Applied Soft Computing 71 (2018) 608–621

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

An effective partitional clustering algorithm based on new clustering validity index

Erzhou Zhu [a,*], Ruhui Ma [b]

[a] School of Computer Science and Technology, Anhui University, Hefei 230601, PR China
[b] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, PR China

ARTICLE INFO

Article history:
Received 22 December 2017
Received in revised form 7 July 2018
Accepted 10 July 2018
Available online 18 July 2018

Keywords:
K-means
Partitional clustering
Clustering validity index

ABSTRACT

As an unsupervised pattern classification method, clustering partitions the input datasets into groups or clusters. It plays an important role in identifying the natural structure of the target datasets. Now, it has been widely used in data mining, pattern recognition, image processing and so on. However, due to different settings of the parameters and random selection of initial centers, traditional clustering algorithms may produce different clustering partitions for a single dataset. Clustering validity index (CVI) is an important method for evaluating the effect of clustering results generated by clustering algorithms. However, many of the existing CVIs suffer from complex computation, low time efficiency and narrow range of applications. In order to make clustering algorithms more stable, traditional *K*-means is firstly improved by the density parameters based initial center selection method other than randomly selecting initial centers. Then, in order to enlarge the application range of clustering and better evaluate the clustering partition results, a new variance based clustering validity index (VCVI) from the point of view of spatial distribution of datasets is designed. Finally, a new partitional clustering algorithm integrated with the improved K-means algorithm and the newly introduced VCVI is designed to optimize and determine the optimal clustering number (K_{opt}) for a wide range of datasets. Furthermore, the commonly used empirical rule $K_{\max} \leqslant \sqrt{n}$ is reasonably explained by the newly designed VCVI. The new algorithm integrated with VCVI is compared with traditional algorithms integrated with five commonly used CVIs. The experimental results show that our new clustering method is more accurate and stable while consuming relatively lower running time.

1. Introduction

Clustering belongs to the unsupervised machine learning method [1]. It partitions the input datasets into groups or clusters. Similar to the principle of "birds of a feather flock together", by the clustering algorithm, the objects within a cluster are similar while objects in different clusters are dissimilar. Clustering plays an important role in finding the natural structure of the target datasets. It is widely used in many fields such as data mining, pattern recognition and image processing [2,3].

By now, researchers have put forward many clustering algorithms to improve data processing efficiency, which can be mainly divided into 4 categories: the partitional clustering algorithm, the hierarchical clustering algorithm, a grid-based clustering algorithm and the density-based clustering algorithm [4]. Among them, the partitional clustering algorithm is the most widely used clustering algorithm. Among various implementations of partitional clustering algorithms, K-means is one of the most popular in reality because of its simplicity and effectiveness. However, due to different settings of the parameters and random selection of initial clustering centers, the traditional K-means algorithm is not stable. It may produce different clustering partitions for a single dataset. This instability also exist in many variant K-means algorithms, such as K-means++ algorithm [5]. Lately, some improved algorithms [6,7] are proposed to select initial cluster centers, but they need to predefine the number of clusters or preset the values of the parameters [8]. The choice of initial centers has an important influence on the clustering effect of the clustering algorithm. In order to solve this problem, this paper revises the traditional K-means algorithm by using the density parameters to select initial centers. By calculating the density parameter of each sample point, the initial clustering centers are no longer randomly selected.

As the number of clusters in a dataset is usually in a fuzzy interval, it is difficult to determine the K_{opt} [9] practically. CVI has always

* Corresponding author.
E-mail address: ezzhu@ahu.edu.cn (E. Zhu).

https://doi.org/10.1016/j.asoc.2018.07.026
1568-4946/

been the focus of cluster analysis and a CVI with good performance is the key and core to optimize and determine the K_{opt} [10]. The researches on CVI mainly use mathematical knowledge to model the validity index. Then, when the clustering number takes different values, the clustering results are evaluated respectively. When the optimal value of the index is set, the corresponding clustering result is the optimal clustering partition of the dataset, and the corresponding clustering number is the optimal clustering number K_{opt}. A lot of studies [11,12] have shown that there is no CVI that can optimally process all the datasets. Meanwhile, the existing CVIs have some shortcomings, such as too complex calculation, low computational efficiency and narrow range of applications [13]. In this paper, a new variance based clustering validity index (VCVI) from point of view of spatial distribution of datasets is introduced. The new VCVI needs relatively less computation resources and exhibits higher efficiency compared with the existing CVIs. The VCVI also exhibits better clustering performance on processing irregular datasets such as non-spherical distribution datasets.

In the process of solving the optimal clustering number, the existing clustering validity indexes utilize the empirical rule (suggested by Bezdek [14]) $K_{\max} \leqslant \sqrt{n}$ to degrade the scale of calculation. However, they did not address the reasonableness of this rule. In this paper, through the combination of the new VCVI and the space fractal geometry knowledge, this empirical rule is reasoned and explained. Generally speaking, this paper makes the following contributions:

(1) *The traditional K-means algorithm is improved*. By replacing the method of randomly selecting the initial clustering centers with the density parameters based initial centers selection method, the traditional K-means is improved. With the improved algorithm, the selection of the initial centers is fixed. The improved algorithm also provides the prerequisite for performing our new VCVI and the optimization and determination algorithm for K_{opt}.
(2) *A new clustering validity index is proposed*. CVI is an important means to evaluate the optimal clustering partitions for the target datasets. During the process of the clustering, the optimal partitions of non-labelled datasets can be determined by observing their spatial distributions. For tagged datasets, the optimal number of clusters can be calculated by the label category. For a new CVI, its performance can be verified by applying it to the tagged datasets. If the new CVI can get the optimal clustering number or the near optimal clustering number for all the tagged datasets, it can be applied to the non-labelled datasets [15]. By extending the variance theory to the multi-dimensional space (Euclidean Space), a new clustering validity index (VCVI) from the point of view of spatial distribution of datasets is proposed. The new VCVI proposed can extend the application range of the partitional clustering algorithms and better evaluate the effect of clustering results. Meanwhile, by utilizing the new VCVI and the spatial fractal geometry theory in mathematics, the rationale of the commonly used empirical rule $K_{\max} \leqslant \sqrt{n}$ where n represents the number of sample points in a dataset is explained.
(3) *A new partitional clustering algorithm is designed*. At the initial stage of the conventional partitional clustering algorithms (such as K-means, K-medoids and FCM (Fuzzy Clustering Algorithm)), it is needed to set the number of the optimal clustering partitions (i.e. the value of K_{opt}). However, the unreasonable value of the K_{opt} will result in poor partition of the target dataset. In this paper, a new partitional clustering algorithm integrated with the improved K-means algorithm and the newly introduced VCVI are designed to optimize and determine K_{opt} and optimal clustering partitions.

The new algorithm integrated with VCVI is compared with traditional algorithms integrated with 5 commonly used CVIs (Dunn's index (abbreviated by DI-index in this paper) [16], Davies–Bouldin index (abbreviated by DBI-index) [17], Calinski Harabasz index (abbreviated by CH-index) [18], Bandyopadhy index (abbreviated by I-index) [19] and Ibai Gurrutxaga index (abbreviated by COP-index) [20]). The experimental results have shown that the new algorithm is more stable and accurate while consuming relatively lower running time.

The remainder of this paper is organized as follows: Section 2 discusses the related work. Section 3 introduces prerequisite background knowledge. Section 4 gives the implementations of the new VCVI and the improved algorithm. Section 5 evaluates the experimental results. Finally, Section 6 briefly concludes this paper and outlines our future work.

2. Related work

By now, a variety of clustering algorithms has been proposed and applied to various fields [21]. During the execution of the clustering algorithms, the procedure adopted for choosing initial clustering centers is extremely important as it has direct impact on the formations of the final clusters [6]. However, the random method of selecting initial clustering centers makes the traditional K-means algorithms unstable. The K-means++ algorithm [5] is proposed to resolve the problem. In this algorithm, only the first cluster center is randomly selected while the remainder initial cluster centers are selected as far as possible from the first point. The MinMax K-means algorithm [8] is a method starting from a randomly picked set of centers and tries to minimize the maximum intra-cluster variance instead of the sum of the intra-cluster variances. Based on the two predefined principal variables that best describe the change in the dataset, Erisolgu [6] proposed an incremental approach for computing initial clustering centers. In this approach, the reduced dataset is partitioned one by one until the number of clusters equals to the predefined number of clusters. By Erisolgus approach, the problem of randomly selecting initial clustering centers is avoided. However, the incremental approach will fall the clustering results into local optima. In order to cope with the local optima problem, the global K-means [22,23] algorithm is proposed. It is the incremental approach that starts from a single cluster and at each step a new cluster is deterministically added to the solution according to a predefined criterion. This algorithm is not susceptible to local optima, but it is more expensive in computation. Different from the above method, our paper utilizes the density parameter to form the initial clustering centers rather than selecting them randomly. The improved K-means algorithm based on density parameter is much more stable and not susceptible to local optima.

In clustering analysis, CVI is an important means to evaluate the effect of the clustering results generated by clustering algorithm. A CVI with good performance is a key to optimize and determine K_{opt} [10]. For better evaluating the clustering results and determining K_{opt}, many CVIs have been proposed. Generally, the commonly used CVIs can be divided into three categories: the indexes based on the fuzzy division of datasets, the indexes based on the statistical information of datasets and the indexes based on the geometric structure of a datasets. Xie-Beni [24] is a fuzzy division based CVI that combines the objective function of fuzzy clustering, the structure of the dataset itself and the nature of the fuzzy membership degree. The fuzzy division based CVIs can objectively evaluate the clustering results, but they are not suitable to evaluate the results of hard clustering algorithms [25]. The improved K-means algorithm in our paper is one of the hard clustering algorithms, so the fuzzy division based CVIs are not suitable to evaluate the results of our algorithm. IGP [26] is a representative CVI based on the statistical

information of datasets. It uses the in-group radio of the intra data points to evaluate the clustering results. Since it only focuses on the adjacent consistency [27], this kind of CVIs is not stable for many datasets. This means the number of clusters generated by the IGP index is usually less than the actual number.

