Clustering Validity Evaluation Method Based on Two Typical Clustering Algorithms

Guan Wang, Cheng Xing *, Jie-Sheng Wang, Hong-Yu Wang, Jia-Xu Liu

Abstract—**Cluster analysis is one of the important methods of data research. The rationality of cluster effect and the determination of cluster number play a very important role in the data analysis. The study of clustering validity establishes the clustering validity index and adopts appropriate clustering algorithm in different data areas to obtain the optimal division. Based on the K-means clustering algorithm and partition around medoids (PAM) clustering algorithm, 11 evaluation indexes of clustering effectiveness were studied and their data were verified, namely Rand index (Ri), Adjusted Rand index (Ari), Mirkin index (Mi), Hubert index (Hi), Silhouette index (SiL), Davies-Bouldin index (DB), Calinski-Harabasz index (CH), Krzanowski-lai index (KL), Hartigan index (HI), Weighted inter- to intra-cluster ratio (Wint) and Homogeneity-Separation (HS). Finally, the K-means clustering algorithm, PAM clustering algorithm and 11 clustering validity indexes were used to carry out simulation experiments on the leuk72_3K data set and wine data set, and the performance of different clustering algorithms and validity functions were compared.**

Index Terms—**K-means clustering algorithm, PAM clustering algorithm, clustering validity index, optimal cluster number**

I. INTRODUCTION

CLUSTER is one of the important research contents in the fields of pattern recognition, data mining, machine learning and so on, and plays an extremely important role in identifying the inherent structure of data [1]. Based on the principle of birds of a feather clustering, clustering classifies a group of individuals into several categories according to similarity, so that the differences between individuals of the same category are as small as possible, while the differences between individuals of different classes are as large as possible [2]. According to the similarity measure and clustering evaluation criteria, clustering algorithms can be divided into two categories: hard clustering algorithm and fuzzy clustering algorithm. K-means clustering algorithm is one of the most classical and widely used hard clustering algorithms, which has reliable theory, simple algorithm, fast convergence speed and can effectively deal with a large number of data sets [3]. Its clustering is based on the idea of "zero equals one", and it can clearly divide each sample into different sub-classes. However, the traditional K-means clustering algorithm excessively relies on the initial conditions, such as the number of clusters k value needs to be given in advance, the clustering results depend on the initial cluster center, and different sample input order will change the clustering results. Because of this, K-means clustering algorithm is always new. For example, in 2018, Yu et al. proposed two improved K-means algorithms, which greatly improved the efficiency of the traditional K-means clustering algorithm [4]. Sinaga et al. proposed a new unsupervised K-means clustering algorithm in 2020 [5]. In 2021, Rezaee et al. proposed GBK-means algorithm based on the improvement of K-means algorithm [6]. Partition Around Medoids (PAM) algorithm can better optimize the cluster center. In order to judge the optimal clustering number of data sets and improve the quality of clustering results, the research on clustering validity has become an important branch of clustering problems and attracted the attention of scholars at home and abroad. At present, many clustering validity functions have been proposed and used. In 2019, Zhu et al. proposed a clustering validity function based on ratio form [7]. In 2021, Wang et al. proposed a new validity function based on intra-class compactness and inter-class separation [8]. However, no clustering validity function can be applied to all data sets due to the ever-changing structure of data sets and different attributes and sizes.

In order to achieve the optimal division of cluster number, a good cluster validity index is indispensable. After years of in-depth research, scholars can divide these indexes into three categories: internal validity index, external validity index and relative validity index. Among them, internal validity indexes are evaluated on the premise of uncertain classification, such as Rand index (Ri), Adjusted Rand index (Ari), Mirkin index (Mi), Hubert index (Hi), etc [9-10]. External validity index refers to evaluating the performance of different clustering algorithms by comparing the division of clustering with the matching degree of external criteria on the premise that the external structure information of the data set is available, such as Silhouette index (SiL), Davies-Bouldin index (DB), Calinski-Harabasz index (CH),

Manuscript received October 28, 2021; revised March 4, 2022. This work was supported by the Basic Scientific Research Project of Institution of Higher Learning of Liaoning Province (Grant No. LJKZ0293), and the Project by Liaoning Provincial Natural Science Foundation of China (Grant No. 20180550700).

Guan Wang is a postgraduate student of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China (e-mail: 480433838@qq.com).

Cheng Xing is a Ph.D candidate in School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114044, P. R. China (Corresponding author, phone: 86-0412-2538246; fax: 86-0412-2538244; e-mail: xingcheng0811@163.com).

Jie-Sheng Wang is a professor of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China (e-mail: wang_jiesheng@126.com).

Hong-Yu Wang is a postgraduate student of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China (e-mail: wanghongyuww@126.com).

Jia-Xu Liu is a postgraduate student of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China (e-mail: 1849226542 @qq.com).

Krzanowski-Lai index (KL), Hartigan index (HI), Weighted inter- to intra-cluster ratio (Wint) and Homogeneity-Separation index (HS) [11-12]. The relative validity index is to evaluate the different coefficient settings of the clustering algorithm so as to select the most suitable parameter settings. This paper introduces two common clustering algorithms: K-means clustering algorithm and PAM clustering algorithm. Then 11 typical clustering validate indexes were used to carry out simulation experiments on the leuk72_3K data set and wine data set, and finally the optimal clustering number was obtained.

II. Data Clustering Algorithm

A. K-means Clustering Algorithm

K-means clustering algorithm is the most classical algorithm based on distance, which adopts distance as the size of similarity. In other words, the smaller the distance between two objects, the more ideal the effect will be [13]. The K-means clustering algorithm process is described as follows.

(1) Randomly select K from the sample data as the clustering centers at the beginning.

(2) Calculate the value of the distance between each object and the center of each cluster, and divide them into the cluster with the smallest distance.

(3) After all objects complete this step, K cluster centers are calculated again.

(4) Observe the K cluster centers calculated in the first time. If the cluster centers change, divide the objects again; Otherwise, output the results.

(5) Stop and output the final result when the center no longer changes.

