ISSN 2616-5775 Vol. 6, Issue 13: 82-91, DOI: 10.25236/AJCIS.2023.061313

# An Outlier Detection Algorithm Based on Spectral Clustering

**Tingfeng Liu[1,a,*], Weifeng Zhu[1,b]**

[1]*School of Information Engineering, Henan Vocational University of Science and Technology, Zhoukou, China*
[a]*3137118581@qq.com*, [b]*395288252@qq.com*
[*]*Corresponding author*

***Abstract:*** *Outliers refer to data that is far away from conventional data objects, resulting from different mechanisms, and there is a clear gap from conventional data. Outlier detection algorithms include statistics-based, distance-based, density-based, and cluster-based algorithms. Spectral clustering has become a research hotspot in the field of clustering because it can cluster on arbitrary-shaped data samples. Through the verification of artificial data and real data sets, the experiments show that the outlier detection algorithm based on spectral clustering is superior to the traditional outlier detection algorithm. Spectral clustering algorithm is a clustering method based on structural mapping. In a data set containing outliers, spectral clustering can be used to describe the structural characteristics of the data set and detect the presence of outliers. The essence of spectral clustering is to convert the clustering problem into the optimal partitioning problem of graphs. The graph partitioning criteria include mincut, RatioCut, Ncut etc. The spectral clustering algorithm is generally divided into two steps. Firstly, a relationship graph is established based on the similarity measure of data sample points.Secondly, an algorithm is constructed to perform cluster segmentation on the graph. This paper has improved in two aspects: on the one hand, the Self-Tuning algorithm proposed by Zelnik-Manor and Perona is used to recalculate the scale parameters, and on the other hand, the k-means algorithm is optimized. The algorithm is compared with traditional outlier detection algorithm. The effectiveness analysis and results show that the algorithm proposed in this paper has more advantages in detecting outliers.*

***Keywords:*** *Spectral Clustering, Self-Tuning, Scale Parameters, Factor of Outlier*

## 1. Introduction

Outlier detection is an important branch of data mining. With the rapid development of data mining technology, people pay more and more attention to outliers that deviate significantly from the overall data trend while paying attention to the overall data trend, because these data points often contain more important information. D. Hawkins' definition of outlier : "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a diffffferent mechanism"[1]. The purpose of outlier detection is to effectively identify abnormal data in the data set and mine meaningful potential information in the data set. In recent years, outlier detection has been widely used in intrusion detection systems, credit-card fraud, medical diagnosis,interesting sensor events and so on[2]. The methods of outlier detection mainly include statistics-based, distance-based, density-based, and clustering-based methods. The clustering technique is usually called unsupervised learning, because unlike supervised learning, there is no classification or grouping information that represents the data category in clustering. After years of development, clustering algorithms have appeared a large number of classic algorithms such as: k-means[3], DBSCAN[4], EM[5] algorithms and so on. However, clustering-based methods are only suitable for certain problems and have certain limitations, such as sensitivity to input parameters and spatial distribution shape of the data set. The recent spectral clustering algorithm can overcome the shortcomings of traditional algorithms, ensure convergence to the global optimal solution, have greater adaptability to data distribution, and have excellent clustering effects, so they have received widespread attention[6-7].

ISSN 2616-5775 Vol. 6, Issue 13: 82-91, DOI: 10.25236/AJCIS.2023.061313

### *1.1. Similarity Matrix*

The similarity matrix is a measure of the distance between any two points in the sample point, the closer the points are, the higher the similarity.Let G = (V, E) be an undirected graph with vertex set $V = \{v_1, v_2, ...v_n\}$ .In the following we assume that the graph G is weighted, that is each edge between two vertices $v_j$ and $v_j$ carries a non-negative weight wij ≥ 0. The weighted adjacency matrix of the graph is the matrix $W = (w_{ij})_{i,j=1,...n}$ .If $w_{ij}$=0 this means that the vertices $v_j$ and $v_j$ are not connected.As G is undirected we require $w_{ij} = w_{ji}$ . The degree of a vertex $v_i \in V$ is defined as

$$d_i = \sum_{j=1}^{n} w_{ij} \tag{1}$$

The degree matrix D is defifined as the diagonal matrix with the degrees $d_1$,...., $d_n$ on the diagonal.

