A New Initialization Method for K-means Algorithm Based on Clustering Coefficient

Zhonghua Jiang*
Shanghai JingAn District College, Shanghai, China

*Corresponding author e-mail: jiangzh@sou.edu.cn

Abstract. The traditional k-means algorithm has been widely used as a simple and efficient clustering method. However, the algorithm uses a random initialization method that does not guarantee unique clustering results. In this paper, a novel method for selecting initial cluster centers on the basis of a complex network is proposed. The data set to be clustered is represented as a complex network. The clustering coefficients of a vertex and those of its adjacent vertices are used to introduce the novel concepts of “valley point” and “peak point,” which can be used to construct a set of candidates of the initial cluster centers. Finally, an algorithm for selecting the initial cluster centers from the constructed set of candidates is proposed. The time complexity of the proposed algorithm is $O(n^2)$, where $n$ is the number of data points. The proposed algorithm is applied to four data sets with different dimensions to compute the initial cluster centers for the k-means algorithm. Compared with the random initialization and maximin methods, the proposed algorithm demonstrates superior clustering performance in obtaining the initial cluster centers for the k-means algorithm.

Keywords: K-means, Complex Network, Initial Cluster Center, Clustering Coefficient, Degree

1. Introduction
The purpose of the clustering analysis is to divide data points into clusters such that the data points in the same cluster tend to be more similar [1]. The common clustering algorithms can be classified into partitional and hierarchical clustering algorithms [2]. The k-means algorithm has been widely applied in various fields due to its simplicity and efficiency. But its clustering result is prone to the initial cluster centers, thus this method tends to be trapped into local minima due to a bad initialization.

Many methods have been proposed to solve this problem. A kind of initialization methods aims at improving the method in which initial cluster centers are selected randomly. MacQueen selected the first $K$ objects as the initial cluster centers from the data [3]. However, this method is sensitivity to the ordering of data. The data point which has the largest minimum distance to the selected cluster centers is selected in the maximin method [4]. And this method is more likely to obtain the unstable clustering results since it selects the first center randomly. Jain and Dubes used the mean of the centers as the final cluster centers by applying the k-means algorithm multiple times [5]. Sun proposed a refinement...
algorithm for selecting the initial centers [6]. The second type of initialization method changes the
global search procedure into a computationally efficient local search procedure by shrinking the search
space for the initial centers. Redmond proposed an initialization method based on kd-trees [5]. Liang
proposed a new method for constructing potential exemplars [7]. The third type of initialization
method combines several optimization methods, such as genetic algorithms [8], density of the object
[9], and rough sets [10], with a k-means algorithm. In summary, a large amount of improved methods
have not been used widely because the determination of initial cluster centers can be considered as an
NP problem [11]. Therefore, it is required to further study the determination of initial cluster centers.

The small-world network model was proposed by Watts and Strogatz [12], inspiring scientific
research on complex systems in the context of complex networks. The data set to be clustered can be
represented as a complex network with vertices and edges constituted by data points and their
connecting relationships. The vertices and their corresponding measures, including the degree,
adjacent vertices, and clustering coefficients, can be exploited to illustrate the properties of the
complex network and further analyze the properties of data sets.

We propose a complex network method for the selection of initial cluster centers in this paper. The
objects in the data set are exploited to constitute a complex network, which is represented by an
undirected unweighted graph. A set of candidates for the initial centers is built by introducing the
concepts of “valley point” and “peak point” according to the clustering coefficient of the vertex and
those of its adjacent vertices. The initial cluster centers are determined by selecting data points from
the set of candidates according to specific selection criteria. Experimental results obtained using the
Wine, Iris, Balance-scale, and Soybean-small data sets show that the clustering performance of the
proposed algorithm is superior to other methods including random initialization and maximin methods.

2. Proposed Algorithm

In this section, we will describe the algorithm for the selection of the initial cluster centers on the basis
of a complex network. First, several concepts for converting the data set into a complex network are
presented.

2.1. Data set as a Complex Network

Let \( X \) be a data set with \( n \) data points and \( K \) clusters; that is, \( X = \{x_i \mid x_i \in R^n, i = 1, 2, \ldots, n\} \).