(1) Data types and metrics

The K-means clustering algorithm generally requires three distances. 1) Measure the distance between samples; 2) The distance between sample and cluster can be expressed by the distance $d(e_i, x)$; 3) The distance between clusters is expressed by the distance (e_i, e_j). The data matrix of n samples represented by p attributes is described as follows:

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

The Euclidean distance can be calculated by:

$$d(i,j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\ldots+(x_{ip}-x_{jp})^2} \quad (1)$$

The Manhattan distance can be calculated by:

$$d(i,j)=\left|x_{i1}-x_{j1}\right|+\left|x_{i2}-x_{j2}\right|+\ldots+\left|x_{ip}-x_{jp}\right| \quad (2)$$

The Minkowski distance can be calculated by:

$$d(i,j)=\sqrt[q]{(\left|x_{i1}-x_{j1}\right|)^q+(\left|x_{i2}-x_{j2}\right|)^q+\ldots+(\left|x_{ip}-x_{jp}\right|)^q} \quad (3)$$

where, q is a non-negative integer. When q=1, it is the Manhattan distance; When q=2, it is the Euclidean distance.

Data distance $d(i,j)=\left|x_{i1}-x_{j1}\right|+\left|x_{i2}-x_{j2}\right|+\ldots+\left|x_{ip}-x_{jp}\right|$ is defined as a document-word matrix, as shown in Table 1. The calculation formula of distance similarity between two documents is defined as:

$$\mathrm{d}(\mathrm{i},j)=\cos(i,j)=\frac{\vec{i}g\vec{j}}{\left\|\vec{i}\right\|g\left\|\vec{j}\right\|} \quad (4)$$

(2) Objective function

The square sum of error (SSE) is taken as the objective function. According to these two different classifications, smaller classifications should be selected. The SSE calculation of continuous attributes is realized by:

$$SSE=\sum_{i=1}^{k}\sum_{x\in E_i} dist(e_i,x)^2 \quad (5)$$

The SSE calculation of document data can be realized by:

$$SSE=\sum_{i=1}^{k}\sum_{x\in E_i} \cos ine(e_i,x)^2 \quad (6)$$

The calculation of cluster center e_i of cluster E_i is described as follows.

$$e_i=\frac{1}{n_i}\sum_{x\in E_i} x \quad (7)$$

The symbols and meanings in the above formulas are listed in Table 2.

(3) Validity of K-means clustering

The basic idea of K-means clustering method to determine the most suitable number of clusters is described as follows. For a given data set, in the known cluster number search area, the clustering algorithm is run to generate clustering results with different number of clusters, the validity index is selected to analyze the clustering results, and the optimal cluster number is found according to the judgment of the results.

Table 1. Document data content

	lost	win	team	score	music	happy	sad	…	coach
Document 1	14	2	8	0	8	7	10	…	6
Document 2	1	13	3	4	1	16	4	…	7
Document 3	9	6	7	7	3	14	8	…	5

Table 2. Symbols meanings

Symbol	Meaning
K	Number of clustering clusters
E_i	The i-th a cluster
x	Object (sample)
e_i	Cluster center of cluster E_i
n	The number of samples in the dataset
n_i	The number of samples in the i-th cluster

Based on the clustering analysis, the clustering criterion function converges according to the important criteria of maximum similarity and minimum difference among a class, minimum similarity and maximum difference among different classes.

$$\mathrm{E}=\sqrt{\sum_{i=1}^{k}\sum_{p\in C_j}(|p-m_i|)^2} \tag{8}$$

where, E is the sum of the squares of errors, P is the data sample, and m_i is the average value of class C_i. Based on the above basic ideas, the distance evaluation function is constructed. Let the data sample set $T=\{m_1,m_2,\cdots,m_n\}$ and the number of clustering be K. Let $I=\{T,K\}$ be the clustering space.

$$\mathrm{D}_{out}=\sqrt{\sum_{i=1}^{k}|m_i-m|^2} \tag{9}$$

where, D_{out} is the distance between classes, m is the sample mean, and m_i is all the sample mean in class C_i. Let $I=\{T,K\}$ be the clustering space.

$$D_{in}=\sqrt{\sum_{i=1}^{k}\sum_{p\in C_i}|p-m_i|^2} \tag{10}$$

where, D_{in} is the clustering distance, p is any spatial object, and m_i is the mean of all samples in class C_i.

Let $I=\{T,K\}$ be the clustering space. When D_{out} is approximately equal to D_{in}, the clustering number is optimal. Therefore, the distance evaluation function is defined as:

$$\mathrm{F}(T,K)=\left|\frac{D_{out}}{D_{in}}-1\right| \tag{11}$$

When this function is applied, since D_{out} is the monotonically increasing function of cluster K and D_{in} is the monotonically decreasing function of cluster K, it can be seen that the defined distance evaluation function must have a minimum value. Therefore, when this minimum value is determined, it means that the spatial clustering effect is the best. Therefore, the most suitable choice of K is obtained by:

$$\min\{F(T,K)\}\quad(K=1,2,3.....N) \tag{12}$$

B. PAM Clustering Algorithm

The core idea of PAM clustering algorithm and K-means clustering algorithm is roughly similar in structure. The biggest difference is that when changing the cluster center, PAM clustering algorithm is to calculate the minimum number of clustering from every point of the cluster center to all other points to optimize the new cluster center. Although PAM clustering algorithm overcomes these problems, but it has a premise, that is to increase the time of cluster analysis. PAM clustering algorithm needs to shorten the clustering completion time. In this view, PAM can only use numerical clustering for small low-dimensional data, when n and k are small. The value k must be specified when used. The steps of PAM clustering algorithm are described as follows.

(1) Set the sample as $X\{x(1),x(2),...\}$.

(2) Firstly, K clustering centers are randomly selected from the data.

(3) Then remove the points outside the cluster center, calculate the distance to each cluster center, and divide the selected number to the sample point with the smallest distance from the sample center, so as to achieve the initial effect.

(4) Then calculate the minimum value of the distance to all other points of each class (excluding points outside the center of the class), and take this value point as the center of the initial cluster. After completing these works, the optimal adjustment of the first effect will be completed.

