Journal of Physics: Conference Series

PAPER • OPEN ACCESS

An Improved K-Means Algorithm Based on Kurtosis Test

To cite this article: Tingxuan Wang and Junyao Gao 2019 *J. Phys.: Conf. Ser.* **1267** 012027

An Improved K-Means Algorithm Based on Kurtosis Test

Tingxuan Wang[*] and Junyao Gao

Beijing Institute of Technology, 5 South Zhongguancun Street, Haidian District, Beijing, China.

[*]460336022@qq.com

Abstract. Clustering is a process of classifying data into different classes and has become an important tool in data mining. Among many clustering algorithms, the K-means clustering algorithm is widely used because of its simplicity and high efficiency. However, the traditional K-means algorithm can only find spherical clusters, and is also susceptible to noise points and isolated points, which makes the clustering results affected. To solve these problems, this paper proposes an improved K-means algorithm based on kurtosis test. The improved algorithm can improve the adaptability of clustering algorithm to complex shape datasets while reducing the impact of outlier data on clustering results, so that the algorithm results can be more accurate. The method used in our study is known as kurtosis test and Monte Carlo method. We validate our theoretical results in experiments on a variety of datasets. The experimental results show that the proposed algorithm has larger external indicators of clustering performance metrics, which means that the accuracy of clustering results is significantly improved.

1. Introduction

Data mining is a hot research direction nowadays. It is the process of mining valuable information and knowledge from a large number of irregular data [1]. Now, clustering is an important exploratory analysis method in data mining and a statistical analysis method for classification problems. It is a method that divides some physical or abstract objects into several clusters, which satisfies the conditions that the similarity of objects in the same cluster is high, while the similarity of objects in the different clusters is low.

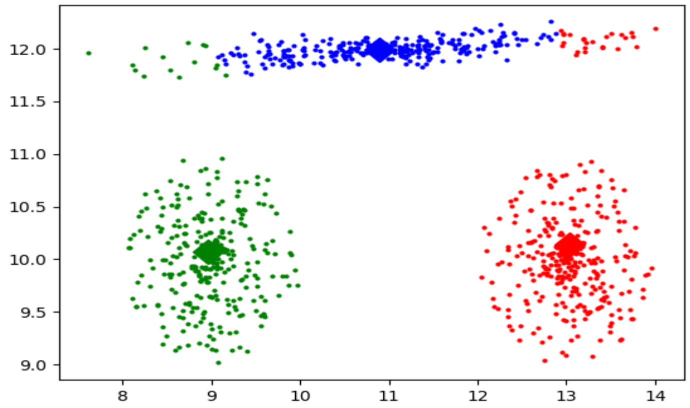

Figure 1. Limitations of the traditional K-means algorithm.

The K-means algorithm is one of the most classical clustering algorithms. This algorithm places the data to be clustered in the European space and considers that each cluster is composed of the data which is close to each other. The algorithm is widely used because of its simplicity, ease to be understood, high efficiency, and its applicability to large datasets [2]. However, the K-means algorithm also has some defects: for example, it is susceptible to noise points and isolated points. At the same time, the Euclidean distance characteristic determines that the K-means algorithm can only be applied to spherical datasets with similar size and density. Figure 1 shows the limitations of the traditional K-means algorithm.

Aiming at the limitation of the traditional K-means algorithm, researchers have made some attempts [3]. In the literature [4], the author uses the I-divergence measure and the Max Entropy measure as similarity measure index, which makes the algorithm suitable for datasets of various complex and irregular shapes. In the literature [5], the author proposes an improved outlier detection method in the iteration process of updating cluster centers, which improves the accuracy of clustering. In the literature [6], based on the original K-means, the author gives a weight to each dimension of the eigenvector. In this way, for the dimension of noise, its weight should be very small, so that the overall cost function value will be lowest. To sum up, in view of the limitations of the traditional K-means algorithm, domestic and foreign scholars have made a lot of improvement studies and achieved some remarkable results. However, the most obvious disadvantage of the current research is that most of the improved algorithms only focus on the improvement of one aspect of the traditional K-means algorithm, which results in the improvement effect of K-means accuracy rate is not obvious. In the process of reading relevant literature, we learn that data kurtosis is a measure of outlier degree of outlier data [7]. The larger the outlier degree of outlier data is, the larger the kurtosis value of the whole data is. We decided to apply this conclusion to the improvement of the K-means algorithm.