The goal when constructing similarity graphs is to model the local neighborhood relationships between the data points.There are three types of methods for constructing the similarity matrix W: ε-neighborhood graph, k-nearest nerghbor graph, fully connected graph.The fully connected graph is most commonly used, and different kernel functions can be selected to define edge weights. Commonly used are polynomial kernel function, Gaussian kernel function, and Sigmoid kernel function. Here, Gaussian kernel function is used.At this time, the similarity matrix and the adjacency matrix are the same:

$$W_{ij} = S_{ij} = \exp(-\frac{\| x_i - x_j \|^2}{2\delta^2}) \tag{2}$$

### *1.2. Graph Laplacians*

There are usually three methods for constructing Laplacian matrices, the unnormalized graph Laplacian matrix is defifined as

$$L=D-A \tag{3}$$

The normalized graph Laplacians includes two forms.We denote the first matrix by Lsym as it is a symmetric matrix, and the second one by Lrw as it is closely connected to a random walk.

$$L_{sys} = D^{\frac{1}{2}} L D^{\frac{1}{2}} = E - D^{\frac{1}{2}} A D^{\frac{1}{2}} \tag{4}$$

$$L_{rw} = D^{-1} L = E - D^{-1} A \tag{5}$$

D is the degree matrix, A is the similarity matrix, and its element value is $A_{ij}$.

### *1.3. Spectral Clustering*

The theoretical basis of the spectral clustering algorithm is graph theory, which regards data clustering as a graph division problem, and uses graphs to model the objects in the data set and their interrelationships[8]. The vertices in the graph represent the set For objects, the edges in the graph represent a certain relationship between objects, and the strength of the relationship is expressed by the weight of the edges.

Spectral clustering algorithm maps data points to weighted undirected graphs, using sample data points as nodes of the graph, and the edge weights represent the similarity between data point features, and then use some segmentation criteria to get the best data Two-dimensional division[9]. The standard spectral clustering algorithm deals with the dichotomy problem, which uses only one feature vector to divide, which is inefficient and unstable. The emergence of K-way spectral clustering algorithm[10] overcomes the shortcomings of the bipartite spectral clustering algorithm. Using multiple feature vectors for partitioning at the same time can not only approximate the optimal k-way partition, but also avoid the loss of information caused by Instability.

Spectral clustering works well in some areas, but there are also problems that need to be resolved. The algorithm needs to use the Gaussian function to construct the similarity matrix, which will involve

ISSN 2616-5775 Vol. 6, Issue 13: 82-91, DOI: 10.25236/AJCIS.2023.061313

the scale parameters. The scale parameters have no range and regularity when they are selected, which makes the spectral clustering sensitive to the parameters. If the selection is improper, the clustering effect drops sharply. Many scholars optimize the spectral clustering algorithm to make it better. TaoTong et al. proposed a one-step spectral clustering algorithm based on self-speed learning. This method uses the missing value mapping matrix to deal with the missing data, uses a one-step clustering method to reduce the error accumulation problem caused by multiple steps in traditional spectral clustering, and uses a custom step size rule to avoid the effects of noise and outliers; Zelnik-Manor proposes that we propose an adaptive spectral clustering algorithm that automatically calculates scales and group numbers, and can handle multi-scale data with problems in previous methods; Afzalan and Jazizade proposed and evaluated iterative feature search multi-scale, Heuristic algorithm for high-dimensional feature space automatic clustering. The proposed heuristic algorithm does not require a priori assumptions about the number of clusters, nor does it require assumptions about the scaling parameters of the affinity metric containing a series of values of the number of clusters.

In summary, the current problems of spectral clustering mainly include: (1) the number of categories when clustering with spectral clustering algorithm is difficult to determine automatically; (2) the spectral clustering algorithm has a large dependence on parameters, the parameters The selection has a great influence on the clustering results.

## 2. Optimization of self-tuning spectral clustering algorithm

### *2.1. Optimization of scale parameters*

Given a sample set $X=\{x_1,x_2,...,x_n\}$, each sample x has m-dimensional features, $x_i=(x_{i1},x_{i2},...x_{im}),(i=1,2,... .,n)$. Define a local scale parameter $\sigma_i$ for each data point in the self-turning algorithm,

$$\sigma_i = d(x_i, x_k) \tag{6}$$

The parameter $\sigma_i$ takes the Euclidean distance from point xi to the k-th point $x_k$ in its nearest neighbor, but $\sigma_i$ may be affected by outliers. Therefore, the scale parameter $\sigma_i$ in this paper takes the weighted distance of k points in the k nearest neighbor of point $x_i$.

$$\sigma_i = \sum_{j=1}^{K} w_{ij} d_{ij} \tag{7}$$

Note that, $w_{ij} = 1 - \ln(d_{ij} / \sum_{k=1}^{K} d_{jk})$, because $0 < d_{ij} / \sum_{k=1}^{K} d_{jk} < 1$, $w_{ij}>1$.