(5) Repeat Step 4 until the center of the first and second time does not change, and finally achieve the goal. The core differences between these two algorithms are mainly reflected in Step 4.

III. Clustering Validity Indicators

A. External Clustering Validity Indicators

(1) Rand index (Ri)

$$\mathrm{Ri}=\frac{a+b}{C_2^n samples} \tag{13}$$

where, C represents clustering results, a and b represent logarithms of data. R_i is between 0 and 1. The larger R_i is, the better the clustering effect is, which means that the optimal division can be achieved.

(2) Adjusted Rand index (ARi)

$$ARi=\frac{Ri-E(Ri)}{\max(Ri)-E(Ri)} \tag{14}$$

ARi is greater than -1 and less than 1. The larger the same value means the better the clustering effect.

(3) Hubert index (Hi)

$$Hi=\frac{\mathrm{S}(C)-S_{\min}(C)}{\mathrm{S}_{\max}(C)-S_{\min}(C)} \tag{15}$$

where , $S(C)=\sum_{C_i\in Cx_i}\sum_{x_j\in C_j}d(x_i,x_j)$, $S_{min}(C)=\sum\min(n_w)x_i, x_j=X\left[d(x_i,x_j)\right]$, $S_{max}(C)=\sum\max(n_w)x_i$

where, $S(C)=\sum_{C_i\in Cx_i}\sum_{x_j\in C_j}d(x_i,x_j)$, $x_j=X\left[d(x_i,x_j)\right]$, $S_{min}(C)=\sum\min(n_w)x_i$, $S_{max}(C)=\sum\max(n_w)x_i$.

The index is simple in structure and easy to calculate, but it also has some shortcomings. It ignores the separation and other measures, so it affects its performance. In addition, there are also Mirkin index (Mi) index and other indicators. During the simulation experiment, these four external indicators are mainly verified.

B. Internal Clustering Validity Indicators

(1) Silhouette index (SiL)

$$SiL=\frac{1}{n}\sum_{j=1}^{n}S_j \tag{16}$$

where, $S_j = \frac{b_j - a_j}{\max(a_j, b_j)}$, $a_j = \left(\frac{1}{n_i}\right)\sum_{l=1,2\ldots} d(X_j, X_l)$ is the average distance between sample X_j and class C_j, and the minimum distance $b_j = \left(\frac{1}{n_i}\right)\min_{(h=1,2,\ldots k; k\neq i)}\sum_{x_i \epsilon c_h} d(X_j, X_l)$ is between sample X_j and all classes. Eq. (16) is the form of standardized summation, which mainly reflects the compactness and separability of the data set.

(2) Davies-Bouldin index (DB)

$$DB = \frac{1}{\mathrm{N}}\sum_{i=1}^{N} D_i \tag{17}$$

$$D_i = \max_{i\neq j} R_{(i,j)} \tag{18}$$

$$R_{(i,j)} = \frac{S_i + S_j}{M_{(i,j)}} \tag{19}$$

$$DB = \frac{1}{n}\sum_{i=1}^{N}\max_{i\neq j}\left(\frac{S_i + S_j}{d(C_i, C_j)}\right) \tag{20}$$

where, S_i is the average value, which can be Euclidean distance; $d(C_i, C_j)$ refers to the distance between cluster i and cluster j.

(3) Calinski-Harabasz index (CH)

$$CH = \frac{Tr(S_b)/(k-1)}{Tr(S_w)/(n-k)} \tag{21}$$

$$Tr(S_b) = \sum_{i=1}^{k} n_i * d(V_i, \bar{V}) \tag{22}$$

$$Tr(S_w) = \sum_{i=1}^{k}\sum_{j=1}^{n} d(X_j, V_i) \tag{23}$$

where, CH indicator is the ratio of compactness and separation degree, n is the number of clustering, K is the sample size so far, $Tr(S_b)$ is the trace of different data areas, and $Tr(S_w)$ is the trace of data areas in the same interval.

(4) Homogeneity-Separation (HS)

$$HS = \frac{2}{k^2 - K}\sum_{i=1}^{k}\sum_{j=j+1}^{k}\left\|V_i - X_j\right\| \tag{24}$$

The higher the HS value is, the less ideal the clustering is. But this index does not take into account the effect of the same class.

(5) Weighted inter- to intra-cluster ratio (Wint)

$$W\,\mathrm{int} = 1 - \frac{1}{\sum_{i=1}^{k} n_i * \mathrm{int}\,er(i)}\sum_{i=1}^{k}\frac{n_i}{n - n_j}\sum_{j=1, i\neq j}^{k} n_j * \mathrm{int}\,er(i,j) \tag{25}$$

where, $\mathrm{int}\,er(i)$ represents the size of similarity within a class, and $\mathrm{int}\,er(i,j)$ represents the size of similarity between classes. In Eq. (25), the maximum value is the optimal number of clustering. In addition to the above listed indicators, there are KL indicators, HI indicators, and so on.

IV. Simulation Experiment and Result Analysis

In order to verify the effectiveness of K-means clustering algorithm and PAM clustering algorithm, leuk72_3K and Wine data sets were used for validation. A total of 11 cluster validity index functions were used, including Ri, ARi, Mi, Hi and other external indicators. SiL index, DB index, CH index, KL index, HI index, Wint index and HS index. Cluster simulation is carried out according to these indicators, and the obtained curves represents the number of classes. Simulation verification is carried out for two kinds of data sets according to the above listed indicators, and the most appropriate clustering effect are achieved by considering the validity of the maximum and minimum values of the values, as well as the targeted data sets.

A. Clustering Validity Verification Based on K-means Clustering Algorithm

Four external validity indexes and seven internal validity indexes were used for clustering analysis on the leuk72_3K dataset and wine dataset based on K-means clustering algorithm. When using these indicators, it should focus on the analysis of the maximum and minimum values of the indicators, according to which we can judge whether it is the best clustering effect. The function of K-means clustering algorithm is to cluster the original data matrix X into K class to minimize the sum of the Euclidean distance between the samples and the centers of gravity of the class.