Based on the geometric structure of a dataset, many CVIs have been proposed. The DI-index [16] is a ratio-type index where the cohesion is estimated based on the distances from the points in a cluster to the centroid. Since the DI-index is too sensitive to the noise data, it is difficult to find K_{opt} from the datasets with outliers. Meanwhile, the computational complexity of the index function increases exponentially with the increasing size of the dataset and the value of K, so the DI-index is not suitable for dealing with large scale datasets. The DBI-index [17] is a commonly used CVI. It estimates the cohesion based on the distance from the points in a cluster to the centroid and the separation based on the distance in between. The DBI-index function is suitable for measuring the clusters of datasets with within-cluster compactness, between-cluster separation, so the greater the overlap of dataset, the worse the performance of DBI-index clustering evaluation. In [28], Milligan and Cooper deeply discussed the clustering performance of the CH-index [18] and the existing CVIs. Through extensive contrasts in different datasets, they found that, in most cases, the CH-index is superior to most CVIs in evaluating clustering performance and determining the optimal number of clusters. The I-index [19] is mainly composed of three factors: $1/K, E_1/E_K$ and $\max_{1\leq i,j\leq K} d(v_i, v_j)$. The factor $1/K$ decreases with the increase of the cluster number K, and the other two factors increase with the increase of K. The three factors of the I-index can balance each other, so as to ensure better clustering. This index is suitable for dealing with some datasets with less class numbers, but it relies too much on user settings for some parameters. By calculating the average distance between the sample points of each cluster to its center, the COP-index [20] measures the compactness of the sample distribution within each cluster, while the separability among clusters is measured by the farthest distance. Accordingly, the smaller the COP-index, the better clustering results of the dataset. The COP-index was initially proposed to be used in conjunction with a cluster hierarchy post-processing algorithm. But it can be used as an ordinary CVI.

The optimal number of clusters derived by CVIs, DI-index, DBI-index, CH-index, I-index, COP-index and many other clustering validity indexes are based on the assumption that the clustering algorithms have already made the optimal partition for the target datasets. However, in many cases the optimal clustering partition is unknown. So, in order to overcome this defect, this paper firstly uses the improved K-means algorithm to drive the optimal partition before the proposed VCVI starting to work.

In the era of big data, the expansion of data scale will lead to the huge calculation of CVIs. With the big data, existing CVIs, like DI-index and COP-index usually lead to the imbalance of index performance and computation efficiency [29]. So, they cannot solve the optimal clustering number K_{opt} efficiently for all datasets. As a matter of fact, most of the existing CVIs can effectively process the dataset of within-cluster compactness, between-cluster separation and each cluster presenting a spherical distribution [30]. However, it is difficult for them to find K_{opt} for the non-spherical distribution datasets, datasets with a large number of outliers and overlapping and datasets with different cluster sizes and densities. Of course, there are already many CVIs for datasets with non-spherical distribution, different clusters with different sample sizes and densities and large degree overlap among clusters [31,32]. But these CVIs usually incur high processing time or exhibit relatively poor clustering accuracy [33]. The new VCVI proposed in this paper needs relatively less computation and exhibits higher efficiency compared with existing CVIs. The VCVI also exhibits better clustering performance on processing the above irregular datasets.

In the process of solving the optimal clustering number, the existing clustering validity indexes utilize the empirical rules $K_{\max} \leqslant \sqrt{n}$ to degrade the scale of the problem. However, they did not address the reasonableness of empirical rules. In this paper, through the combination of the new VCVI and the space fractal geometry knowledge, this empirical rule is reasoned and explained.

3. Background knowledge

This section introduces the preliminary knowledge (includes the basic theory of the clustering algorithm and the definition of five existing CVIs) for better presenting our new VCVI and its corresponding algorithms.

3.1. Clustering algorithm description

In the Euclidean space R^m, a dataset containing n sample points $D = \{x_1, x_2, \ldots, x_n\}$ is given. Of which, each sample point $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ has m attributes. By the clustering algorithm, D is divided into K clusters $C = \{C_1, C_2, \ldots, C_K\}$, a $K \times n$ partition matrix, represented as $U(D) = [U_{ki}]$, is derived where $k = 1, 2, \ldots, K$; $i = 1, 2, \ldots, n$; u_{ki} is the degree of membership of sample point x_i to cluster C_k. Clustering can be divided into hard clustering and soft clustering (fuzzy clustering). In this paper, hard clustering is drawn to implement our VCVI, so the membership degree u_{ki} should be satisfied with formula (1). In the formula, the results of clustering must be satisfied with $C_i \neq \phi, D = U_{i=1}^{K} C_i, C_i \bigcap C_j \neq \phi, (i \neq j, i, j = 1, 2, \ldots, K)$.

$$u_{ki} = \begin{cases} 1, & \text{if } x_i \in C_k \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Definition 1. The Euclidean space distance between sample points $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ and $x_j = \{x_{j1}, x_{j2}, \ldots, x_{jm}\}$ can be calculated as (each sample point contains m feature attributes):

$$d(x_i, x_j) = ||x_i - x_j|| = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{im} - x_{jm})^2} \tag{2}$$

Definition 2. The criterion function is defined as:

$$E = \sum_{i=1}^{K} \sum_{x \in C_i} d(x, \bar{v}_i)^2 \tag{3}$$

where E is the sum of the mean errors of all sample points; x is the sample point of cluster C_i; and $\bar{v}_i$ is the center point of cluster C_i.

3.2. Commonly used clustering validity indexes

(1) *Dunn index (DI-index).* This CVI is proposed by Dunn [16], which can be described as:

$$DI(K) = \min_{1\leq i\leq K}\{\min_{1\leq j\leq K \& i\neq j}\left(\frac{D_{\min}(C_i, C_j)}{\max_{1\leq p\leq K}\omega(C_p)}\right) \tag{4}$$

where $D_{\min}(C_i, C_j) = \min_{x\in C_i \& y\in C_j} d(x, y)$ and $\omega(C_p) = \max_{1\leq i\leq |C_p|} d(x_i, y_j)$. The optimal clustering number K_{opt} is given by the following formula:****

$$K_{opt} = \{K | \max_{2\leq K\leq n-1}\{DI(K)\}\} \tag{5}$$

(2)*Davies–Bouldin index (DBI-index).* This CVI is proposed by Davies–Bouldin [17], which can be described as:

$$DBI(K) = \frac{1}{K}\sum_{i=1}^{K} \max_{j\neq i}\left(\frac{avg(C_i) + avg(C_j)}{\delta(C_i, C_j)}\right) \tag{6}$$

where $avg(C_i) = \frac{1}{|C|}\sum_{x\in C_i} d(x, v_i)$, v is center of the cluster C and $|C|$ is the number of sample points in cluster C. $\delta(C_i, C_j) = d(v_i, v_j)$; v_i is

the center of cluster C_i. The value of the optimum cluster number K_{opt} is given by the following formula:

$$K_{opt} = \{K|\min_{2\le K\le n-1}\{DBI(K)\}\} \tag{7}$$

(3) *Calinski–Harabasz index (CH-index).* This CVI is proposed by Calinski–Harabasz [18], which can be described as:

$$CH(K) = \frac{\left[\frac{\sum_{i=1}^{K}|C_i|d(v_i,v)^2}{K-1}\right]}{\frac{\sum_{i=1}^{K}\sum_{x\in C_i}d(x,v_i)^2}{n-K}} \tag{8}$$

In the formula above, v_i is the central point of cluster C_i, and v is the global centers vector of all samples in the dataset. The optimal clustering number K_{opt} is given as:

$$K_{opt} = \{K|\max_{2\le K\le n-1}\{CH(K)\}\} \tag{9}$$

(4) *Bandyopadhy index (I-index).* This CVI (proposed by Bandyopadhy [19]) can be described as:

$$I(K) = \left[\left(\frac{1}{K}\times\frac{E_1}{E_K}\right)\times\max_{1\le i,j\le K}d(v_i,v_j)\right]^2 \tag{10}$$

Here, $E_K = \sum_{k=1}^{K}\sum_{j=1}^{n}u_{kj}\times d(v_j,v_K)$, $p\ge 1$, n is the sample number of the dataset. $U=[U_{ki}]_{K\times n}$ is the partition matrix of the dataset. v_{jk} is the clustering center of cluster C_i. p is set as 2 by the authors who proposed this index. As long as the value of p is no less than 1, there is no effect on the final result. In this paper, p is set as 1 for simplicity. The value of the optimum cluster number K_{opt} is given by the following formula:

$$K_{opt} = \{K|\max_{2\le K\le n-1}\{I(K)\}\} \tag{11}$$

(5) *Ibai Gurrutxaga index (COP-index).* This CVI is proposed by Ibai Gurrutxaga [20], which can be described as formula (12) where n is the sample number of the dataset, and v_i is the center point of cluster C_i. The corresponding K_{opt} can be calculated by formula (13)

$$COP(K) = \frac{1}{n}\sum_{i=1}^{K}|C_i|\frac{(1/|C_i|)\sum_{x\in C_i}d(x,v_i)}{\min_{y\notin C_i}\max_{z\in C_i}d(y,z)} \tag{12}$$

$$K_{opt} = \{K|\min_{2\le K\le n-1}\{COP(K)\}\} \tag{13}$$

4. New approaches for effective clustering

In this section, an improved K-means clustering algorithm based on density parameters for selecting initial clustering centers is firstly proposed. Then, VCVI, a new variance based clustering validity index that from the point of view of spatial distribution of datasets, is defined. Finally, a new optimization and determination algorithm for K_{opt} that combines the improved clustering algorithm and new VCVI is designed.