In this paper, we propose an improved K-means algorithm based on kurtosis test. This method performs kurtosis test on the traditional K-means running results, and finds the "abnormal data" that are mistakenly divided into each cluster. Then, the kurtosis calculation is used to find suitable clusters for these data. For those data which can’t find the appropriate cluster, we consider them to be outlier data in the dataset. In theory, the method can take into account both the shape of the dataset and the influence of the outlier data on the clustering results, so that the accuracy of the algorithm can be improved. The comparison experiments show that compared with the results in other literatures, the external indicators of clustering performance metrics of the new algorithm are improved. The proposed algorithm is advanced and innovative. The results of this paper can be applied to improve the clustering effect and classification accuracy.

The rest of this paper is organized as follows: Second 2 introduces the principle and method of normality test based on kurtosis, Monte Carlo simulation experiment and the Kurtosis-Kmeans algorithm flow. In Second 3, the experimental results are presented, which show the superiority of the new algorithm. Second 4 is the summary and conclusion.

2. Methods

According to the literature [7], data kurtosis is a measure of outlier degree of outlier data. We assume that when the number of samples is large enough, the samples in each cluster obtained by clustering algorithm are approximately normal distribution, which is also in line with our objective cognition. In this way, we can apply statistical normality test based on kurtosis to K-means clustering algorithm [8]. From Section I, we know that for complex and irregular shape datasets, the traditional K-means algorithm does not work well, as shown by dividing many data originally belonging to another cluster into the wrong cluster according to the Euclidean distance. These error-divided data are abnormal for the cluster, like the outlier data, which will cause the sudden change of the kurtosis value of the cluster. According to this character, this paper proposes an improved K-means algorithm, which can detect these "abnormal data" by using kurtosis value, and then classify these data correctly and find out the true outlier data among them, until we think that all the clusters obey the multi-dimensional normal distribution.

2.1. Multi-dimensional normality test based on M-type multi-dimensional kurtosis

The multi-dimensional normality test of the sample is an important basis for determining whether the algorithm still needs to be iterated. However, there is no uniform definition of multi-dimensional kurtosis. At present, many statisticians have given the generalization of one-dimensional kurtosis in multi-dimensional cases. This paper intends to adopt the definition of multi-dimensional kurtosis proposed by statistician Mardia. In 1970, Mardia first gave a generalization of one-dimensional kurtosis in multi-dimensional cases, called M-type multi-dimensional kurtosis, and constructed multi-dimensional normality test statistics and their asymptotic distribution.

Let **X** denote a set of n points in R^{ρ}. Its mean $E(X)=\mu$, and its covariance matrix $\Sigma=E\left[(X-\mu)(X-\mu)^{T}\right]$ is non-singular matrix. We define equation (1) and equation (2) as the sample mean and sample covariance matrix respectively.

$$\bar{\chi}=\frac{1}{n}\sum_{i=1}^{n}\chi_i \tag{1}$$

$$S=\frac{1}{n}\sum_{i=1}^{n}\left(\chi_i-\bar{\chi}\right)\left(\chi_i-\bar{\chi}\right)^{T} \tag{2}$$

After deducing the formula, Mardia obtains the measure of M-type multidimensional kurtosis β_{ρ}, as shown in equation (3).

$$\beta_{\rho}=E\left[(X-\mu)^{T}\Sigma^{-1}(X-\mu)\right]^{2} \tag{3}$$

The corresponding M-type multi-dimensional sample kurtosis is shown in equation (4).

$$b_{\rho}=\frac{1}{n}\sum_{i=1}^{n}\left[\left(\chi_i-\bar{\chi}\right)^{T}S^{-1}\left(\chi_i-\bar{\chi}\right)\right]^{2} \tag{4}$$

Mardia not only gives the definition of multidimensional kurtosis, but also obtains the asymptotic distribution of related statistics in multi-dimensional normal case, as shown in equation (5).

$$M_2=\frac{b_p-\rho(\rho+2)}{\left[8\rho(\rho+2)n^{-1}\right]^{1/2}}\xrightarrow{D}N(0,1)(n\to\infty) \tag{5}$$

According to the multidimensional kurtosis definition and related statistics proposed by Mardia, we can get the method of multidimensional normality test: take the significance level α, compare the statistics M_2 and the upper α quantile of the corresponding distribution, when the former is larger than the latter, reject the hypothesis: $H_0:X\sim N_{\rho}(\mu,\Sigma)$. In addition, we can also test the hypothesis: $H_0:\beta_{\rho}=\rho(\rho+2)$ directly. The decision to accept or reject the hypothesis is made by calculating b_{ρ} and comparing it with the significance level quantile of the corresponding empirical distribution.