(1) External validity indicators on leuk72_3K dataset

The simulation results of the leuk72_3K dataset based on K-means clustering algorithm and four external validity indicators are shown in Fig. 1. According to the simulation diagram in Fig. 1, only Mi is effective at the lowest point, while the other three indicators are effective at the highest point, that is to say that the optimal number of clusters is 3.

(2) Internal validity indicators on leuk72_3K dataset

The simulation results of the leuk72_3K dataset based on K-means clustering algorithm and 7 internal validity indexes are shown in Fig. 2.

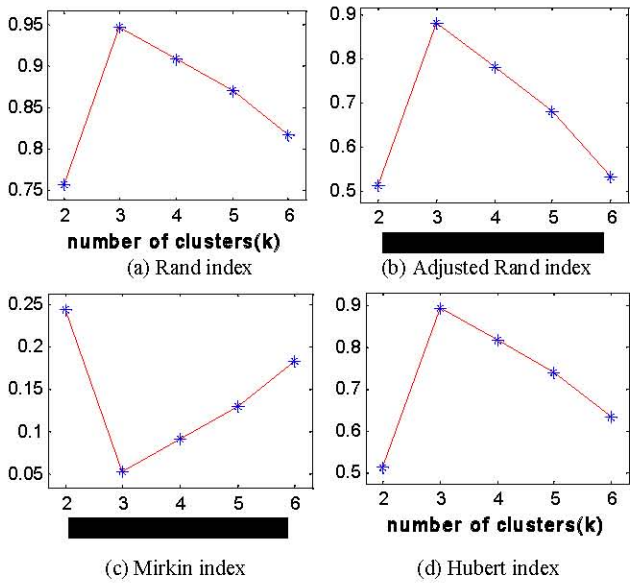


Fig. 1 External validity indicators under leuk72_3K dataset.

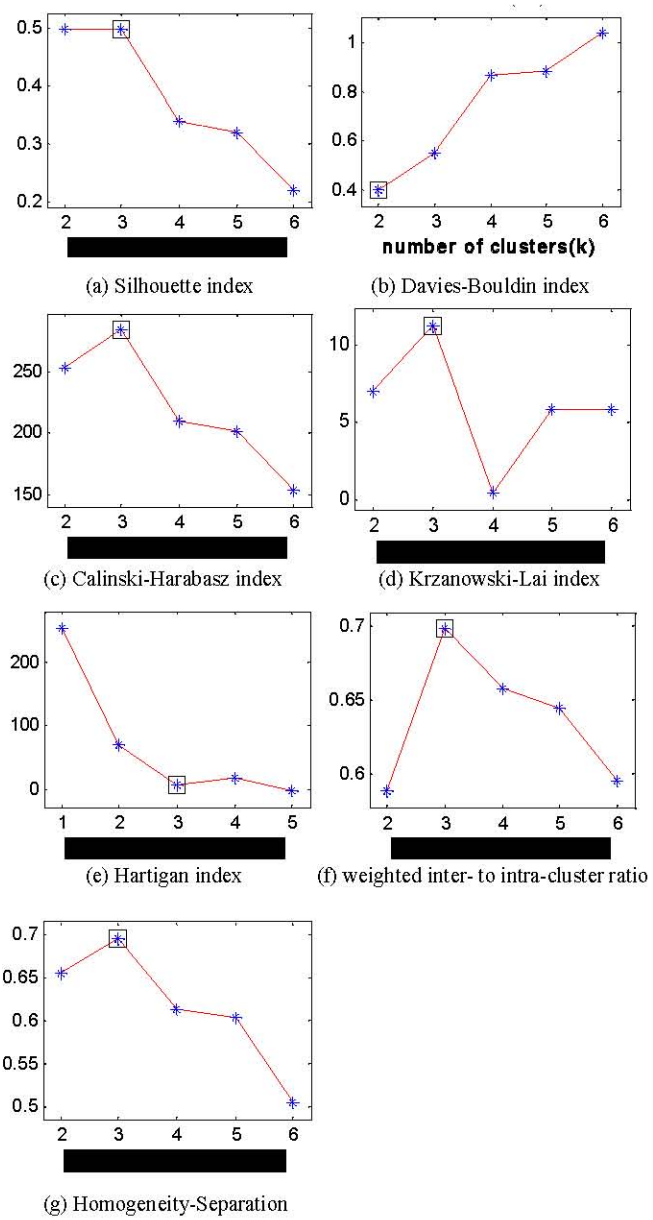
Here, values often have maximum and minimum values, so there are also validity problems. However, in these graphs, the number of classes is smaller than the number of classes in the Wine data set, so the number of points is only five. The clustering performance index data is listed in Table 3. According to the simulation results of above 7 internal indicators, it can be seen that the optimal cluster number is generally found, but the optimal cluster number is not obtained for one indicator. According to the analysis, it can be concluded that it may be related to the selection of the initial cluster center, leading to the deviation of the results. According to the above simulation results and data table, it can be seen that the K-means clustering algorithm has high stability and high accuracy in the classification on the leuk72_3K data set, with an average error rate of about 5%, which can be said to achieve a better clustering effect.

(3) External validity indicators on Wine dataset

The simulation results of wine data set based on K-means clustering algorithm and four external validity indicators are shown in Fig. 3. As can be seen from the simulation diagram in Fig. 3, Mi index is effective at the lowest point, and the other three indexes are effective at the maximum.

(4) Internal validity indicators on Wine dataset

The simulation results of wine data set based on K-means clustering algorithm and 7 internal validity indicators are shown in Fig. 4. As can be seen from the simulation results in Fig. 4, Si, CH, KL, Wint and HS indicators are all effective at the highest point, while DB and HI indicators are effective at the lowest value. It can be seen from these simulation curves that the optimal class number is basically found, but the clustering effect is not optimal due to the high error rate due to the high dimension or complex structure of the data. According to the simulation results of above 7 internal indicators, including the validity of maximum and minimum values, the performance indicator data are listed in Table 4 and Table 5. According to the above simulation results, among the external indicators, only Mi is effective at the minimum value, while the rest are effective at the peak value. In internal indicators, the number of clusters is not very concentrated, so that the error rate is higher than that of K-means clustering algorithm. In general, the accuracy of K-means clustering algorithm for the classification of leuk72_3K data sets is higher than that for the classification of wine data sets.