4.1. Improved K-means algorithm

The traditional K-means algorithm randomly selects the number (K) of initial clustering centers. However, the poor selection of initial centers may cause slow convergence of clustering. In order to improve the accuracy and clustering effect of the K-means algorithm, as well as the optimization of the K value, the traditional K-means algorithm needs to be improved.

Definition 3. In datasets $D=\{x_1, x_2, \ldots, x_n\}$, the average distance between all sample points pair (x_i, x_j) is defined as:

$$AveDist = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}d(x_i,x_j) \tag{14}$$

Definition 4. In Euclidean space R^m, a dataset containing n samples $D=\{x_1, x_2, \ldots, x_n\}$ is given. Of which, sample point $x_i (i=1,2,\ldots, n)$ is the center of the circle, and the region with ε radius is called the neighborhood of sample point x_i. And then the neighborhood value of each sample point in the dataset is defined as:

$$\varepsilon = \alpha\times AveDist,\quad 0<\alpha\le 1 \tag{15}$$

where α is called the influence factor of neighborhood value, and *AveDist* is calculated by formula (14). In order to achieve the better clustering partition effect of the K-means algorithm, the value of the impact factor needs to be properly adjusted according to the space distribution characteristics of the dataset. (1) When the datasets with within-cluster compactness, between-cluster separation are encountered, the value of is restricted to the interval of [0.05, 0.2], and its common values are 0.05, 0.1, 0.125, 0.15, 0.175 and 0.2. (2) When the datasets with within-cluster separation, between-cluster (relatively) compactness are encountered, the value of is restricted to the interval of [0.2, 0.5], and its common values are 0.2, 0.3, 0.4 and 0.5.

Definition 5. In Euclidean space R^m, a dataset containing n samples $D=\{x_1, x_2, \ldots, x_n\}$ is given. Of which, the number of sample points in the field with $x_i (i=1, 2, \ldots, n)$ as the center and ε as the radius is called the density parameter of sample points x_i. Specifically, the density parameter is defined by formula (16). The corresponding jump function $sgn(x)$ is defined by formula (17). By this density parameter, the initial clustering centers are not randomly selected.

$$\rho(x_i,\varepsilon) = \sum_{j=1,j\ne i}^{n} sgn(\varepsilon - d(x_i,x_j)) \tag{16}$$

$$sgn(x) = \begin{cases} 1, & x\ge 0 \\ 0, & x<0 \end{cases} \tag{17}$$

Based on the above definitions, we can present the improved K-means algorithm based on density parameters for selecting the initial cluster centers (described as Fig. 1). Firstly, according to formula (14), average distance *AveDist* between all the sample points pair (x_i, x_j) is calculated. Meanwhile, according to formula (15), formula (16) and formula (17), the density parameter of each sample point is obtained $(x_i, \varepsilon)(i=1,2,\ldots,n, x_i\in D)$. Secondly, selecting the sample point that has the largest density parameter as the first initial clustering center point, meanwhile, all the samples in its field ε are deleted from dataset D. Finally, repeating the above steps until K initial cluster centers are obtained. Specifically, line (1) calculates the average distance between all sample points pair. Based on the average distance obtained by line (1), lines (2)–(4) use Eqs. (15) and (16) to calculate the density parameter $\rho(x_i, \varepsilon)$ of each point x_i in the range of the circle which takes as the radius. Lines (5)–(8) firstly take the data point with the highest density as the first initial cluster center, and remove the points in its neighborhood; then take the data point with the highest density in the remainder of the dataset as the second initial clustering center and remove the points in its neighborhood; repeat the above steps until all the initial cluster centers are found. Lines (9)–(21) are the clustering process of the traditional K-means algorithm. The improved K-means algorithm no longer randomly selects the initial clustering center point, so the stable clustering result can be obtained.

4.2. Design of new VCVI

In probability and statistics, variance is an important tool to measure the dispersion degree between datasets. The smaller the variance is, the more concentrated the data. But variance is mainly used to measure the deviation degree of one-dimensional datasets.

Input: Dataset $D = \{x_1, x_2, \ldots, x_n\}$; The value of the cluster number K.
Output: K clusters are divided into $C = \{C_1, C_2, \ldots, C_K\}$.

Clustering process:
(1) Calculate the average distance between all the sample points pair (x_i, x_j) in D according to $AveDist$ defined by formula (14);
(2) for $i = 1, 2, \ldots, n$ do
(3) Calculate the density parameter $\rho(x_i, \varepsilon)$ of sample point x_i on the basis of formula (15) and (16) and put it into set $S = \{\rho(x_i, \varepsilon)\}$;
(4) end for;
(5) for $j = 1, 2, \ldots, K$ do
(6) Select the sample point with the j^{th} largest density parameter, $(x_j = \max\{S\})$, from set S. And set it as the j^{th} initial clustering center v_j.
(7) Remove all density parameters of the sample points in the neighborhood x_i from S.
(8) end for;
(9) Get the initial cluster center set: $\{v_1, v_2, \ldots, v_K\}$;
(10) Repeat
(11) Let $C_i = \emptyset$ $(1 \le i \le K)$;
(12) for $j = 1, 2, \ldots, n$ do
(13) According to formula (2), the distances from sample point x_j of the dataset to all the cluster center $v_i (1 \le i \le K)$are calculated: $d(x_j, v_i)$;
(14) According to the nearest distance principle, sample point x_j is partitioned into the corresponding class clusters;
(15) end for
(16) for $j = 1, 2, \ldots, K$ do
(17) Calculate the new cluster centers: $v_j' = \frac{1}{|C_j|}\sum_{x \in C_j} x$;
(18) if $v_j \ne v_j'$
(19) Update cluster center point: $v_j \leftarrow v_j'$;
(20) end for
(21) Until the criterion function defined by formula (3) converges to a constant. //End Repeat

Fig. 1. Improved K-means algorithm.

Therefore, it is necessary to extend it to multidimensional space (Euclidean Space), so as to calculate the dispersion degree for multidimensional data. Based on the knowledge of variance properties, a new mathematical model of clustering validity index, VCVI, is proposed. Some concepts related to the new clustering validity index VCVI are defined as follows:

Definition 6. In Euclidean space R^m, a dataset containing n samples $D = \{x_1, x_2, \ldots, x_n\}$ is given. Of which, the clustering algorithm is used to divide D into K clusters, and the corresponding partition matrix is U_K. Then, CVI V_K is used to evaluate the effect of clustering. When clustering parameter K takes different values, the corresponding partition matrix of clustering results is $U_i (i = 1, 2, \ldots, n-1)$, its corresponding clustering index value is $V_i (i = 1, 2, \ldots, n-1)$. According to the property of CVI, the optimal clustering partition of the dataset is obtained:

$$U_m = U_2 \otimes U_3 \otimes \cdots \otimes U_{n-1} \tag{18}$$

Here $U_p = U_i$ oplus U_j indicates that the partition matrix of the optimal clustering result is assigned to U_p. Therefore, U_m in formula (18) is the optimal partition matrix of the dataset (that is, the optimal clustering partition of the dataset), and m is the value of the optimal clustering number K_{opt}.

Definition 7. Suppose in Euclidean space R^m, dataset $D = \{x_1, x_2, \ldots, x_n\}$ containing n sample points is given. The mean square sum of the Euclidean distance between all sample points in D to its central point v is called variance of the dataset D (marked as S).

$$S = \frac{1}{n}\sum_{i=1}^{n}(d(x_i, v))^2 \tag{19}$$

The value of S reflects the dispersion degree of the sample points in dataset D, which is the same as the one dimensional data variance principle. For this reason, this paper proposes to construct CVI by variance, and calculates the value of K_{opt} based on the minimum variance principle.

Definition 8. Suppose in Euclidean space R^m, dataset $D = \{x_1, x_2, \ldots, x_n\}$ containing n sample points is given. The clustering algorithm divides dataset D into K clusters $C = \{C_1, C_2, \ldots, C_K\}$, and then the corresponding clustering centers are $\{v_1, v_2, \ldots, v_K\}$ and v is the global clustering center of D. The mean square sum of the distance between the centers of each cluster to the global center is called the inter cluster variance (marked as G).

$$G = \frac{1}{K}\sum_{i=1}^{K}(d(x_i, v))^2 \tag{20}$$

Definition 9. Suppose in Euclidean space R^m, dataset $D = \{x_1, x_2, \ldots, x_n\}$ containing n sample points is given. The clustering algorithm divides dataset D into K clusters $C = \{C_1, C_2, \ldots, C_K\}$, and the corresponding clustering centers are $\{v_1, v_2, \ldots, v_K\}$. Then, the sum of squared distances between the sample points in the cluster C_i and the clustering center v_i is calculated as:

$$T_i = \sum_{j=1}^{|C_i|}(d(x_j, v_i))^2 \tag{21}$$

Definition 10. Suppose in Euclidean space R^m, dataset $D = \{x_1, x_2, \ldots, x_n\}$ containing n sample points is given. The clustering algorithm divides dataset D into K clusters $C = \{C_1, C_2, \ldots, C_K\}$. The intra cluster variance is defined as:

$$T = \frac{1}{n}\sum_{i=1}^{K}T_i \tag{22}$$

Definition 11. Suppose in Euclidean space R^m, dataset $D = \{x_1, x_2, \ldots, x_n\}$ containing n sample points is given. The clustering algo-

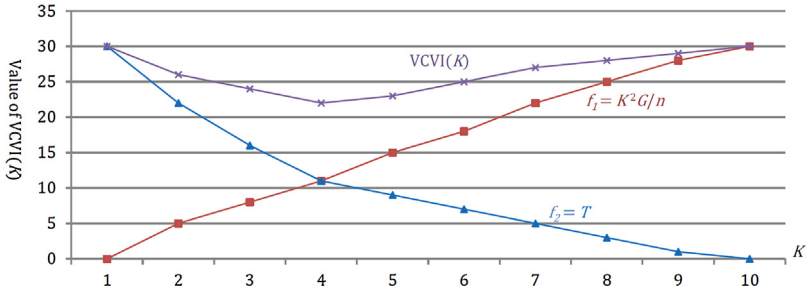

Fig. 2. The growth trend graph of the index function VCVI(K).