### *2.2. Optimization of k-means algorithm*

Since k-means clustering is more sensitive to the selection of initial clustering centers, the k-means algorithm is unstable, and the spectral clustering algorithm is unstable. This paper proposes a dichotomy k-means algorithm based on the k-means algorithm. That is, at the beginning of the algorithm, the two objects with the furthest distance are used as the initial center to form two initial clusters, and these clusters continue to split. This continues until k cluster. This method makes the algorithm less susceptible to initialization because it has only two centroids per step. In this way, objects in the same class can be guaranteed to have great similarity in subsequent processing, and there is extremely low similarity between different classes.

We introduce Davies-Bouldin Index (DBI), an index to evaluate the advantages and disadvantages of clustering algorithms. The smaller the DBI value, the smaller the DBI value means that the clustering result is close to the same cluster, and the different clusters are separated farther. That is, the smaller the intra-class distance, the larger the inter-class distance.

The core of the optimized algorithm is as follows: (1) Calculate the two furthest objects in the data set: $x_1$, $x_2$. (2) Calculate the distance between the remaining data objects in the data set to $x_l$, $x_2$, divide them into the class with the smallest center, mark them, and record the minimum distance. (3) After the division is completed, the cluster centers are recalculated to obtain $c_1$ and $c_2$.(4) Use the maximum and minimum distance method, $\max\{\min(d(c_1, j), d(c_2, j))\}$, j = 1, 2, ... n, get $x_j$, use the DBI formula

ISSN 2616-5775 Vol. 6, Issue 13: 82-91, DOI: 10.25236/AJCIS.2023.061313

to determine whether it is the best clustering center; (5) Go to (2), the data set the center again; (6) According to the DBI formula, calculate DBInew and compare it with the last calculated DBIold. If $DBI_{new}<DBI_{old}$, then find a reasonable $x_j$, and the value of k is increased by 1 on the basis of the last time, otherwise Find a new clustering center that meets the conditions, and the clustering ends. (7) By analogy, until no new clustering center that meets the condition can be found, the clustering result is finally output.

$$DBI = \frac{1}{N}\sum_{i=1}^{N}\max(R_{ij}) \tag{8}$$

$$R_{ij} = \frac{S_i + S_j}{M_{ij}} \tag{9}$$

Measure the similarity between category i and category j. $S_i$ calculates the average distance from the cluster data to the cluster centroid $S_i = \left\{\frac{1}{T}\sum_{j=1}^{T_i}| X_j - A_i |^q\right\}^{1/q}$, $X_j$ represents the j-th data point in the i-th category; $A_i$ represents the center of the i-th category; Ti represents the number of data points in the i-th category; q takes 1 to represent the average value of the distance from each point to the center, and q takes 2 to represent The standard deviation of the distance from each point to the center can be used to measure the degree of dispersion. $M_{ij} = \left\{\sum_{k=1}^{N}|a_{ki}-a_{kj}|^p\right\}^{1/p}$ aki represents the value of the K-th attribute of the center point of the i-th category, and Mij is the distance between the center of the i-th category and the j-th category.

### *2.3. Factor of outlier*

After the data is clustered, outliers need to be mined, for which an outlier index is introduced. First we have the following definition:

Definition 1: Hypothesis $C=\{C_1,...,C_k\}$ is a set of clustering sequences and $|C_1| \geq ... \geq |C_k|$, where $|C_i|$ represents the number of objects in the cluster $C_i (i = 1,...k)$,and k is the number of clusters. Given two positive numbers α and β, if the following formula holds, we define b as the boundary of size clustering.

$$|C_1| + ... + |C_b| \geq |D|\alpha \tag{10}$$

$$|C_b| / |C_{b+1}| \geq \beta \tag{11}$$

Thus, the set of large clusters can be defined as $LC = \{C_i \mid i \leq b\}$, and the set of small clusters can be defined as $SC = \{C_j \mid j > b\}$.

Definition 1 gives a quantitative method for distinguishing between large and small clusters. Equation (10) indicates that the number of data in the large cluster is greater than the ratio α of the entire data. Therefore, a cluster containing a large number of data objects is a large cluster. Equation (11) indicates that large clusters are β times larger than small clusters, indicating that large and small clusters must have sufficient gaps in scale. In this way, the effect of small clustering on the data set can be ignored.

For any object x in the dataset, the factor of outlier (FOO):

$$FOO(x) = \frac{1}{|C_i|} dist(x, C_i) \tag{12}$$

Where dist(x, $C_i$) is the Euclidean distance between the object x and the center of the cluster $C_i$, and $|C_i|$ is the number of objects in the cluster $C_i$.