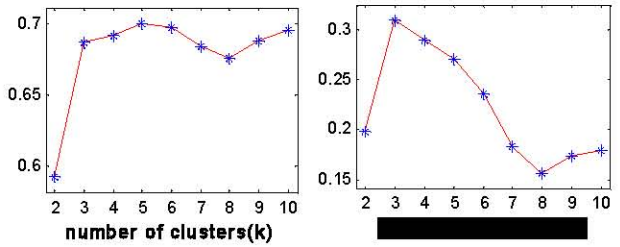


(a) Silhouette index (b) Davies-Bouldin index (c) Calinski-Harabasz index (d) Krzanowski-Lai index (e) Hartigan index (f) weighted inter- to intra-cluster ratio (g) Homogeneity-Separation

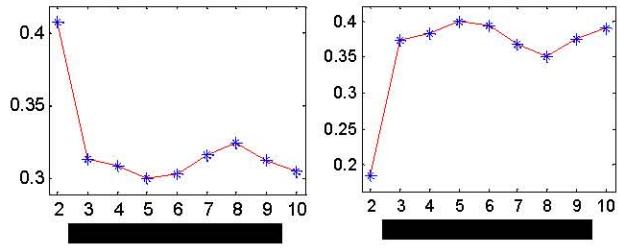
Fig. 2 Internal validity indicators under leuk72_3K dataset.

TABLE 3. PERFORMANCE INDICATOR DATA

Performance indicators	Number of cluster centers				
	2	3	4	5	6
Ri	0.7563	**0.9472**	0.9081	0.8705	0.8173
ARi	0.5134	**0.8809**	0.7089	0.6817	0.5317
Mi	0.2473	**0.0528**	0.0919	0.1295	0.1827
Hi	0.5125	**0.8944**	0.8161	0.7410	0.6346
SiL	0.4969	**0.4970**	0.3389	0.3192	0.2183
DB	**0.3980**	0.5531	0.8668	0.8809	1.0412
CH	253.0578	**284.5089**	209.8667	201.6904	153.3078
KL	7.0431	**11.2587**	0.4562	5.8922	5.8922
HI	253.0578	**69.2454**	7.4437	18.1731	-2.1620
Wint	0.5882	**0.6986**	0.6582	0.6440	0.5954
HS	0.6558	**0.6957**	0.6138	0.6044	0.5050



(a) Rand index (b) Adjusted Rand index

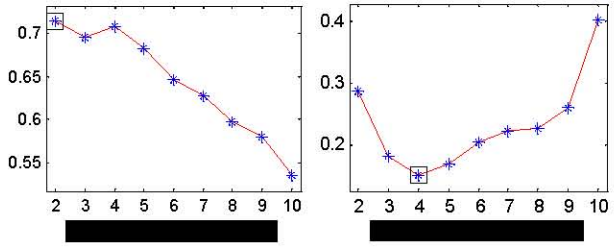


(c) Mirkin index (d) Hubert index

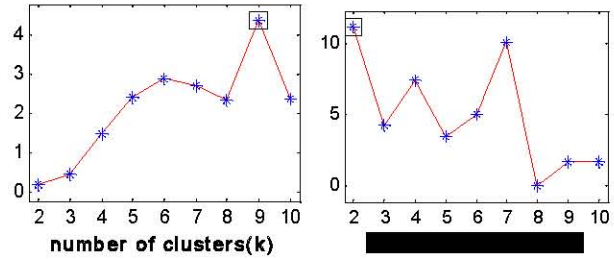
Fig. 3 External validity indicators under Wine dataset.

TABLE 4. PERFORMANCE INDICATOR DATA

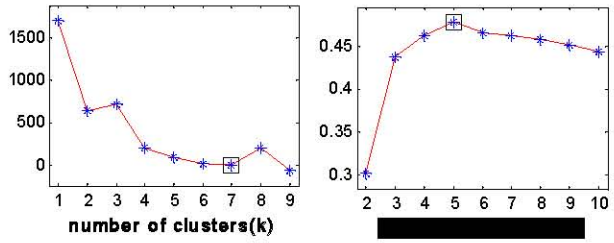
Number of cluster centers	Performance indicators				
	Ri	ARi	Mi	Hi	SiL
2	0.5875	0.1995	0.4125	0.0251	**0.7130**
3	0.6762	**0.3125**	0.3128	0.3748	0.6986
4	0.6770	0.2801	0.3112	0.3796	0.7096
5	**0.7001**	0.2652	**0.3002**	**0.3968**	0.6753
6	0.6751	0.2259	0.3015	0.3806	0.6483
7	0.6750	0.1805	0.3198	0.3695	0.6254
8	0.6736	0.1580	0.3259	0.3509	0.5998
9	0.6763	0.1753	0.3130	0.3740	0.5760
10	0.6752	0.1765	0.3020	0.3801	0.2750



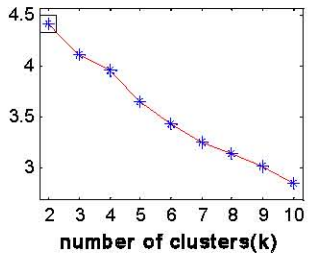
(a) Silhouette index (b) Davies-Bouldin index



(c) Calinski-Harabasz index (d) Krzanowski-Lai index



(e) Hartigan index (f) weighted inter- to intra-cluster ratio



(g) Homogeneity-Separation

Fig. 4 Internal validity indicators under Wine dataset.