rithm divides dataset D into K clusters $C = \{C_1, C_2, \ldots, C_K\}$. The new clustering validity index based on variance is defined as:

$$VCVI(K) = \frac{1}{n}(K^2G + nT) = \frac{1}{n}\{K \times \sum_{i=1}^{K}(d(v_i, v))^2 + \sum_{i=1}^{K}\sum_{j=1}^{|C_i|}(d(x_j, v_i))^2\} \quad (23)$$

The value of the optimum clustering number Kopt is calculated by formula (24).

$$K_{opt} = \{K|\min_{2\le K\le n-1}\{VCVI(K)\}\} \quad (24)$$

Fig. 2 describes the change trend of CVI function VCVI(K) with clustering number K. From which, we can see that the VCVI index has two main components, where $f_1 = K^2G/n$ reflects the dispersion degree between clusters of datasets, and the value of f_1 grows with the increasing number of K (clustering number); $f_2 = T$ reflects the dispersion degree between sample points in all clusters in a dataset, the value of f_2 decreases with the increase of the clusters number K. From the mathematical knowledge it can be seen, usually, when the function value satisfies $f_1 = f_2$, the corresponding VCVI(K) index is the smallest. So in the process of solving the optimal clustering number for a dataset, the value of the optimal clustering number K_{opt} can be quickly determined by means of mathematical image analysis. This solution is not necessarily the optimal clustering partition for the dataset, but at least the near optimal clustering partition. Through the above analysis, we can draw the following conclusion: the new clustering validity index VCVI reflects the dispersion degree of inner and inter clusters. That is, the VCVI index takes into account the discrete degree of all samples in dataset D from the global (inter clusters) perspective and the local (intra cluster) perspective respectively.

Inference 1. Suppose in Euclidean space R^m, dataset $D = \{x_1, x_2, \ldots, x_n\}$ containing n sample points is given. Of which the clustering algorithm divides dataset D into K clusters $C = \{C_1, C_2, \ldots, C_K\}$. Let G stands for inter cluster variance and C stands for intra cluster variance, then $K_{\max} \le \sqrt{n}$.

Proof. Let $\bar{g}$ be the mean inter cluster variance: $\bar{g} = G/K$; $\bar{c}$ be the mean intra cluster variance: $\bar{c} = T/n$. The use of mathematical knowledge of spatial fractal geometry (in a graph, if the shape of a part is similar to that of the whole, then it is called a fractal), when the spatial structure of the sample points in each cluster (corresponds to the part in the spatial fractal geometry) and the spatial structure of center points of K clusters (corresponds to the whole in the spatial fractal geometry) have similar spatial fractal geometry features, we can get:

$$\bar{c}/(T/K) = \bar{g}/G \quad (25)$$

However, in practical applications, it is difficult to guarantee the spatial distribution of datasets with fractal geometry characteristics. So we only consider the generality and universality of the problem, that is, when the spatial distribution of a dataset follows the principle of within-cluster compactness, between-cluster separation, then the following relationship can be drawn:

$$\bar{c}/(T/K) \le \bar{g}/G \quad (26)$$

From the above analysis, we know:

$$G = K \times \bar{g}, \quad T = n \times \bar{c} \quad (27)$$

Meanwhile, when $G = T$, we can get:

$$K \times \bar{g} = n \times \bar{c} \quad (28)$$

Combining (26)–(28) and formula (23), we can immediately get: $K_{\max} \le \sqrt{n}$.

By now, the rationale of empirical rule $K_{\max} \le \sqrt{n}$ is proved by using the mathematical knowledge. So, in the clustering analysis, the range of cluster number K is: $2 \le K \le \sqrt{n}$. It can be seen from the proof part of inference 1 that the validity of empirical rule $K_{\max} \le \sqrt{n}$ can be proved by cluster validity index function VCVI(K). Furthermore, if the spatial distribution of the dataset after clustering follows the principle of within-cluster compactness, between-cluster separation, we can make the following conclusion: when the value of the VCVI index reaches the minimum, the value of K is the optimal number of clusters or at least the near optimal number of clusters.

4.3. K value optimization and determination algorithm based on new VCVI

Combining the empirical rule $2 \le K_{\max} \le \sqrt{n}$ with the new clustering validity index VCVI, a new K value optimization and determination algorithm based on VCVI index is proposed. The algorithm not only can remarkably compress the solution range of the cluster number K, but also has the ability to perform optimal cluster partition for different types of datasets, and then rapidly determine the value of optimal number K_{opt} of clusters. The K value optimization and determination algorithm based on the VCVI index is described in Fig. 3. In the algorithm, line (1) calculates the range of K_{max} by the number of sample points (n) in dataset $D = \{x_1, x_2, \ldots, x_n\}$ and the empirical rule $2 \le K_{\max} \le \sqrt{n}$. For different K in $[2, \sqrt{n}]$, lines (2)–(5) calculate the value of VCVI(K) based on the improved K-means algorithm (as shown in Fig. 1) and formula (23). Lines (6)–(13) select the minimal VCVI(K), $K \in [2, \sqrt{n}]$, and consequently, the corresponding optimal clustering number K_{opt} of dataset D is acquired. Lines (14) gives the optimal division C ($C = \{C_1, C_2, \ldots, C_{K_{opt}}\}$) of dataset D.

5. Experimental results

This section presents the details on the simulation for experimental evaluation of the performance on determining K_{opt} and performing optimal clustering partition based on the new VCVI proposed. The test environment of this section consists of an Intel Pentium CPU (E6700 at 3.2 GHz), 2.0 GB RAM and Windows 7 OS.

Input: Dataset $D = \{x_1, x_2, \ldots, x_n\}$.
Output: The optimal clustering number K_{opt};
The optimal clustering partition $C=\{C_1, C_2, \ldots, C_{K_{opt}}\}$ of dataset D.

Process:
(1) According to the number of sample points n in dataset $D = \{x_1, x_2, \ldots, x_n\}$ and the empirical rule, we can get: $2 \le K_{max} \le \sqrt{n}$;
(2) for $K = 2, 3, \ldots, \sqrt{n}$ do
(3) Using the improved K-means algorithm (as shown in Fig.1) on D;
(4) Evaluate the clustering result according to the new VCVI(K) described in formula (23);
(5) end for;
(6) Let min = VCVI(2);
(7) for $K = 3, 4, \ldots, \sqrt{n}$ do
(8) if $min >$ VCVI(K)
(9) then $min \leftarrow$ VCVI(K);
(10) $K_{opt} \leftarrow K$;
(11) else Keep min unchanged;
(12) end if;
(13) end for;
(14) The optimal clustering partition $C=\{C_1, C_2, \ldots, C_{K_{opt}}\}$ of dataset D is got when the clustering validity index reaches the minimal value. Consequently, the value of K is the corresponding optimal clustering number K_{opt}.

Fig. 3. K value optimization and determination algorithm based on VCVI index.

Table 1
Description of the datasets.

Dataset name	Points	Clusters	Dimensions	Range of K
4k2	400	4	2	[2, 20]
Aggregation	754	6	2	[2, 27]
S1	5000	15	2	[2, 70]
Iris	150	3	4	[2, 12]
Wine	178	3	13	[2, 13]
Hayes-Roth	132	3	5	[2, 11]

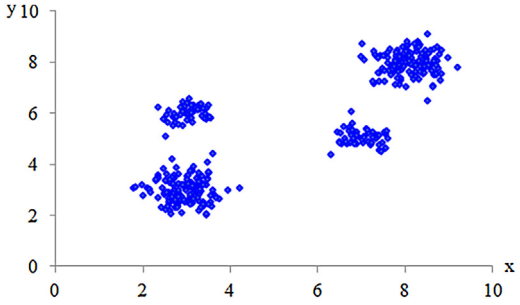

Fig. 4. Spatial distribution of 4k2 dataset.

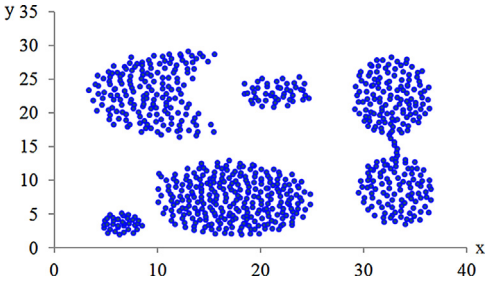

Fig. 5. Spatial distribution of Aggregation dataset.

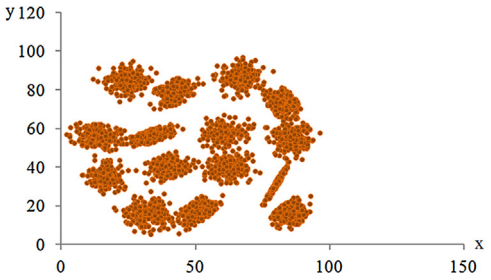

Fig. 6. Spatial distribution of S1 dataset.

Meanwhile, MyEclipse 8.6 with jdk 1.8 is selected for running our Java programs. In this section, three simulated datasets and three UCI machine learning datasets are selected to evaluate the clustering effect (efficiency and accuracy) of the new algorithm integrated with VCVI. Specifically, the three simulated datasets are 4k2, Aggregation and S1 (http://cs.joensuu.fi/sipu/datasets/); the three UCI machine learning datasets are Iris, Wine and Hayes-Roth (http://archive.ics.uci.edu/ml/datasets.html). A detailed description of the six experimental datasets is shown in Table 1. In Table 1, column "*Points*" specifies the number of sample points in each dataset; "*Clusters*" denotes the number of clusters in each dataset; "*Dimensions*" gives the dimension of each datasets; based on the number of "*Points*" in each dataset and the empirical rule $2 \le K_{\max} \le \sqrt{n}$, "*Range of K*" gives the range of K for each dataset.

