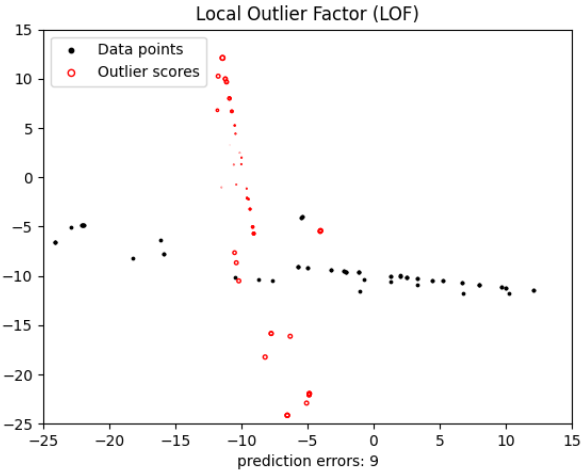

Figure 2: Extracting 6 eigenvalues

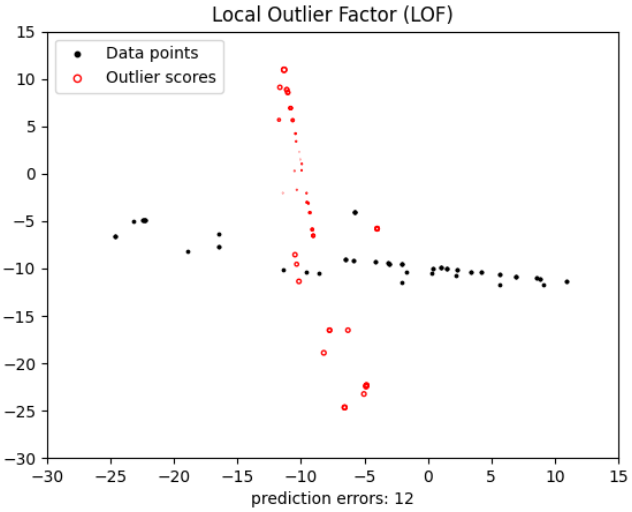

Figure 3: Extracting 8 eigenvalues

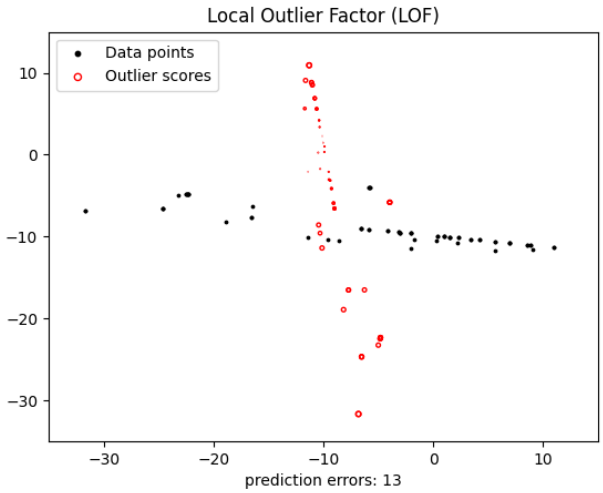

Figure 4: Extracting 10 eigenvalues

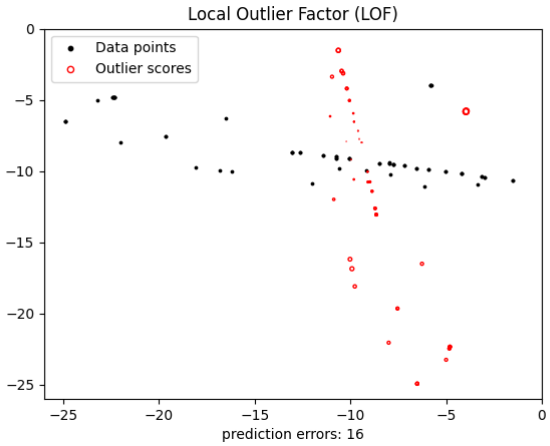

Figure 5: Extracting 13 eigenvalues

The evaluation indexes with different number of feature indexes were calculated respectively, and the results are shown in Table 3.

Table 3: Comparison of evaluation indexes extracted with different number of feature indexes

Number of feature indexes	Accuracy (%)	Recall (%)	F-score (%)
13	75.83	73.79	74.78
10	70.21	68.11	69.08
8	68.34	64.12	66.13
6	62.72	60.63	61.64

As can be seen from the table, with the increase of the number of extracted feature indicators, the accuracy P, recall R and F value of outlier detection are improved to varying degrees.

5. Conclusions

In this paper, an unsupervised learning based abnormal quotation detection method for unit is studied. In the absence of training samples, outliers are identified by analyzing the relationship between samples, population and each sample, which is called abnormal quotation pattern. The

proposed unsupervised learning based outlier bid detection model includes feature extraction, principal component analysis, and local outlier factor calculation modules.

1) Principal component analysis can reduce the dimension of abnormal bid feature set and eliminate the information overlap between original features. Using the first two principal components to represent the quotation pattern of each unit, the normal unit can be mapped to the high density area on the two-dimensional plane, and the abnormal unit can be mapped to the low density area.

2) With the increase of the number of extracted abnormal feature indexes of unit quotation, the number of detected outliers also increases, and the accuracy P, recall R and F value are improved to different degrees.

3) The effect of only using the abnormal feature index of quotation and dimensionality reduction method is not ideal, and the outlier detection algorithm can be improved to a certain extent to improve the detection effect.

References

[1] Han Jiawei, Kamber M.Data Mining: Concepts and techniques. Fan Ming, Meng Xiaofeng, Trans. 3rd Ed. Beijing: China Machine Press, 2012:186-188.

[2] Barnett V, Lewis T, Abeles F. utliers in statistical data. 3rd ed. Hoboken: Wiley, 1994.

[3] Eskin E.A nomaly detection over noisy data using learnt probability distributions//Proc of the 17th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers, 2000: 255-262.

[4] Fawcett T, Provost F. The Adaptive fraud detection. Data Mining & Issue 12 Merlin, et al.: A survey of outlier detection techniques 3525 ·Knowledge Discovery, 1997, 1(3): 291-316.

[5] Ni Wei-Wei, Lu Jie-ping, Chen Geng, Sun Zhi-hui. Data displacement outgroup detection algorithm based on K-means partitioning. Computer Research and Development, 2006 (09): 1639-1643.

[6] Gu Ping, Liu Haibo, Luo Zhiheng. An outlier detection algorithm Based on Multiple reunion classes. Application Research of Computers, 2013, 30 (03): 751-753+756.

[7] Tao Qing, Cao Jinde, Sun Demin. Regression method based on Support Vector Machine classification. Journal of Software, 2002 (05): 1024-1028.

[8] Zhang Yi, Liu Xumin, Sui Ying, Guan Yong. Research and Improvement of Denoising Algorithm Based on K-Nearest Neighbor Point Cloud. Journal of Computer Applications, 2009, 29 (04): 1011-1014.

[9] Wang Zhen. Analysis and Research of distance-based Outlier detection algorithm. Chongqing University, 2011.

[10] Wang Jinghua, Zhao Xinxiang, Zhang Guoyan, Liu Jianyin. NLOF: A novel density-based local outlier detection algorithm. Computer Science, 2013, 40 (08): 181-185.

[11] Thakur S, Gasperis G D. Cartel formation in charging network for electric vehicles// IEEE 16th International Conference on Environment and Electrical Engineering. Florence: Institute of Electrical and Electronics Engineers, 2016:1-6.