This paper intends to use the method of using the significance level quantile of the empirical distribution as the test method. The advantage of this method is that it can be used in the case of large samples and small samples, while the method of using the asymptotic distribution of statistics can only be used in the case of large samples. Obviously, the former is more suitable for the application of clustering algorithm.

2.2. Monte Carlo method

The Monte Carlo method is a very important numerical method. It is based on a probabilistic model, and the results of the simulation experiments are used as approximate solutions to the problem according to the process described by the model [9]. In this paper, for the purpose of obtaining the upper α quantile of the b_{ρ} empirical distribution, the Monte Carlo simulation is used.

Firstly, 10,000 sets of samples from the normal distribution $N_{\rho}(0, I)$ with sample size n are generated by using MATLAB. Then, each set of samples is substituted into the expression of b_{ρ}, and the value of b_{ρ} corresponding to each set of samples is obtained. The 10,000 sets of samples correspond to the values of 10000 b_{ρ}. What we need is the upper α quantile of the b_{ρ} empirical distribution. It is known from its probability meaning that there should be 10,000α samples in the

10,000 sets of samples larger than this value. Finally, the values of the 10000 b_ρ are arranged in descending order, and the value of the sequence number 10000α can be approximated as the upper α quantile of the b_ρ empirical distribution. Table 1 shows the upper 5% quantile of the b_ρ empirical distribution.

Table 1. The upper 5% quantile of the b_ρ empirical distribution.

ρ	n	The upper 5% of the b_ρ	ρ	n	The upper 5% of the b_ρ
2	50	9.46	4	50	25.88
	100	9.17		100	25.74
	150	8.99		150	25.58
	180	8.93		180	25.42
	300	8.73		300	25.15
	1000	8.41		1000	24.69
ρ	n	The upper 5% of the b_ρ	ρ	n	The upper 5% of the b_ρ
3	50	16.70	13	50	193.73
	100	16.49		100	196.75
	150	16.31		150	197.32
	180	16.18		180	197.36
	300	15.98		300	197.29
	1000	15.55		1000	196.69

2.3. The proposed algorithm flow

With the method of multi-dimensional normality test and the upper α quantile of b_ρ empirical distribution obtained by Monte Carlo method, we propose an improved K-means algorithm based on kurtosis test. The proposed algorithm flow is as follows.

Algorithm 1. An improved K-means algorithm based on kurtosis test.

```
Input:  sample set $D=\{\chi_1,\chi_2,\dots,\chi_m\}$, cluster number k.
Output: k clusters, outlier dataset $C'$.
use the traditional K-means algorithm to get the initial k clusters $C=\{C_1,C_2,\dots,C_k\}$;
initialize the outlier dataset $C'=\varnothing$;
repeat
    for i=1; i ≤ k do
        calculate the kurtosis value of $C_i$ ——Kurtosis($C_i$);
        find the upper α quantile of the bρ empirical distribution according to table 1;
        if Kurtosis($C_i$)>bρ then
            for j=1; j ≤ |$C_i$| do
                set threshold $\varepsilon_1$, and find out the data $C_{i_j}$ that will make the kurtosis of the cluster $C_i$ suddenly change in the direction of decreasing;
                $C_i=C_i-\{C_{i_j}\}$, $C'=C'\cup\{C_{i_j}\}$;
            end for
        end if
    end for
```

```
        for p=1; p ≤ $|C'|$ do
            for q=1; q ≤ k do
                set threshold $\varepsilon_2$;
                if $C'_p$ doesn't make the kurtosis of cluster $C_q$ suddenly change then
                    $C' = C' - \{C'_p\}$, $C_q = C_q \cup \{C'_p\}$;
                end if
            end for
        end for
    until the kurtosis values of all clusters are less than the upper α quantile of the $b_\rho$ empirical distribution
```

3. Results

3.1. Experimental environment and datasets

The experimental hardware environment is Intel(R) Core(TM) i5-3320M CPU @ 2.60GHz, 8G RAM, and the software environment is python3.7. The datasets used in the experiment are simulation dataset containing elliptical data, simulation dataset containing outlier data and Iris and Wine datasets in UCI database. The simulation dataset two contains eight outlier data. Table 2 shows the information of the four datasets mentioned above.