5.1. Spatial distribution of datasets

Fig. 4 shows the spatial distribution of sample points in 4k2 analog dataset. From this figure, we can see that the 4k2 dataset can be divided into 4 clusters, and the sample distribution of this dataset is characterized by "within-cluster compactness, between-cluster separation". Fig. 5 describes the state graph of spatial distribution for the Aggregation dataset. From the figure, we can see that the dataset is comprised of 6 clusters, and each cluster has a uniform and compact distribution of sample points. However, the number of sample points contained by each cluster varies greatly, and the spatial distribution patterns of some clusters are also different from those of others. Fig. 6 shows the sample points spatial distribution of S1 dataset. This dataset is composed of 15 clusters. The sample distribution of S1 dataset has the feature of "within-cluster compactness". However, the sample distribution between some clusters has a small degree of cross and overlap.

Iris is a commonly used experimental dataset for clustering and classification. It is collected by Fisher based on the characteristics of iris plants. The Iris dataset is divided into 3 clusters, 50 samples per cluster, and each sample has four attribute values. Since the Iris dataset has four dimensions, it is needed to reduce the dimensions

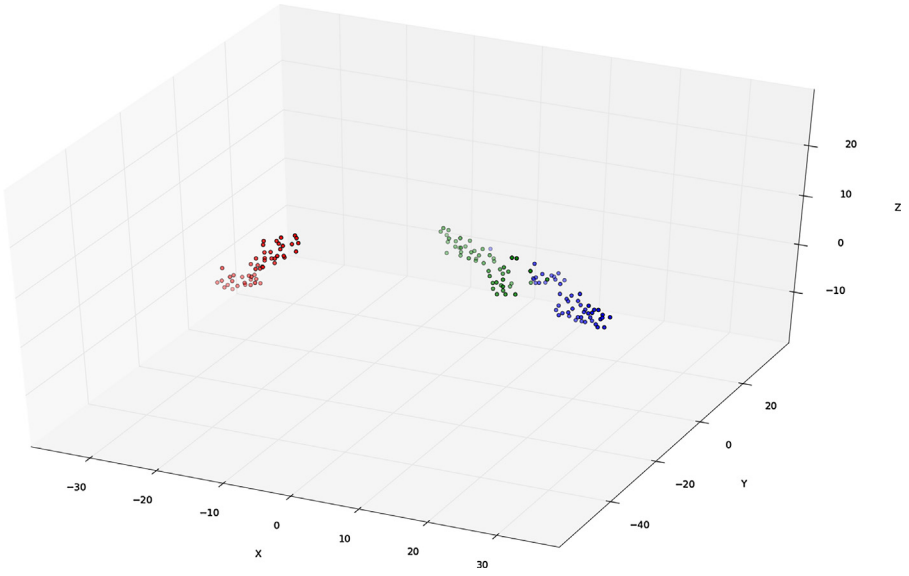

Fig. 7. Spatial distribution of S1 dataset.

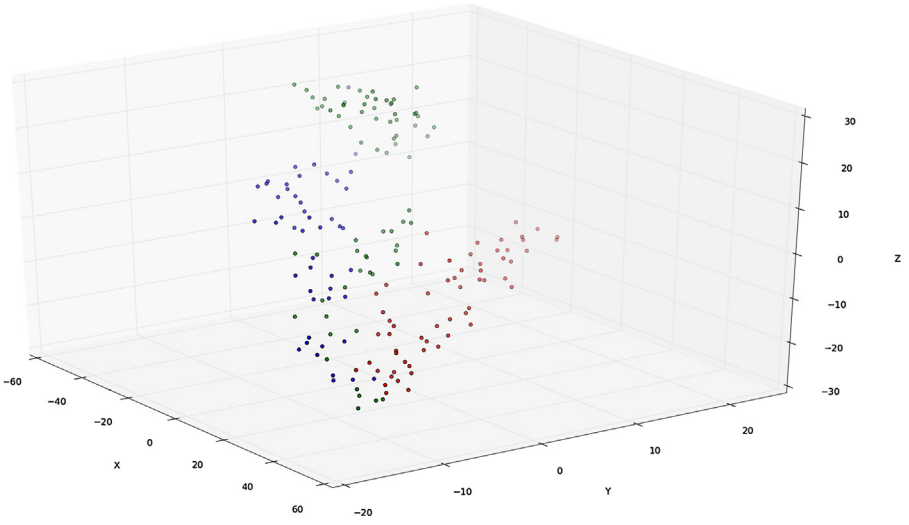

Fig. 8. Spatial distribution of S1 dataset.

to be displayed in the low dimensional space. High dimensional data visualization currently has: PCA, LDA (linear dimensionality reduction), MDS, Lsomap, SNE, T-SNE (nonlinear dimensionality reduction). In this paper, we select the widely used dimensionality reduction tool T-SNE [34]. From Fig. 7, we can see that the spatial distribution of two clusters in the Iris dataset has a small degree of overlap, while the other one is far apart from these two clusters and is linearly separable. The Wine dataset is a collection of data from three different kinds of red wines produced in a region from Italy. It contains 178 samples, each sample has 13 attributes, and each attribute corresponds to one of the chemical constituents of the red wine. Fig. 8 is a three-dimensional spatial distribution graph of the Wine dataset after dimensionality reduction by the T-SNE method. From the figure, we can see that the 3 clusters of the dataset are linearly inseparable, and there is a small degree of overlapping among clusters. The Hayes-Roth dataset is collected and reorganized by Barbara and Frederick. The optimal clustering number of this dataset is 3, and each sample point consists of five attributes: peoples name, hobby, age, education and marital status. Fig. 9 is a three-dimensional spatial distribution graph of the Hayes-Roth dataset after dimensionality reduction by the T-SNE method. Obviously, the dataset has a high degree of overlapping among clusters.

5.2. Clustering effect evaluation for different CVIs

For the 6 datasets listed in Table 1, clustering is performed by using our new algorithm when clustering number K is set as different values. At the end of each clustering, the results of our new algorithm are evaluated by our new VCVI. Meanwhile, the results of clustering processed by our new algorithm integrated with VCVI are compared with the results of traditional algorithms integrated with the other five widely used CVIs. According to the references that proposed these existing CVIs, we choose the hierarchical clustering algorithm for COP-index, K-means algorithm for DI-index, CH-index and DBI-index, and simulated annealing (SA) [35] for I-index. Because of too much difference among the values of the six CVIs, it is needed to standardize the CVI values to facilitate the display and analysis of experimental results. In this paper, the following method is adopted:

$$MaxI = \max_{2\le i\le\sqrt{n}}\{FI(i)\} \tag{29}$$

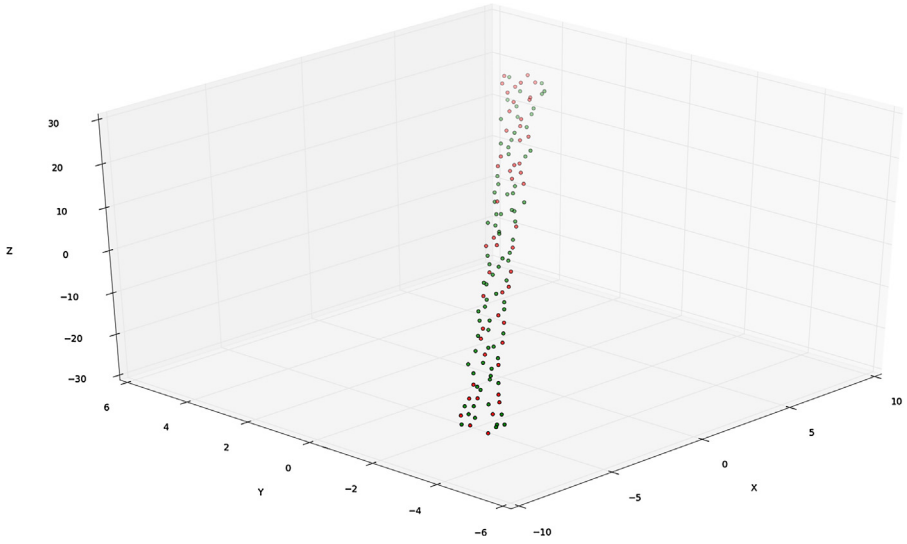

Fig. 9. Spatial distribution of S1 dataset.

Table 2
Standardized CVI values of 4k2 dataset.

K	DI-index	DBI-index	CH-index	I-index	COP-index	VCVI-index
2	**0.4716**	0.4351	1627.7	1.4803	0.2467	2.2367
3	0.1728	0.3272	1530.0	**1.5295**	0.1971	1.4567
4	0.3338	**0.3063**	**4155.1**	0.1787	**0.1630**	**0.6790**
5	0.0092	0.7085	3605.0	0.6361	0.2322	0.8750
6	0.0104	0.9523	3420.0	0.1399	0.2939	1.1574
7	0.0363	0.8158	3396.7	0.1776	0.2697	1.4963
8	0.0170	0.8241	3170.3	0.0961	0.2733	1.9137
9	0.0215	0.7576	3375.8	0.1232	0.2632	2.4280
10	0.0254	0.8054	3125.7	0.1022	0.2728	3.0131
11	0.0327	0.8329	3047.9	0.1142	0.2817	3.6350
12	0.0249	0.8124	2918.1	0.0378	0.2846	4.2294
13	0.0249	0.8204	2947.3	0.0433	0.2895	5.0271
14	0.0198	0.8713	2976.3	0.0202	0.3135	5.6198
15	0.0198	0.8396	2962.7	0.0261	0.3125	6.5390
16	0.0198	0.8282	2909.0	0.0201	0.3121	7.3813
17	0.0198	0.8083	2828.2	0.0244	0.3112	8.5140
18	0.0198	0.8005	2730.9	0.0103	0.3116	9.4893
19	0.0198	0.8161	2677.8	0.0123	0.3150	10.532
20	0.0198	0.8315	2557.4	0.0097	0.3178	11.591

$$FI_s(K) = \frac{FI(K)}{MaxI} \times 30, \quad 2 \le K \le \sqrt{n} \tag{30}$$

Of which, *FI(K)* is the clustering validity index function (like *DI(K)* in formula (4), *DBI(K)* in formula (6), *CH(K)* in formula (8), *I(K)* in formula (10), *COP(K)* in formula (12) and *VCVI(K)* in formula (23)). The inequality $2 \le K \le \sqrt{n}$ is the empirical rule proved in Section 4.2. *FIs(K)* is the standardized CVI values which will be displayed and analyzed in the following subsections. Through the standardization of the above methods, the values of the 6 CVIs will be limited to the interval of [0, 30].