ISSN 2616-5775 Vol. 6, Issue 13: 82-91, DOI: 10.25236/AJCIS.2023.061313

### *2.4. Optimized spectral clustering algorithm steps*

Input: data set containing n data, parameters $\alpha, \beta$.

Output: m outlier candidate sets.

(1) Construct the similarity matrix A, the elements in the matrix

$$A_{ij} = \exp(-\frac{\| x_i - x_j \|^2}{\sigma_i \sigma_j}) \tag{13}$$

Where $\sigma i=dm(x_i,x_k)$ is the weighted Euclidean distance from the sample point to the k-th nearest neighbor point.

(2) Construct the degree matrix D, the element D(i,j) on the main diagonal of the degree matrix D is the sum of the elements of the i-th row of the similarity matrix W, and all other elements are 0;

(3) Construct the Laplace matrix L=D-W as $L = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ ,then calculate the normative Laplacian matrix:

$$L_{sys} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = E - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \tag{14}$$

(4) Let be the first k minimum eigenvalues of $L_{sys}$, $v_1$, $v_2$,..., $v_k$ are the corresponding eigenvectors, then construct the matrix $V=[v_1,v_2,...,v_k]$, vi is the column vector, i=1,2,...K;

(5) Normalize V by row to get matrix U,

$$U = \frac{V_{ij}}{(\sum_j V_{ij}^{\ 2})^{\frac{1}{2}}} \tag{15}$$

(6) Treat each row in U as a point in a sample space, Rk and use the optimized K-means clustering to class k.

(7) Assign the original sample xi to the j-th cluster if and only if the i-th row of the matrix U belongs to the j-th cluster.

(8) According to the parameters $\alpha, \beta$ and the definition of large and small clusters, class k is divided into a large cluster set $C_j$ and a small cluster set $C_i$.

(9) Calculate the outlier index FOO of each data x in the cluster,

(10) Sort the data points according to the size of the outlier index and return to the first m outliers.

## 3. Experimental Results

### *3.1. Synthetic data set*

We will experiment with different data sets to evaluate our proposed algorithm and compare it with two outlier detection algorithms (KNN algorithm, LOF algorithm). We use artificial data sets to prove that the proposed outlier detection method based on spectral clustering can effectively identify outliers. The clustering effect of the artificial data set Data1 is shown in Figure 1, where A, B, C, and D are four normal clusters of different sizes, E is the only outlier class, and some outlier data is distributed around it. The specific information is shown in Table 1. Experiments with outlier detection algorithm, KNN algorithm, and LOF algorithm based on spectral clustering, respectively, the parameters $\alpha$ and $\beta$ based on spectral clustering are set to 0.75 and 5, the parameter MinPts of LOF algorithm is set to 30, and the detection based on KNN The nearest neighbor parameter k of the algorithm is set to 5, and the result is shown in the figure below.

ISSN 2616-5775 Vol. 6, Issue 13: 82-91, DOI: 10.25236/AJCIS.2023.061313

*Table 1: Synthetic data set Data1 information*

| Data1 | A | B | C | D | E |
|---|---|---|---|---|---|
| Number of data | 50 | 56 | 116 | 66 | 6 |
| Number of ouliers | 4 | 0 | 8 | 7 | 6 |