Table 2. Experimental datasets.

No.	Dataset	Number of instances	Number of attributes	Size of classes
1	Simulation dataset one	900	2	300,300,300
2	Simulation dataset two	908	2	300,300,300,8
3	Iris	150	4	50,50,50
4	Wine	178	13	59,71,48

3.2. Simulation experiment results and analysis

In order to verify that the Kurtosis-Kmeans algorithm can break through the defect that the traditional K-means algorithm is only applicable to spherical datasets with similar size and density, it is applied to the simulation dataset one. Figure 2 shows the running process of Kurtosis-Kmeans algorithm and the results of each step. Red, green and blue point sets represent clustering results, and black point set represents outlier dataset.

It can be seen from figure 2 (e) that after the Kurtosis-Kmeans algorithm is iterated twice, compared with figure 2 (a), most of the data that are incorrectly divided due to the limitations of Euclidean distance are divided into the correct clusters. The experimental results show that the Kurtosis-Kmeans algorithm is more suitable for complex and irregular shape datasets, which improves the clustering accuracy.

In order to verify that the Kurtosis-Kmeans algorithm can break through the defect that the traditional K-means algorithm is susceptible to noise points and isolated points, it is applied to the simulation dataset two. Figure 3 shows the running process of Kurtosis-Kmeans algorithm and the results of each step. Red, green and blue point sets represent clustering results, and black point set represents outlier dataset.

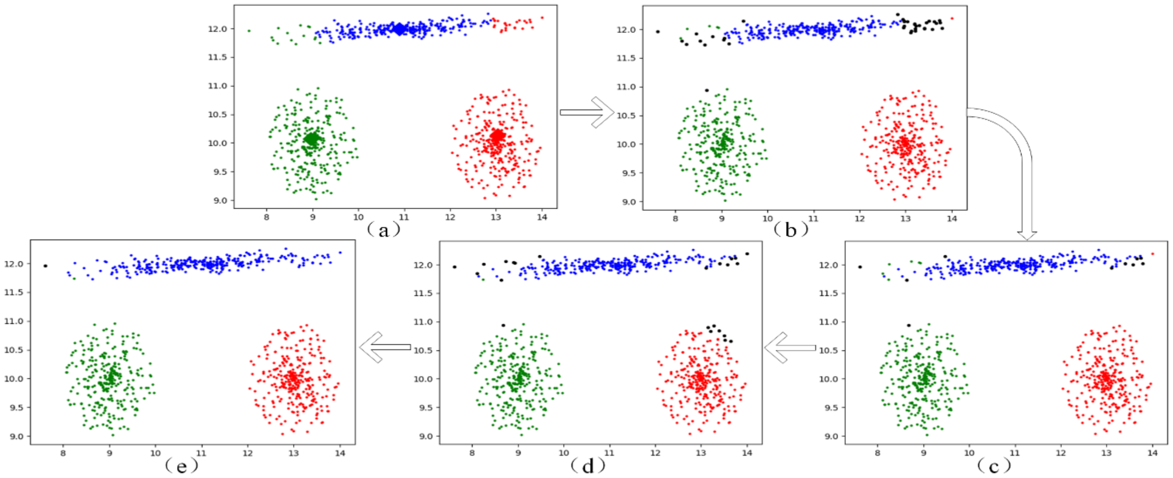

Figure 2. The running process of Kurtosis-Kmeans algorithm on simulation dataset one.