5.2.1. CVI values of 4k2 dataset

Table 2 lists the CVI values of 4k2 dataset evaluated by 6 CVIs. In this table, the first column "*K*" gives the range of K for this dataset. Since there are 400 points in this dataset, the value of K is limited to the interval of [2,20] by the empirical rule of $2 \le K \le \sqrt{n}$. The remaining columns of this table list the CVI values calculated by different CVIs for different values of *K*. In this table, index values with underlined bold fonts specify the optimal cluster numbers K_{opt} calculated by different CVIs. As the optimal partition of 4k2 dataset is 4 ($K = 4$), so CVIs, DBI-index (4, 0.3063) (in this number pair, 4 is the optimal clustering number of this dataset when the value of DBI-index reaches 0.3063, the same as below), CH-index (4, 4155.1), COP-index (4, 0.1630) and VCVI (4, 0.6790) can get the optimal clustering partition. However, the DI-index (2, 0.4716) and I-index (2, 1.5295) cannot obtain the optimal clustering partition (that is, the best clustering numbers 2 and 3 are not the optimal clustering results of the 4k2 dataset).

For better illustration of the experimental results, Fig. 10 gives the standardized CVI values processed by formulas (29) and (30). Through the standardization of the methods of the two formulas and the empirical rule $2 \le K \le \sqrt{n}$, the CVI values of the tested datasets calculated by different CVIs are limited to the interval of [2,30]. Of which, the points that represent the optimal cluster numbers calculated by different CVIs are marked with number pairs. From this figure, we can see that CVIs, DBI-index (4, 9.6439), CH-index (4, 30), COP-index (4, 15.387) and VCVI (4, 1.7574) can get the optimal clustering partition. The DI-index (2, 30) and I-index (3, 30) cannot obtain the optimal clustering partition. In the remainder of this section, only the standardized results presented by figures are given for simplicity.

5.2.2. CVI values of Aggregation dataset

Fig. 11 gives the comparison of experimental results for six CVIs in solving the optimal cluster number on the Aggregation dataset. Through the standardization of the methods of formulas (29) and (30), CVI values calculated by different CVIs are limited to the interval of [0, 30]. The values of K are limited to the interval of [2,27] by the empirical rule $2 \le K \le \sqrt{n}$ (there are 754 sample points in this dataset). Then, in the range of *K*, the Aggregation dataset is clustered by using the improved K-means algorithm. Finally, VCVI and the commonly used five CVIs are used to evaluate the results of partition, so as to find the optimal clustering number K_{opt}. As a matter of fact, the optimal cluster number of the Aggregation dataset is 6. From Fig. 11 we can see that only the cluster validity index VCVI proposed in this paper can get the optimal clustering partition of this dataset. That is to say, we can get the optimal cluster number 6 for this dataset when the value of VCVI reaches 4.422 (6, 4.422). The COP-index (4, 23.798) and DBI-index (4, 18.772) can obtain the near optimal cluster partition. However, I-index (2, 30), CH-index (18, 30) and DI-index (24, 30) cannot get the optimal clustering number.

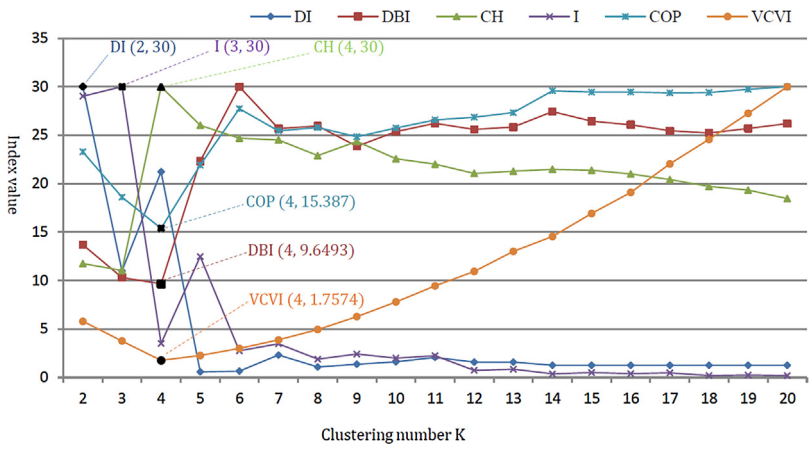

Fig. 10. Standardized CVI values of 4k2 data set.

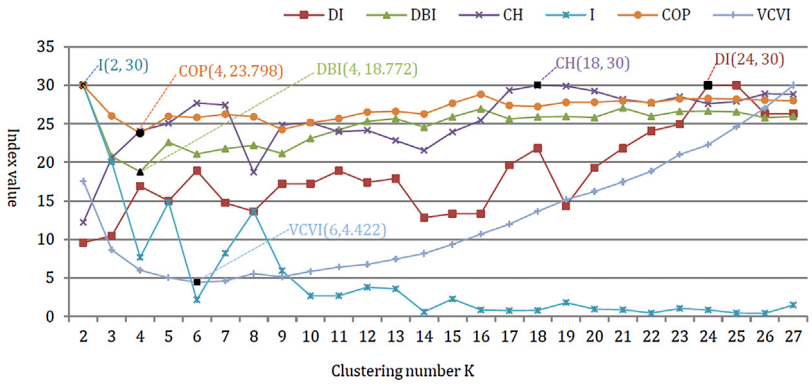

Fig. 11. Standardized CVI values of Aggregation data set.

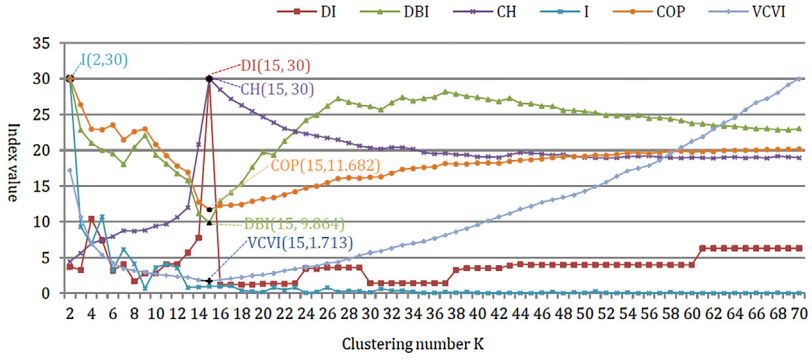

Fig. 12. Standardized CVI values of S1 data set.

5.2.3. CVI values of S1 dataset

The S1 dataset has 5000 sample points, so the range of the cluster number K is $2 \leq K \leq 70$. From the spatial distribution of S1 dataset shown in Fig. 6, we can derive that the optimal clustering number of S1 is 15. Fig. 12 shows the comparison of experimental results on clustering measurements for S1 dataset by different CVIs. According to the properties of the CVIs, we can see that the DI-index (15, 30), DBI-index (15, 9.864), CH-index (15, 30), COP-index (15, 11.682) and our VCVI (15, 1.713) can get the optimal clustering number. The optimal clustering number obtained by I-index is 2, which is not consistent with the actual optimal clustering of the dataset S1. So, I-index (2, 30) cannot get the optimal clustering number for this dataset.

5.2.4. CVI value of Iris dataset

The Iris dataset is divided into 3 clusters (as shown in Fig. 7), 50 samples per cluster, and each sample has four attribute values. From the comparative experimental results shown in Fig. 13, we can see that the Iris dataset can be optimally partitioned by CH-index (3, 30), I-index (3, 30) and our VCVI (3, 5.6871) index. The best clustering number obtained by DI-index is 10 (10, 30) or 11 (11, 30), and the optimal clustering number corresponding to DBI- index (2, 10.4231) and COP-index (2, 14.7133) is 2. Therefore, when the data is overlapped in a great degree, the DI-index cannot get the optimal clustering number. The DBI-index and COP-index merge clusters with larger overlapping into a single one, so they can obtain the near optimal cluster partition.

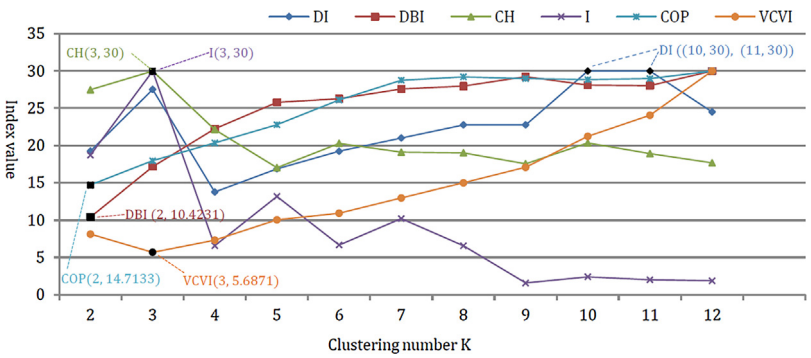

Fig. 13. Standardized CVI values of Iris data set.