Next, we calculate the Euclidean distance between data points \( x_i \) and \( x_j \):

\[
d(x_i, x_j) = \sqrt{(x_i - x_j)^T (x_i - x_j)}
\]  

To apply the theory of complex networks to the current problem, data set \( X \) is denoted by the
undirected unweighted graph \( G=(V, E) \), where \( V=\{v_1, v_2, \ldots, v_n\} \). Each data point \( x_i \) in data set \( X \)
represents a vertex \( v_i \in V \) in graph \( G \). Therefore, The number of data points \( x_i \) in data set \( X \) is equal to
the number of vertices in graph \( G \). And data point \( x_i (i = 1, 2, \ldots, n) \) and vertex \( v_i (i = 1, 2, \ldots, n) \) keeps
one-to-one correspondence.

Suppose a complete graph in which each different vertex pair is connected by a unique undirected
edge has \( n(n-1)/2 \) edges. However, this type of graph increases the computational complexity by a
factor of \( n \). To avoid these potential problems, the threshold \( T \) is defined according to the average
distance between all pairs of data points.

\[
T = \frac{2}{n(n-1)} \sum_{i>j} d(x_i, x_j)
\]

The following equation determines whether there is an edge connecting vertices \( v_i \) and \( v_j \):

\[
e(v_i, v_j) = \begin{cases} \exists, & \text{if } d(x_i, x_j) < T \\ \emptyset, & \text{otherwise} \end{cases}
\]
The adjacency matrix of graph $G$ can be represented as:

$$ A_{ij} = \begin{cases} 1, & \text{if there exists an edge between vertices } v_i \text{ and } v_j \\ 0, & \text{otherwise} \end{cases} \quad (4) $$

The degree of a vertex is the total number of edges directly connected to that vertex. The degree is a critical characteristic of a vertex. In terms of the adjacency matrix $A_{ij}$ of graph $G$, the degree of a vertex can be defined as

$$ d_i = \sum_{j=1}^{n} A_{ij} \quad (5) $$

Suppose that the degree of vertex $v_i$ in the undirected unweighted graph $G$ is $d_i$; this implies that vertex $v_i$ has $d_i$ adjacent vertices, among which there can exist (at most) the maximum number of edges $d_i(d_i-1)/2$. If the actual number of edges among the adjacent vertices of the vertex $v_i$ is $E(i)$, the clustering coefficient of vertex $v_i$, denoted as $C(i)$, can be represented as

$$ C(i) = \frac{E(i)}{d_i(d_i-1)/2} \quad (6) $$

2.2. The Proposed Algorithm

The clustering coefficient indicates to what extent the neighbors of vertex $v_i$ are themselves neighbors, or to what extent the vertices adjacent to $v_i$ are also adjacent to each other. In terms of the data set $X$ with $K$ clusters, $K$ cluster centers are mostly located on the vertices with high clustering coefficients. However, provided that the vertices with high clustering coefficients are directly selected as the cluster center candidates, the most serious problem is that these vertices do not always belong to $K$ distinct clusters; that is, if the cluster number for the selected vertices is lower than $K$, bad clustering results will be obtained. Because the relationship between the vertices with high clustering coefficients and the cluster centers of the data set has multiple-to-one correspondence instead of one-to-one correspondence, the clustering coefficient cannot be considered as an indicator for selecting the initial cluster centers. Numerous clustering experiments revealed that the valley points and peak points can be defined to clarify this problem. Suppose that the clustering coefficient of vertex $v_i$ in graph $G$ is $C(i)$ and that $d_i$ clustering coefficients for the adjacent vertices of $v_i$ constitute the set of clustering coefficients $C_{v_i} = \{C_{v_i}(1), C_{v_i}(2), \ldots, C_{v_i}(d_i)\}$.

In the context of complex networks, whether the vertex $v_i$ can be selected as a candidate for an initial cluster center mainly depends on the value of $C(i)$ relative to the value of each element in $C_{v_i}$.