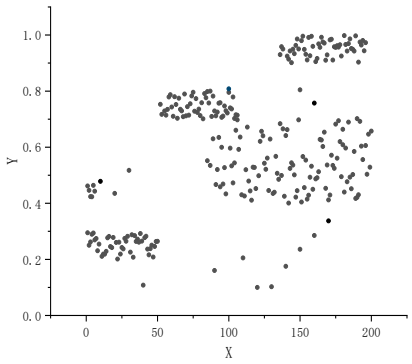



*Figure 1: Synthetic data set.*

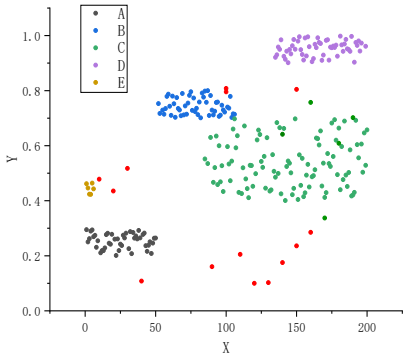



*Figure 2: KNN outlier detection algorithm.*

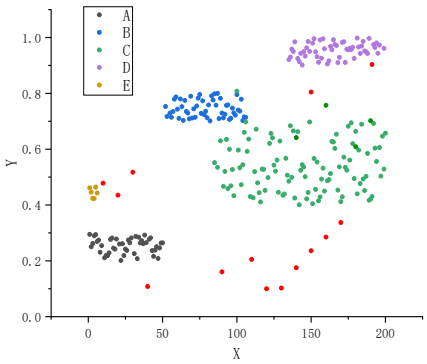



*Figure 3: LOF outlier detection algorithm.*

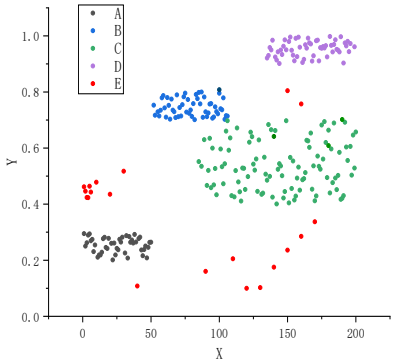



*Figure 4: Outlier detection algorithm based on spectral clustering.*

It can be seen from Figure 2 that the KNN algorithm can detect most of the outliers (represented in red) and there are detection errors, and the outlier class E cannot be correctly detected; from Figure 3, it can be seen that the LOF algorithm is more effective in identifying outliers KNN is slightly better, but also cannot detect outlier E. It can be seen from FIG. 4 that the outlier detection algorithm based on spectral clustering can correctly detect all outliers and the outlier class E can also be correctly identified.

The above-mentioned outlier detection effects are analyzed by using precision and false positive indicators. The accuracy indicates the detection accuracy of the algorithm. The closer the value is to 1, the better the detection effect. The false detection rate is a measure of the degree of false detection of non-related information. The closer the value is to 0, the better the effect.

*Table 2: Comparison of three algorithms detection*

| Algorithms | KNN | LOF | Algorithms based on spectral clustering |
|---|---|---|---|
| Precision | 0.801 | 0.875 | 0.996 |
| False Positive | 0.28 | 0.22 | 0.04 |

It can be seen from the table 2 that the accuracy rate of the KNN algorithm and the LOF algorithm is lower than the outlier detection algorithm based on spectral clustering, and the false detection rate based on the spectral clustering algorithm is significantly lower than that of the KNN algorithm and the LOF algorithm, indicating that the outliers based on spectral clustering. The point detection algorithm is obviously superior to other algorithms.

ISSN 2616-5775 Vol. 6, Issue 13: 82-91, DOI: 10.25236/AJCIS.2023.061313

### *3.2. UCI data set*

In order to test the effectiveness of the algorithm on real data, we selected Breast Cancer Wisconsin (WBC), Arrhythmia, Glass, Wine, Yeast, Cardiotocogrpahy in the UCI standard data set as the experimental data set. In the detection of outliers, it is often impossible to directly process the classification attributes in the mixed data. In some data sets with semantically meaningful outliers, it is necessary to adapt to the pre-processing of the data based on the knowledge background and expert knowledge of the data set. Detect variant data sets at outliers. In the experiment, we preprocess the data set first.

(1) Processing of missing values: Since standard outlier detection cannot handle missing values, the specific number of missing values in each data set is first counted. For all missing data in the same attribute, the sample mean in this attribute is used to fill in.

(2) Data type conversion: When we process numeric data, if we encounter non-numeric categories in the data, we can convert them to numbers, and we use LabelEncoder in sklearn.

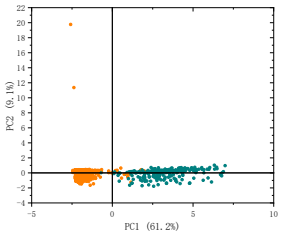
(3) Downsampling: used for data sets with larger size, outliers account for less in the data. The purpose of downsampling is to select a part of the data from the majority set and reassemble the minority set into a new data set to solve the uneven distribution of data.

The WBC data set is Wisconsin-Breast Cancer, which records the measurement results of breast cancer cases. There are two types: benign and malignant. This data set was down-sampled to 21 points, and malignant tumors were regarded as outliers. Ionosphere is the original data of the ionosphere, a binary data set with a dimension of 34, and all the attributes with zero values are deleted, so the total dimension is 33. "Bad" is an outlier cluster, and "good" is a normal cluster. The Wine data set is the result of chemical analysis of three different varieties of wine. The first type is down-sampled to 10 points and considered as outliers, and the rest are normal points. Glass contains several types of glass, of which category 6 is an obvious minority and is considered an outlier. Yeast is a yeast data set that predicts the cell localization sites of proteins and contains 64 outliers. Cardio imaging (Cardio) dataset includes measurements of fetal heart rate (FHR) and uterine contraction (UC) features of cardiography classified by obstetric experts. This is a classification data set, classification is divided into normal, suspicious and pathological. For outlier detection, normal classes form inliers, while pathological classes are down-sampled to 176 points, which are considered outliers. The suspect class is discarded. The basic information of the data set is shown in Table 3.