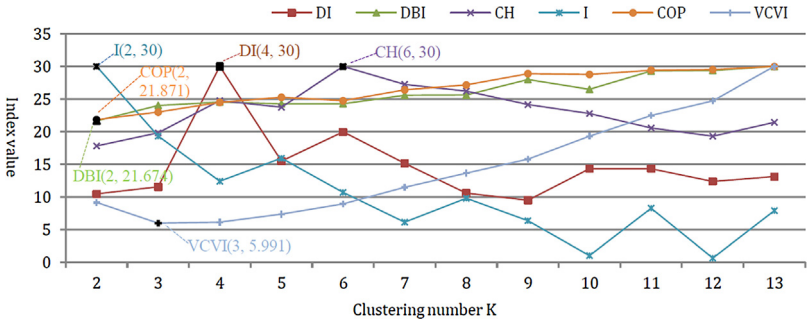

Fig. 14. Standardized CVI value of Wine data set.

5.2.5. CVI value of Wine dataset

In order to better demonstrate the experimental results on the Wine dataset, we also use formulas (29) and (30) to standardize the different index values. By the empirical rule $2 \le K \le \sqrt{n}$ (there are 178 sample points in this dataset), the range of the clustering number K is [2,13]. For different integer values of K in this interval, clustering is performed by the improved K-means algorithm firstly. Then the clustering results are evaluated by using different CVIs and the experimental results are shown in Fig. 14. From Fig. 8, we can see that there are 3 clusters in the Wine dataset. So, the results shown in Fig. 14 demonstrate that only the VCVI index (3, 5.991) proposed in this paper can get the optimal clustering number 3 for this dataset. The DI-index (4, 30), DBI-index (2, 21.674), I-index (2, 30) and COP-index (2, 21.871) can only get the near optimal clustering partition. However, the CH-index (6, 30) cannot get the best cluster number.

5.2.6. CVI values of Hayes-Roth dataset

As listed in Table 1, there are 132 sample points in the Hayes-Roth dataset. So, the range of K is limited to the interval of [2,11] by the empirical rule. Table 1 also tells us that the optimal clustering number of Hayes-Roth dataset is 3. From the comparative experimental results of 6 CVIs on the Hayes-Roth dataset shown in Fig. 15, we can see that our new VCVI (3, 5.721) can get the optimal clustering partition, the I-index (2, 30), DBI-index (2, 25.2311) and COP-index (2, 25.8663) can get the near optimal number of clusters. However, the DI-index (9, 30) and CH-index (11, 30) cannot get the optimal cluster number for this dataset.

5.2.7. Discussion

In general, the improved clustering partition algorithm, by using the new VCVI, can get the optimal cluster number K_{opt} for all different datasets tested. The COP-index and DBI-index, can obtain the optimal or the near optimal cluster number for all tested datasets. The other 3 CVIs cannot get optimal cluster number for all tested datasets. So, VCVI is stable in solving different distribution datasets.

Table 3
Clustering effect evaluated by different CVIs for tested datasets.

CVIs	Dataset					
	4k2	Aggregation	S1	Iris	Wine	Hayes-Roth
DI	No	No	Opt	No	Near	No
DBI	Opt	Near	Opt	No	Near	Near
CH	Opt	No	Opt	Opt	No	No
I	No	No	No	Opt	Near	Near
COP	Opt	Near	Opt	No	Near	Near
VCVI	Opt	Opt	Opt	Opt	Opt	Opt

Table 3 collects the results evaluated by the above six experiments. In the table, "*Opt*" and "*Near*" refer to the corresponding CVI which can get the optimal clustering number or near optimal clustering number respectively. "*No*" refers to the corresponding CVI which cannot get the optimal clustering number for certain dataset.

Since the DI-index is difficult to find the optimal number K_{opt} of clusters for the datasets with outliers, it cannot get K_{opt} for most of the tested datasets. Since the greater the overlap of dataset is, the worse the performance of DBI-index clustering evaluation the DBI-index cannot get K_{opt} for the Iris dataset. The CH-index also has such shortcoming on dealing with overlap of dataset. The I-index is suitable for dealing with some datasets with less class numbers. So, it cannot deal with Aggregation dataset and S1 dataset properly. The COP-index was first proposed to be used in conjunction with a cluster hierarchy post-processing algorithm. But it can also be used as an ordinary CVI. Accordingly, the smaller the COP-index is, the better clustering results of dataset. Since the improved K-means algorithm in this paper utilizes the density parameters based method to select the initial centers other than selecting them randomly, the stable clustering results can be obtained. The new VCVI is constructed from the point of view of spatial distribution of datasets. It precisely reflects the dispersion degrees of points within each cluster (local) or points inter clusters (global). By the global

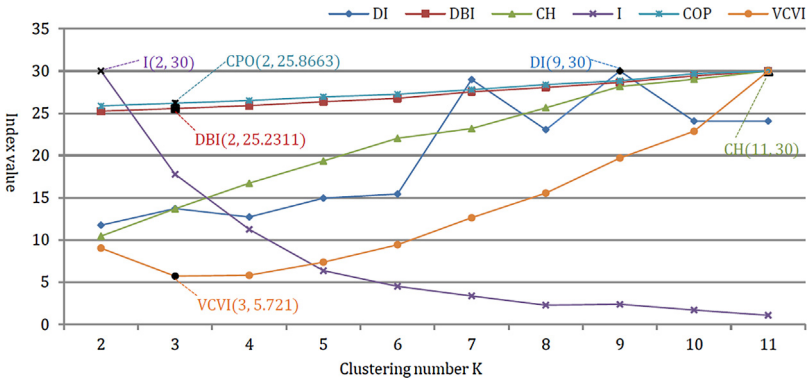

Fig. 15. Standardized CVI value of Hayes–Roth data set.

and local observations on the datasets, VCVI obtains the optimal clustering number stably.

5.3. Efficiency evaluation of 6 CVIs in processing 6 datasets

Because of too much difference among the execution time of the new algorithm integrated with VCVI and traditional algorithms integrated with other 5 CVIs, it is needed to standardize the execution time to display the experimental results properly. By utilizing the similar mechanism as formulas (29) and (30), the results of this experiment are limited to the interval of [0, 30]. For the 6 datasets listed in Table 1, clustering is performed by using our new algorithm when the clustering number K is set with different values. At the end of each clustering, the results of our new algorithm are evaluated by different CVIs. Actually, according to the references that proposed the 5 existing CVIs, different clustering algorithms are employed to process the datasets before these CVIs starting to work such as the COP-index employs the hierarchical clustering algorithm to process the tested datasets, the I-index employs the SA algorithm, the DI-index employs the K-means algorithm. The references that proposed CH-index and DBI-index do not incorporate the clustering algorithms. In this paper, all the datasets are processed only by our revised K-means algorithm for fair comparison. So the results shown in Fig. 16 are only the processing time by different CVIs (not include the execution time of the revised K-means algorithm).

Fig. 16 shows the standardized time costs of the 6 clustering indexes in solving the optimal clustering numbers of the 6 experimental datasets. In this experiment, the results listed in Fig. 16 are the mean time costs for a maximum 100 iterations processed by different CVIs. From this figure, we can see that the DI-index incurs the highest running time in solving K_{opt}, the running time of COP-index is the second highest, the DBI-index, CH-index, I-index and VCVI all need much less running time than the ones of DI-index and COP-index. Meanwhile, our VCVI has the lowest running time.

Table 4
Clustering effect evaluated by different CVIs for tested datasets.

Dataset	4k2	Aggregation	S1	Iris	Wine	Hayes-Roth
α	0.05	0.175	0.05	0.15	0.10	0.20

5.4. Accuracy evaluation of clustering by different CVIs

As mentioned previously, the choice of initial cluster centers has a great impact on the clustering results. Therefore, in order to reduce this influence and make the clustering results more stable, the traditional K-means is improved based on the density parameters for selecting the initial centers (as shown in Fig. 1). In dataset $D = \{x_1, x_2, \ldots, x_n\}$, the algorithm firstly calculates density parameters $\rho(x_i, \varepsilon)$ for a neighborhood of each sample point $x_i (i = 1, 2, \ldots, n)$. Then, it ascendingly orders the sample points with their density parameters. Lastly, the previous smallest K sample points $\{v_1, v_2, \ldots, v_K\}$ are selected as the initial clustering centers. In the process of selecting the initial cluster centers, the key step is how to choose a reasonable value of neighborhood of each sample point, that is, how to set the size of influence factor in formula (15). According to the spatial distribution characteristics of the six experimental datasets listed in Table 1, and after a lot of continuous tuning experiments on α, the reasonable value of α is listed in Table 4.

For traditional partitional clustering algorithms, different clustering results can be obtained by different options for initial clustering centers. In this part, the traditional K-means algorithm is firstly used to randomly select the initial clustering centers for each dataset listed in Table 1. By repeatedly performing experiments 500 times on the datasets, it is found that the selection of the initial clustering centers has a huge impact on the accuracy of the traditional K-means clustering algorithm. In Table 5 (as there is no label for each sample, we cannot evaluate the correctness of S1 dataset by different CVIs), the accuracy of each clustering result is statistically analyzed, and then the average clustering accuracy

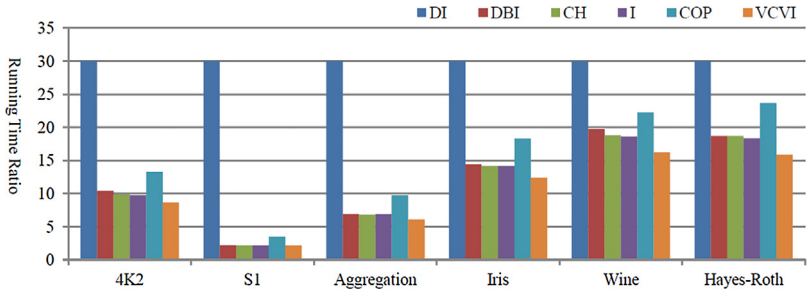

Fig. 16. Efficiency of the 6 CVIs in processing 6 tested data sets.

Table 5
Accuracy of traditional and improved K-means algorithms on processing different datasets.