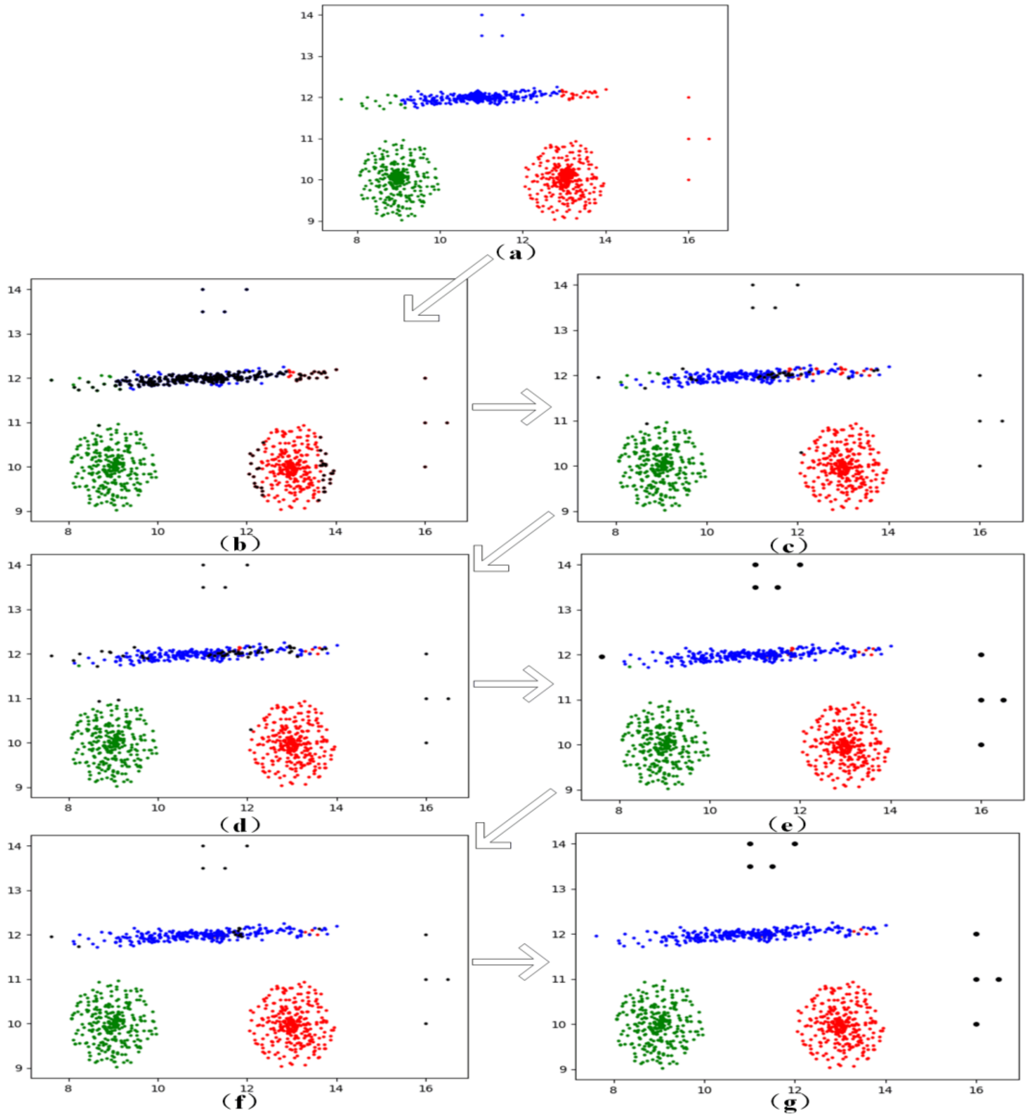

Figure 3. The running process of Kurtosis-Kmeans algorithm on simulation dataset two.

As can be seen from figure 3, due to the influence of the addition of outlier data, the kurtosis value of the cluster containing outlier data becomes sensitive. When looking for outlier data for the first time, many data are considered as outlier data, as shown in figure 3 (b). Therefore, in the following iteration, the threshold ε should be appropriately increased. After the Kurtosis-Kmeans algorithm is iterated three times, the figure 3 (g) is compared with the initial situation figure 3 (a), not only the majority of data are correctly divided as expected, but also the eight outlier data added are correctly detected. The experimental results show that the Kurtosis-Kmeans algorithm can detect the outlier data in the dataset, which improves the clustering effect.

3.3. Comparisons with results in other researches

In order to verify the advancement of the new algorithm, in the same experimental environment, the traditional K-means algorithm, the ID-Kmeans algorithm in reference [4], the improved K-means algorithm for discrete values (DV-Kmeans) in reference [5], and the Kurtosis-Kmeans algorithm based on kurtosis test proposed in this paper are selected for comparative experiments. We apply the above four algorithms to the Iris and Wine datasets of the UCI database. Then, the external indicators of clustering performance metrics of the four algorithms Jaccard Coefficient (JC), Fowlkes and Mallows Index (FMI), and Rand Index (RI) are calculated and compared respectively [10]. The result values of the above performance metrics are all in the [0, 1] interval, and the larger the value is, the better the performance of the algorithm is.