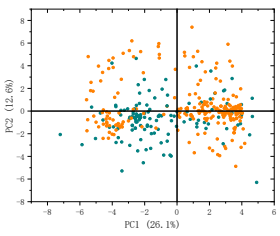
*Table 3: Basic information of experimental data set*

| dataset | Number of samples | dimension | ouliers | Outlier ratio |
|---|---|---|---|---|
| WBC | 278 | 30 | 21 | 5.6% |
| Ionosphere | 351 | 33 | 126 | 36% |
| Wine | 129 | 13 | 10 | 7.7% |
| Glass | 214 | 9 | 9 | 4.2% |
| Yeast | 1364 | 8 | 64 | 4.7% |
| Cardiotocogrpahy | 1831 | 21 | 176 | 9.6% |

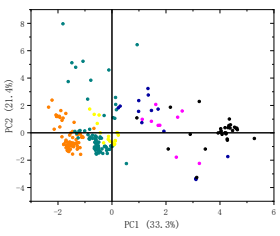
It can be seen from the table that the number of attributes of the data set, that is, the dimension of the data set is relatively large, and it is difficult to visualize the data in a high-dimensional space, so dimension reduction is required to visually reflect the distribution of the data set. The two data sets are displayed in two dimensions using principal component analysis. As shown in Figure 5.




*(a)WBC*

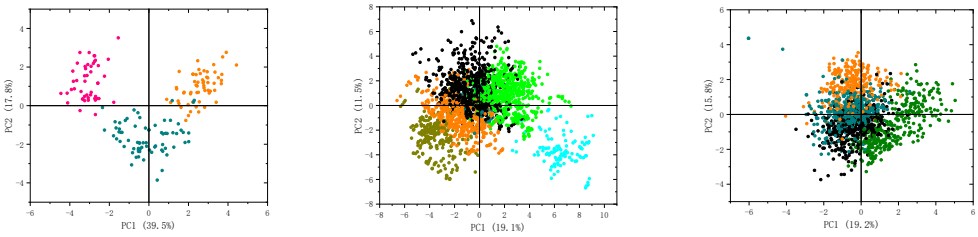



*(b)Ionosphere*




*(c)Wine*

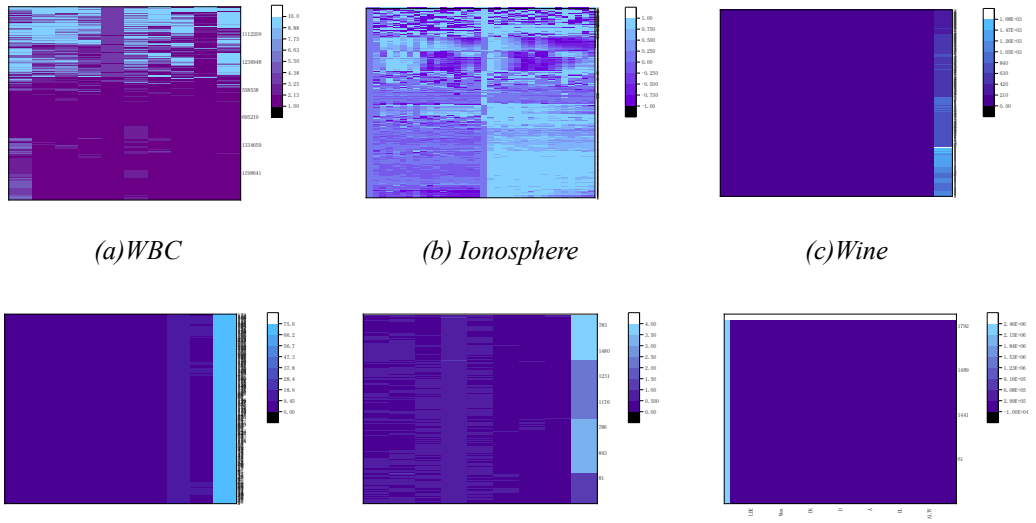
ISSN 2616-5775 Vol. 6, Issue 13: 82-91, DOI: 10.25236/AJCIS.2023.061313




*(d) Glass* *(e)Yeast* *(f) Cardiotocogrpahy*

*Figure 5: Distribution of the data set.*

In the figure we can clearly see the distribution of the data set. In order to visually show the distribution or density of data, we use Heatmap to display the relevant attributes of the WBC, Arrhythmia, Glass, Wine, Yeast, Cardiotocogrpahy datasets. The data set shown in Figure 6 can directly provide reference for selecting the attributes representing all variables for principal component analysis. The heat map shows the areas in the data set and the number of attributes in a special highlight. The more concentrated the number of attributes in a certain area, the darker the color of the data identification on the map, which can be understood as the difference in attributes; On the contrary, if the data of the attributes in the area is relatively small, the lighter the color in the figure, then the data distribution of the attributes in this area is sparse. The heat map can not only observe the characteristic status of each attribute in the data set more clearly, but also observe in which area the similar characteristics of the object attribute in the data set are concentrated.