TABLE 5. PERFORMANCE INDICATOR DATA

Number of cluster centers	Performance indicators					
	DB	CH	KL	HI	Wint	HS
2	0.2897	0.2035	**11.2385**	1500.3215	0.3025	**4.4357**
3	0.1836	0.4532	4.5326	650.2315	0.4352	4.1053
4	**0.0115**	0.4916	7.4329	700.3621	0.4526	3.9852
5	0.1586	2.4876	3.5023	246.1653	**0.4756**	3.7513
6	0.2104	2.9103	5.0001	70.2431	0.4526	3.4102
7	0.2293	2.5362	10.1521	**10.0012**	0.4523	3.2503
8	0.2298	2.4823	0.0124	10.0023	0.4520	3.1985
9	0.2635	**4.3216**	2.5413	702.3613	0.4513	3.0231
10	0.4158	2.4820	2.5413	10.2153	0.4459	1.5263

B. Clustering Validity Verification Based on PAM Clustering Algorithm

Four external validity indexes and seven internal validity indexes were used for clustering analysis of the leuk72_3K dataset and wine dataset based on PAM clustering algorithm. PAM clustering algorithm is very similar to K-means clustering algorithm, except in the selection of the initial clustering centers.

(1) External validity indicators on leuk72_3k dataset

The simulation results of the leuk72_3K data set based on PAM clustering algorithm and four external validity indicators are shown in Fig. 5. According to the simulation results in Fig. 5, PAM algorithm was used to analyze the leuk72_3K data set, and the optimal clustering number was reflected in the extreme value of the curves, and the accuracy reached 100%.

(2) Internal validity indicators on leuk72_3k dataset

The simulation results of the leuk72_3K data set based on PAM clustering algorithm and 7 internal validity indexes are shown in Fig. 6, in which the values often have maximum and minimum values, so there are also validity problems. In these simulations, the number of classes represented is much higher than before. According to the above simulation results, it can be seen that only the cluster number of DB index deviates from the optimal effect. Other indicators are accurate to obtain the optimal cluster number, so the accuracy is very high. Table 6 and Table 7 show the clustering performance index parameters. According to the simulation results, the classification error rate of PAM clustering algorithm for the leuk72_3K data set is close to zero, and basically all the indexes get the optimal results. However, one index (DB index) did not obtain the optimal cluster number.

(3) External validity indicators on Wine dataset

The simulation results of wine data set based on PAM clustering algorithm and four external validity indicators are shown in Fig. 7. As can be seen from the simulation results in Fig. 7, only Mi is effective at the lowest point, while Ri, ARi and Hi are effective at the maximum value. In addition, Mi index, Ri index and Hi index all achieved the best clustering effect, while ARi index did not. The reason may be that the algorithm has a great influence on the selection of the initial clustering centers of wine data set when discriminating with this index, or it may be related to the dimension of wine data set, spatial structure and other factors. In general, classification effect of PAM clustering algorithm on this data set is very accurate.

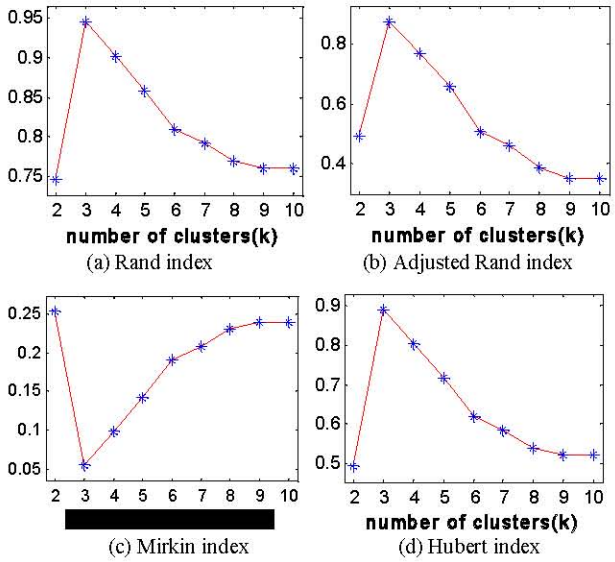


Fig. 5 External validity indicators under leuk72_3 dataset.