Dataset	Method	
	Traditional	Improved
4k2	92.38%	100%
Aggregation	93.72%	96.83%
Iris	88.30%	94.67%
Wine	69.82%	80.22%
Hayes-Roth	47.00%	67.73%

Table 6
Accuracy of VCVI by another 10 UCI machine learning datasets.

Dataset	Points number	Attribute number	Clusters number	K_{opt}	Accuracy
Cancer	699	9	2	2	95.37%
Haberman	306	3	2	2	93.95%
Glass	214	9	7	6	67.85%
Parkinsons	195	22	2	2	88.38%
Ecoli	336	7	8	8	93.76%
Spectf	267	44	2	2	94.04%
Energy_efficiency	768	8	12	12	95.75%
statlog (German)	1000	24	2	2	89.00%
Ionosphere	351	34	2	2	83.39%
Libras movement	360	90	15	15	78.39%

of the traditional K-means algorithm is calculated (as listed in the 2*nd* column of the table). In this table, column "*Traditional*" stands for the accuracy of the traditional K-means algorithms (randomly select the initial cluster centers); column "*Improved*" are the accuracy of our improved algorithm (using density parameter based initial center selection method). From the experimental results, it can be explained that the traditional K-means algorithm is less stable and less accurate in clustering.

Table 5 (the 3rd column) also lists the experimental results of our new algorithm shown in Fig. 3. In the experiments, the data listed in Table 4 is used to set the sizes of α for different datasets. From the experimental results shown in Table 5, the subsequent conclusions can be drawn. For each clustering, the selection of the initial centers is fixed, therefore, the stability and robustness of the improved K-means algorithm is better than the random selection of the initial clustering centers K-means algorithm. And then, the improved algorithm is more accurate and more stable than the traditional ones. In order to further verify the accuracy of the new algorithm proposed, another 10 datasets from UCI machine learning datasets (http://archive.ics.uci.edu/ml/datasets.html) are tested. Table 6 lists the experimental results. Meanwhile, this table also shows that, except Glass, our VCVI can get all K_{opt} for the tested datasets.

6. Conclusion and future works

When the traditional cluster partition algorithms solve the clustering problems, it is necessary to set the value of the clustering number K in advance. But in practice, the numbers of clusters (K values) are usually in fuzzy intervals which seriously limit the further applications of the traditional cluster partition algorithms. In order to solve this problem, this paper first proposed an improved K-means algorithm based on density parameters for the initial centers selection, which can quickly find the cluster centers, improve the stability and accuracy of the clustering algorithm, and reduce the number of iterations of the algorithm. Then, a new clustering validity index, VCVI, was introduced to extend the application of the clustering partition algorithms. VCVI can find the optimal clustering number more accurately than some cluster validity indices proposed in the past. The experimental results by testing different datasets demonstrated that our new VCVI and algorithm can effectively get the optimal clustering number and the optimal clustering partition, especially for spatial distribution datasets with "within-cluster compactness, between-cluster separation". However, the proposed algorithm does not solve the intrinsic defects of the traditional K-means algorithm, so it still cannot process the dataset with a large number of noise points. The efficiency of the algorithm still needs to be improved when it is used to process large scale datasets. At the same time, the initial cluster centers are derived by calculating the average distance among all the sample points pairs. This method improves the accuracy of the algorithm, but at the expense of time efficiency. Finally, for different types of datasets, it is sometimes needed to select a reasonable neighborhood value for each sample. Therefore, these defects and deficiencies should be improved as our future work.

Acknowledgements

We are grateful for Professor Yun Yang from Swinburne University of Technology in Australia for English proofreading. This work was supported by the National Natural Science Foundation of China [Grant No. 61300169] and the Natural Science Foundation of Education Department of Anhui province (China) [Grant No. KJ2018A0022].

References

[1] J. Huang, Z.L. Yu, Z. Gu, A clustering method based on extreme learning machine, Neurocomputing 277 (2018) 108–119.
[2] F. Nie, X. Wang, M.I. Jordan, H. Huang, The constrained Laplacian rank algorithm for graph-based clustering, in: Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI 2016), Phoenix, USA, February 12–17, 2016, pp. 1969–1976.
[3] F. Nie, C. Ding, D. Luo, H. Huang, Improved MinMax cut graph clustering with nonnegative relaxation, in: Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2010): Part II, Barcelona, Spain, September 20–24, 2010, pp. 451–466.
[4] R. Xu, D. Wunsch II, Survey of clustering algorithms, IEEE Trans. Neural Netw. 16 (3) (2005) 654–678.
[5] D. Arthur, S. Vassilvitskii, K-Means++: the advantages of careful seeding, in: Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2007), New Orleans, Louisiana, USA, January 7–9, 2007, pp. 1027–1035.
[6] M. Erisoglu, N. Calis, S. Sakallioglu, A new algorithm for initial cluster centers in k-means algorithm, Pattern Recognit. Lett. 32 (14) (2011) 1701–1705.
[7] Y. Liu, Z. Ma, F. Yu, Adaptive density peak clustering based on K-nearest neighbors with aggregating strategy, Knowl. Based Syst. 133 (2017) 208–220.
[8] G. Tzortzis, A. Likas, The MinMax k-means clustering algorithm, Pattern Recognit. 47 (7) (2014) 2505–2516.
[9] J. Liang, X. Zhao, D. Li, F. Cao, C. Dang, Determining the number of clusters using information entropy for mixed data, Pattern Recognit. 45 (6) (2012) 2251–2265.
[10] S. Yue, J. Wang, J. Wang, X. Bao, A new validity index for evaluating the clustering results by partitional clustering algorithms, Soft Comput. 20 (3) (2016) 1127–1138.
[11] A. Ben Said, R. Hadjidj, S. Foufou, Cluster validity index based on Jeffrey divergence, Pattern Anal. Appl. 20 (1) (2017) 21–31.
[12] S. Angel Latha Mary, A.N. Sivagami, M. Usha Rani, Cluster validity measures dynamic clustering algorithms, J. Eng. Appl. Sci. 10 (9) (2015) 4009–4012.
[13] A. Filchenko, S. Muravyov, V. Parfenov, Towards cluster validity index evaluation and selection, in: Proceedings of the 2016 IEEE Artificial Intelligence and Natural Language Conference (AINL 2016), St. Petersburg, Russia, November 10–12, 2016, pp. 37–44.
[14] J.C. Bezdek, Pattern Recognition in Handbook of Fuzzy Computation, IOP Publishing Ltd, 1998.
[15] Y.-S. Chang, F. Nie, Z. Li, X. Chang, H. Huang, Refined spectral clustering via embedded label propagation, Neural Comput. 29 (12) (2017) 3381–3396.
[16] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybern. 3 (3) (1973) 32–57.
[17] D.L. Davies, B. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 1 (2) (1979) 224–227.
[18] T. Caliski, J. Harabasz, A dendrite method for cluster analysis, Commun. Stat. 3 (1) (1974) 1–27.
[19] U. Maulik, S. Bandyopadhyay, Performance evaluation of some clustering algorithms and validity indices, IEEE Trans. Pattern Anal. Mach. Intell. 12 (24) (2001) 1650–1654.
[20] I. Gurrutxaga, I. Albisua, O. Arbelaitz, J.I. Martin, J. Muguerza, J.M. Perez, I. Perona, SEP/COP: an efficient method to find the best partition in hierarchical

clustering based on a new cluster validity index, Pattern Recognit. 43 (10) (2010) 3364–3373.
[21] M.G.H. Omran, A.P. Engelbrecht, A. Salman, An overview of clustering methods, Intell. Data Anal. 11 (6) (2007) 583–605.
[22] A.M. Bagirov, J. Ugon, D. Webb, Fast modified global k-means algorithm for incremental cluster construction, Pattern Recognit. 44 (4) (2011) 866–876.
[23] F. Nie, Z. Zeng, I.W. Tsang, D. Xu, C. Zhang, Spectral embedded clustering: a framework for in-sample and out-of-sample spectral clustering, IEEE Trans. Neural Netw. 22 (11) (2011) 1796–1808.
[24] X.L. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell. 13 (8) (1991) 841–847.
[25] F. Nie, D. Xu, X. Li, Initialization independent clustering with actively self-training method, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 42 (1) (2012) 17–27.
[26] A.V. Kapp, Are clusters found in one dataset present another dataset? Biostatistics 8 (1) (2006) 9–31.
[27] F. Nie, X. Wang, H. Huang, Clustering and projected clustering with adaptive neighbors., in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2014), New York City, USA, August 24–27, 2014, pp. 977–986.
[28] G.W. Milligan, M.C. Cooper, An examination of procedures for determining the number of clusters in a dataset, Psychometrika 50 (2) (1985) 59–179.
[29] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J.M. Perez, I. Perona, An extensive comparative study of cluster validity indices, Pattern Recognit. 46 (1) (2013) 243–256.
[30] M. Ramze Rezaee, B.P.F. Lelieveldt, J.H.C. Reiber, A new cluster validity indexes for the fuzzy c-mean, Pattern Recognit. Lett. 19 (3–4) (1998) 237–246.
[31] N.R. Pal, J. Biswas, Cluster validation using graph theoretic concepts, Pattern Recognit. 30 (6) (1997) 847–857.
[32] T. Pei, A. Jasra, J. David, A. Hand, X. Zhu, C. Zhou, DECODE: a new method for discovering clusters of different densities in spatial data, Data Min. Knowl. Discov. 18 (3) (2009) 337–369.
[33] R.J.G.B. Campello, Generalized external indexes for comparing data partitions with overlapping categories, Pattern Recognit. Lett. 31 (9) (2010) 966–975.
[34] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008) 2579–2605.
[35] S. Bandyopadhyay, U. Maulk, M.K. Pakhira, Clustering using simulated annealing with probabilistic redistribution, Int. J. Pattern Recognit. Artif. Intell. 15 (2) (2001) 269–285.