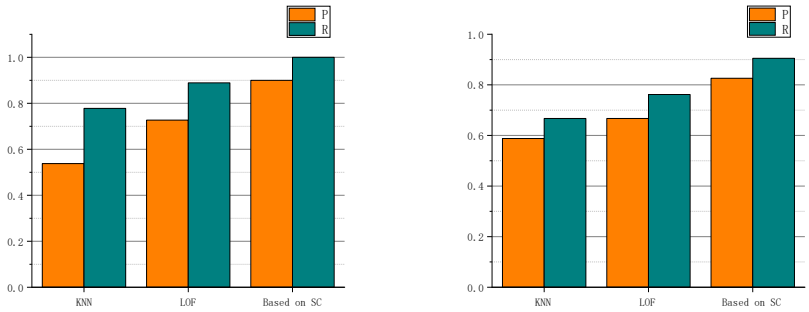


*(a)WBC* *(b) Ionosphere* *(c)Wine*

*(d)Glass* *(e)Yeast* *(f)Cardiotocogrpahy*

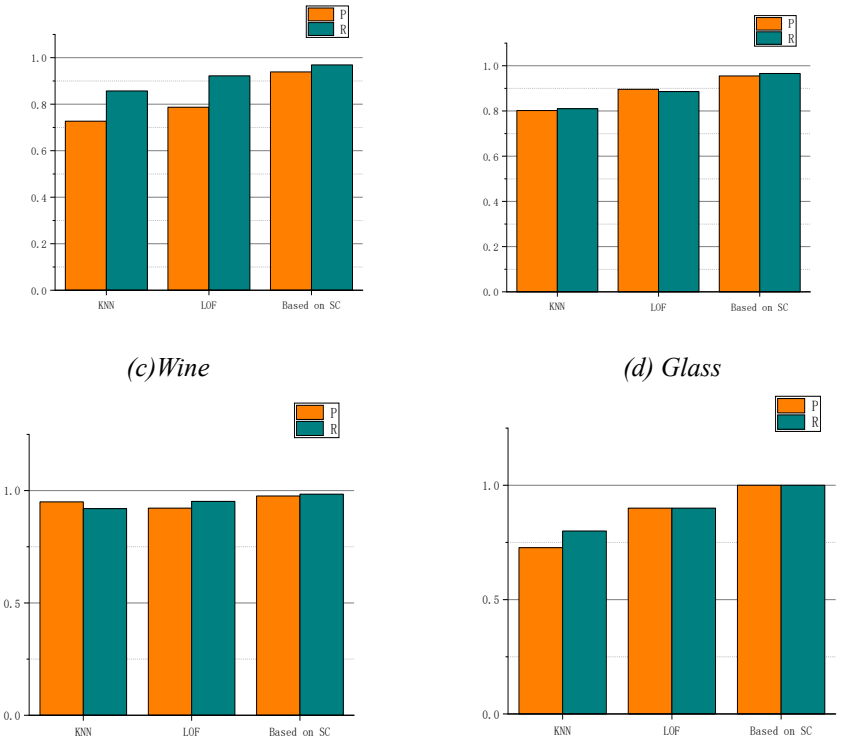
*Figure 6: Heatmap of the dataset.*

Analyze KNN, LOF, and spectral clustering-based algorithms through the accuracy rate P and recall rate R to evaluate the actual effects of the three algorithms on these six data sets.




*(a)WBC* *(b)Inosophere*

ISSN 2616-5775 Vol. 6, Issue 13: 82-91, DOI: 10.25236/AJCIS.2023.061313




*(c)Wine* *(d) Glass*

*(e)Yeast* *(f) Cardiotocogrpahy*

*Figure 7: The detection effect of the three algorithms on different data sets.*

It can be seen from Figure 7 that the outlier detection algorithm based on spectral clustering is superior to the other two algorithms in the accuracy of these six data sets, and it is more sensitive to outliers and can identify most outliers. . The KNN and LOF algorithms are greatly affected by the data dimension. When the data dimension is high, the algorithm effect is not ideal.