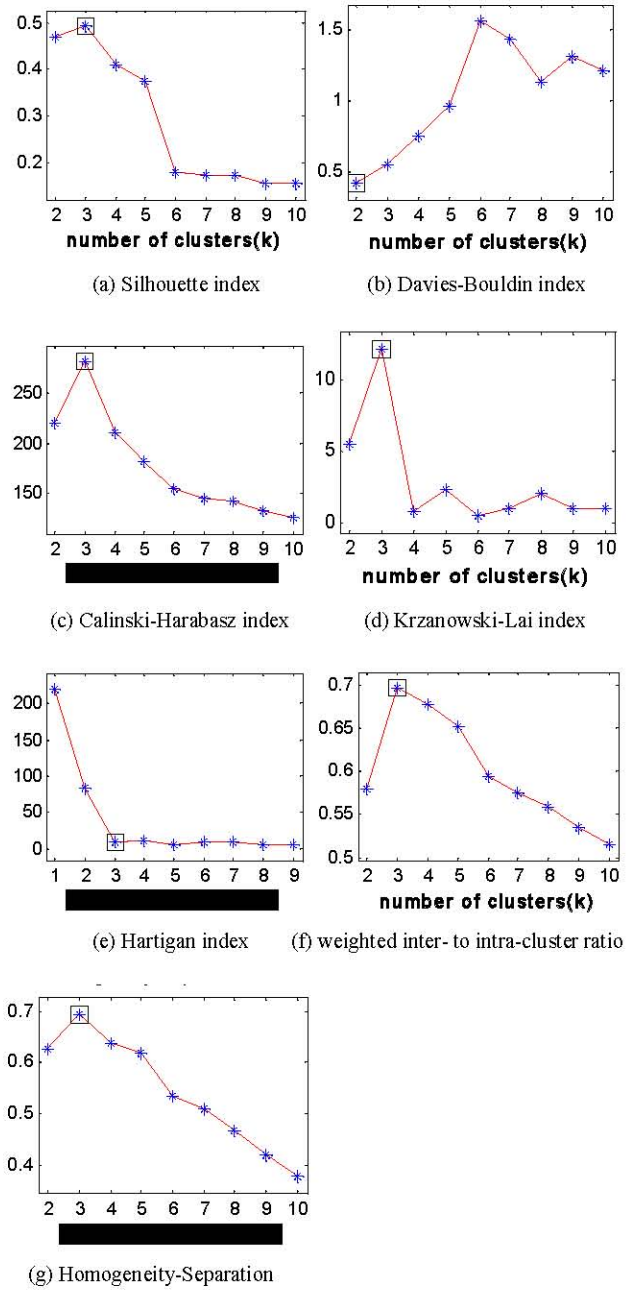


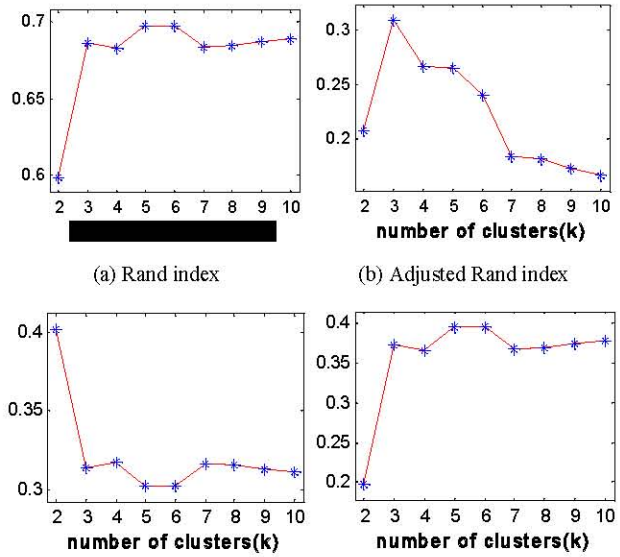
Fig. 6 Internal validity indicators under leuk72_3 dataset.

TABLE 6 PERFORMANCE INDICATOR DATA

Number of cluster centers	Performance indicators				
	Ri	ARi	Mi	Hi	SiL
2	0.7469	0.4925	0.2531	0.4937	0.4690
3	0.9448	0.8758	0.0552	0.8897	0.4934
4	0.9010	0.7685	0.0990	0.8020	0.4095
5	0.8588	0.6593	0.1412	0.7175	0.3754
6	0.8087	0.5083	0.1913	0.6174	0.1797
7	0.7923	0.4598	0.2077	0.5845	0.1733
8	0.7700	0.3872	0.2300	0.5399	0.1726
9	0.7613	0.3509	0.2387	0.5227	0.1551
10	0.7617	0.3499	0.2383	0.5253	0.1563

TABLE 7 PERFORMANCE INDICATOR DATA

Serial number	Performance indicators					
	DB	CH	KL	HI	Wint	HS
2	0.4170	219.8464	5.5747	219.8464	0.5791	0.6276
3	0.5486	280.8172	12.1824	83.3027	0.6969	0.6939
4	0.7499	209.7426	0.8564	8.2862	0.6778	0.6365
5	0.9554	181.3576	2.3999	10.2850	0.6518	0.6187
6	1.5576	154.1780	0.5424	4.7591	0.5934	0.5329
7	1.4279	145.5354	1.0473	8.9906	0.5745	0.5084
8	1.1239	142.4573	2.0843	9.5207	0.5585	0.4656
9	1.3104	133.0196	1.0773	4.9777	0.5343	0.4207
10	1.2057	125.9736	1.0773	4.8346	0.5152	0.3770



(a) Rand index (b) Adjusted Rand index

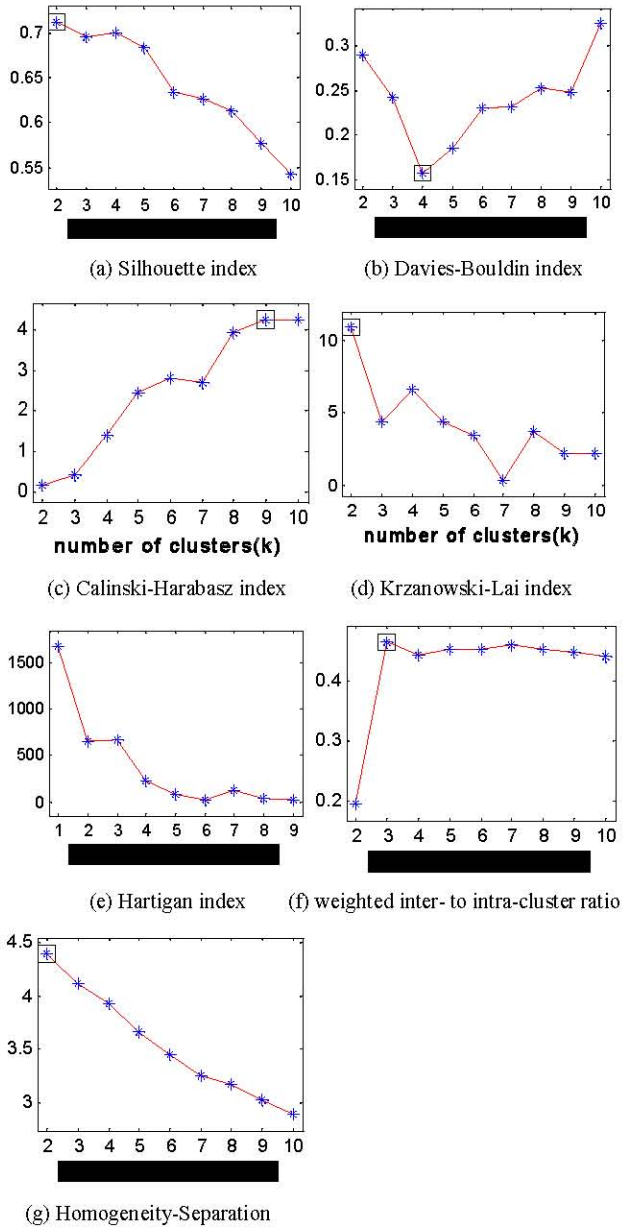
(c) Mirkin index (d) Hubert index

Fig. 7 External validity indicators under Wine dataset.