The results of the external indicators of clustering performance metrics are shown in table 3 - table 5.

Table 3. JC of four algorithms in Iris and Wine datasets.

Dataset	Traditional K-means	ID-Kmeans	DV-Kmeans	Kurtosis—Kmeans
Iris	68.23%	71.00%	64.59%	73.85%
Wine	41.20%	41.44%	41.74%	45.94%

Table 4. FMI of four algorithms in Iris and Wine datasets.

Dataset	Traditional K-means	ID-Kmeans	DV-Kmeans	Kurtosis—Kmeans
Iris	81.12%	83.06%	78.57%	84.98%
Wine	58.35%	58.59%	58.89%	62.96%

Table 5. RI of four algorithms in Iris and Wine datasets.

Dataset	Traditional K-means	ID-Kmeans	DV-Kmeans	Kurtosis—Kmeans
Iris	87.37%	88.59%	85.15%	89.88%
Wine	71.87%	72.04%	72.21%	75.11%

As can be seen from table 3 - table 5, the Kurtosis—Kmeans algorithm has improved the external indicators of clustering performance metrics Jaccard Coefficient, Fowlkes and Mallows Index and Rand Index compared with the other three algorithms. The comparison results show that the Kurtosis-Kmeans algorithm increases the similarity between the clustering results and the original samples, which means that the accuracy of the Kurtosis-Kmeans algorithm is higher, and the effect is better than the other three algorithms. The experimental results reflects the advancement of the Kurtosis—Kmeans algorithm.

Through simulation and comparison experiments, it has been proved that the new algorithm proposed by us can improve the clustering effect and eliminate the influence of outlier data. It is advanced and innovative.

4. Conclusion

As a classical unsupervised learning algorithm, the improvement of K-means algorithm is of great significance. In this paper, we use the statistical multi-dimensional normality test method to improve the clustering effect, and propose an improved K-means algorithm based on kurtosis test. This algorithm can improve the adaptability of clustering results to complex shape datasets while reducing the impact of outlier data on clustering results. The experimental results show that the Kurtosis-Kmeans algorithm has larger external indicators of clustering performance metrics than other algorithms, and has achieved good clustering results. The effectiveness and advancement of the Kurtosis-Kmeans algorithm are also demonstrated by the experimental results. In future work, more attention will be paid to the threshold setting in the process of the algorithm. In addition, using the kurtosis value directly as a dynamic indicator in the algorithm iteration process is also a positive research direction and is worth researching.

References

[1] Huang Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values[J]. *Data Mining and Knowledge Discovery*, 1998, **2**(3): 283-304.

[2] Jain A K. Data clustering: 50 years beyond K-means[J]. *Pattern recognition letters*, 2010, **31**(8): 651-666.

[3] Fahim A M, Salem A M, Torkey F A, et al. An efficient enhanced k-means clustering algorithm[J]. *Journal of Zhejiang University Science A*, 2006, **7**(10): 1626-33.

[4] Li H. Improved K-means clustering method and its application[D]. Northeast Agricultural University, 2014.

[5] Zhu J. Research and Application of K-means algorithm[D]. Dalian University of Technology, 2013.

[6] Huang J Z, Ng M K, Rong H, et al. Automated variable weighting in k-means type clustering[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, **27**(5): 657-668.

[7] Xiao-Ran Z. Discussions of kurtosis' statistical meanings[J]. *Journal of Yanshan University*, 2006, **30**: 57-60.

[8] Tian Y. Tests for normality based on Skewness and Kurtosis[D]. Shanghai Jiao Tong University, 2012.

[9] Pagès G. The Monte Carlo Method and Applications to Option Pricing[M]//Numerical Probability. Springer, Cham, 2018: 27-47.

[10] Zhou Y, Ji C, Zhang C. A Clustering Algorithm based on Internal Constrained Multi-view K-means[C]//Proceedings of the 12th Chinese Conference on Computer Supported Cooperative Work and Social Computing. ACM, 2017: 137-144.