In the outlier detection algorithm, the difference in the data set is relatively large, and the evaluation algorithm such as accuracy rate, false detection rate or recall rate cannot be expressed very accurately, so we choose F1-score as the evaluation index. F1-score is the harmonic average of precision and recall. When F1 is higher, it can indicate that the algorithm is more effective.

$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{M + TP - TN} \tag{16}$$

Where: P is Precision, R is Recall, TP is True Positive, TN is True Negative, and M is the total number of samples. The accuracy rate is and the recall rate is. The confusion matrix includes TP, TN, false positive rate FP (False Positive), false negative rate FN (False Negative), and the formula for calculating the total number of samples is $M = TP + TN + FP + FN$. Figure 8 shows the F1 curve of three algorithms for detecting outliers in six data sets.

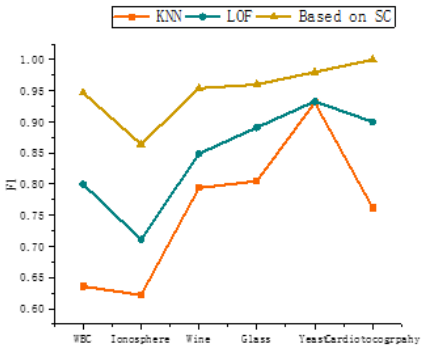



*Figure 8: F1-score of the data set.*

ISSN 2616-5775 Vol. 6, Issue 13: 82-91, DOI: 10.25236/AJCIS.2023.061313

It can be seen from Fig. 8 that the advantages of outlier detection algorithm based on spectral clustering are obvious and relatively stable. The KNN algorithm and LOF algorithm have large differences in the detection of outliers in different data sets, and the effect is poor.

## 4. Conclusions

This paper presents an effective outlier detection method based on adaptive spectral clustering, which can detect global and local outliers at the same time, and solves the problems that spectral clustering has a large dependence on parameters and the number of clusters is difficult to determine. This algorithm uses self-tuning spectral clustering algorithm to re-improve the scale parameters, while improving the k-means algorithm to make the clustering stable. The definition of size clustering and the outlier index are used to identify outliers. Experiments with artificial data sets and real data sets verify the feasibility and stability of the algorithm.

## References

*[1] Huang Qiang, Ye Qing, Nie Bin, et al. Research on Outlier Identification Methods [J]. Software Guide, 2019 (6): 35-41.*
*[2] Zelnik Manor L, Perona P. Self-connecting spectral clustering [C] Advancements in Neural Information Processing Systems (NIPS) Cambridge, MA: MIT Press, 2004.*
*[3] Liu Xiangxiang, Zhang Linlin, Zhu Siqiao, et al. A Multi form Flexible Resource Aggregation Model for Virtual Power Plants Based on Grey Target Theory and Spectral Clustering [J]. China Electric Power, 2023, 56 (11): 104-112.*
*[4] Jean Claude C. From spectral cluster to uniform solvent estimates on compact manifolds [J]. Journal of Functional Analysis, 2024286 (2).*
*[5] Kunshuo X, Ge T, Boyu L, et al. Sub Classification and Type Identification of Weathering Silicate Glasses Based on Spectral Clustering [J]. Journal of Physics: Conference Series, 20232607 (1).*
*[6] Hongyuan Z, Xuelong L. Discretize Relaxed Solution of Spectral Clustering via a Non European Algorithm [J]. IEEE transactions on neural networks and learning systems, 2023.*
*[7] Xiao G, Hai Z, Xiangyu C. On the efficiency of higher order spectral clustering under weighted stochastic block models [J]. Computational Statistics and Data Analysis, 2024190.*
*[8] Li Fangshuo, Liu Lina, Cheng Zhijiong, et al. Low voltage household transformer relationship recognition based on improved spectral clustering algorithm [J]. China Test, 2023, 49 (10): 128-134.*
*[9] Ma Zhaohui, Hou Jichao, Xie Chengxin, et al. Research on Spectral Clustering Algorithm Based on User Characteristics [J]. Information Technology and Informatization, 2023, (10): 104-108.*
*[10] Guo Yingya, Wang Lijuan, Geng Haijun. Edge Server Placement Algorithm Based on Spectral Clustering [J]. Computer Science, 2023, 50 (10): 248-257.*
*[11] Jiayi T, Yuping L, Xinwang L. Multiview Spectral Clustering Based on Consensus Neighbor Strategy [J]. IEEE transactions on neural networks and learning systems, 2023, PP.*