(4) Internal validity indicators on Wine dataset

The simulation results of wine data set based on PAM clustering algorithm and 7 internal validity indicators are shown in Fig. 8. As can be seen from the simulation results in Fig. 8, Si, CH, KL, Wint and HS indicators are all effective at the highest point, while DB and HI indicators are effective at the lowest value. Table 8 and Table 9 show the performance indicators. It can be seen from these simulation diagrams that the clustering numbers are basically found, but the error rate may be high due to the high dimension of wine data set, data structure and other problems.

According to above data table and simulation diagram, the clustering effect is not very obvious. The error rate is very high, which may be related to the selection of the initial cluster centers. In general, the classification effect of PAM algorithm on the leuk72_3K data set is more reasonable and accurate than that of wine data set.



(a) Silhouette index (b) Davies-Bouldin index

(c) Calinski-Harabasz index (d) Krzanowski-Lai index

(e) Hartigan index (f) weighted inter- to intra-cluster ratio

(g) Homogeneity-Separation

Fig. 8 Internal validity indicators under Wine dataset.

TABLE 8. PERFORMANCE INDICATOR DATA

Number of cluster centers	Performance indicators					
	DB	CH	KL	HI	Wint	HS
2	0.2846	0.1964	11.2431	1623.5486	0.1983	4.4531
3	0.2419	0.3143	4.0216	669.2348	0.4356	4.2130
4	0.1501	1.2105	7.1024	681.9834	0.4297	3.9216
5	0.1853	2.4943	4.9215	260.1654	0.4305	3.6913
6	0.2317	2.8916	3.9423	115.5983	0.4306	3.4920
7	0.2320	2.8898	0.9983	180.1365	0.4210	3.2862
8	0.2511	3.9910	3.9862	140.1597	0.4315	3.2013
9	0.2496	4.1350	2.9543	120.2156	0.4295	3.1057
10	0.3215	4.1345	2.9541	118.3540	0.4238	2.8160

TABLE 9. PERFORMANCE INDICATOR DATA

Number of cluster centers	Performance indicators				
	Ri	ARi	Mi	Hi	SiL
2	0.5968	0.2139	0.4016	0.1986	0.7213
3	0.6835	0.3150	0.3219	0.3864	0.6916
4	0.6745	0.2563	0.3360	0.3619	0.7023
5	0.6943	0.2550	0.3001	0.3913	0.6813
6	0.6942	0.2346	0.3101	0.3910	0.6231
7	0.6740	0.1983	0.3410	0.3601	0.6129
8	0.6768	0.1846	0.3408	0.3605	0.6017
9	0.6790	0.1023	0.3403	0.3613	0.5753
10	0.6915	0.0916	0.3309	0.3769	0.5473

V. CONCLUSIONS

Based on K-means clustering and PAM clustering methods, this paper uses 11 clustering validity indexes, including 4 external indexes and 7 internal indexes, to carry out simulation experiments on the leuk72_3K data set and wine data set. The experimental results show that the K-means clustering algorithm based on external and internal indexes has a general classification effect on wine data set, but the classification accuracy of leuk72_3K data set is much higher than that of wine data set. The classification result of the PAM clustering algorithm based on external indexes is very accurate, but the classification effect of PAM clustering algorithm based on internal indexes is not good. Similarly, for the leuk72_3K data set, similar results were obtained by PAM clustering algorithm based on external indexes or internal indexes.

REFERENCES

[1] S. Askari, "Fuzzy C-Means Clustering Algorithm for Data with Unequal Cluster Sizes and Contaminated with Noise and Outliers: Review and Development," *Expert Systems with Applications*, vol. 165, pp. 113856, 2021.

[2] A. Ishizaka, B. Lokman, and M. Tasiou, "A Stochastic Multi-criteria Divisive Hierarchical Clustering Algorithm," *Omega*, vol. 103, pp. 102370, 2021.

[3] J. A. Hartigan, and M. A. Wong, "A K-means Clustering Algorithm," *Journal of the Royal Statistical Society*, vol. 28, no. 1, pp. 100-108, 1979.

[4] S. S. Yu, S. W. Chu, C. M. Wang, Y. K. Chan, and T. C. Chang, "Two Improved K-means Algorithms," *Applied Soft Computing*, vol. 68, pp. 747-755, 2018.

[5] K. P. Sinaga, and M. S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 80716-80727, 2020.

[6] M. J. Rezaee, M. Eshkevari, M. Saberi, and O. Hussain, "GBK-means Clustering Algorithm: An Improvement to the K-means Algorithm Based on the Bargaining Game," *Knowledge-Based Systems*, vol. 213, pp. 1066722, 2021.

[7] L. F. Zhu, J. S. Wang, and H. Y. Wang, "A Novel Clustering Validity Function of FCM Clustering Algorithm," *IEEE Access*, vol. 7, pp. 152289-152315, 2019.

[8] H. Y. Wang, J. S. Wang, and L. F. Zhu, "A New Validity Function of FCM Clustering Algorithm Based on Intra-class Compactness and Inter-class Separation," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 6, pp. 12411-12432, 2021.

[9] L. P. Panapakidis, and G. C. Christoforidis, "Implementation of Modified Versions of the K-means Algorithm in Power Load Curves Profiling," *Sustainable Cities and Society*, vol. 35, pp. 83-93, 2017.

[10] S. Douglas, M. J. Brusco, and L. Hubert, "The Variance of the Adjusted Rand Index," *Psychological Methods*, vol. 2, pp. 261, 2016.

[11] Q. P. Zhao, and P. Fränti, "WB-index: A Sum-of-squares Based Index for Cluster Validity," *Data & Knowledge Engineering*, vol. 92, pp. 77-89, 2014.

[12] C. Songul, "Integrated K-means Clustering with Data Envelopment Analysis of Public Hospital Efficiency," *Health Care*, vol. 23, pp. 325-338, 2020.

[13] T. Venera, M. Fordellone, and M. Vichi, "Building Well-Being Composite Indicator for Micro-Territorial Areas Through PLS-SEM and K-Means Approach," *Social Indicators Research*, vol. 153, pp. 407-429, 2021.