**Definition 1** Let $C(i)$ denote the clustering coefficient of vertex $v_i$ in the undirected unweighted graph $G$, and let $C_{v_i} = \{C_{v_i}(1), C_{v_i}(2), \ldots, C_{v_i}(d_i)\}$ denote the set of clustering coefficients of $d_i$ adjacent vertices of $v_i$. Vertex $v_i$ is then defined as a valley point if it satisfies $\nabla C_{v_i}(j) \geq C(i)$, where $C_{v_i}(j) \nabla C_{v_i}$.

**Definition 2** Let $C(i)$ denote the clustering coefficient of vertex $v_i$ in the undirected unweighted graph $G$ and let $C_{v_i} = \{C_{v_i}(1), C_{v_i}(2), \ldots, C_{v_i}(d_i)\}$ be the clustering coefficient set of $d_i$ adjacent vertices of $v_i$. Vertex $v_i$ is then defined as a peak point if it satisfies $\triangledown C_{v_i}(j) \leq C(i)$, where $C_{v_i}(j) \triangledown C_{v_i}$.

The procedure for defining vertex $v_i$ as a valley or peak point is essentially the procedure of illustrating the relationship between the clustering coefficients of a vertex and its adjacent vertices. The set constituted by valley points and their adjacent vertices, or the set constituted by peak points and their adjacent vertices, can be considered small clusters, and the data point corresponding to the valley point or peak point can be further considered the representative point of its small cluster. Based on these concepts, the initial cluster centers can be selected from these valley points and peak points.
To illustrate the proposed algorithm, the concept of distance between vertex and graph also should be defined.

**Definition 3** Suppose that vertex \( v_i \) satisfies \( v_i \notin V \), where \( V \) is the set of vertices in the undirected unweighted graph \( G \). The distance \( d(v_i, G) \) between vertex \( v_i \) and the undirected unweighted graph \( G=(V, E) \) can be defined as the minimum distance between vertex \( v_i \) and all vertices \( V=\{v_1, v_2, ..., v_n\} \) in graph \( G \); that is, \( d(v_i, G)=\min(d(v_i, v_1), d(v_i, v_2), ..., d(v_i, v_n)) \).

The algorithm for the determination of the initial cluster centers is described by the following steps.

**Algorithm 1:** The determination of the initial cluster centers

**Step 1:** Input the data set \( X \) with \( K \) clusters. The initial cluster centers are required to determine. Set \( p = 0 \) and \( k = 0 \).

**Step 2:** The data set \( X \) is represented as the undirected unweighted graph \( G \) according to Eqs. (1)–(3).

**Step 3:** Calculate the clustering coefficients for all the vertices in \( V \).

**Step 4:** Determine the set of valley points and set of peak points from graph \( G \) according to Definitions 1 and 2. A set of candidates for the initial cluster centers is constituted by the set of valley points and peak points.

**Step 5:** Select the valley point with the highest clustering coefficient (denoted as \( C_{\max}(i) \)) from the set of candidates as the first initial cluster center \( c_k \). (If more than one vertex meets the condition, count the number of its adjacent vertices whose clustering coefficients are greater than \( C_{\max}(i) \). The vertex with the largest number is selected first.) Update \( k = k + 1 \). Delete the selected vertex from the set of candidates. Update \( p = p + 1 \) and define \( G_p \) as the set of all \( c_k \); that is, \( G_p=\{13\} \) (For convenience, here we neglect the edges of \( G_p \)).

**Step 6:** Calculate the distances between \( c_k \) and the remaining vertices in the set of candidates. Update \( k = k + 1 \). Select the vertex with the largest distance as the second initial cluster center \( c_k \), and delete that vertex from the set of candidates. Update \( p = p + 1 \) and define \( G_p \) by union with \( G_{p-1} \); that is, \( G_p = \{c_k\} \cup G_{p-1} \).

**Step 7:** Calculate the distance between the remaining vertices in the set of candidates and graph \( G_p \) according to Definition 3. Update \( k = k + 1 \). Select the vertex with the largest distance as the \( k \)th initial cluster center \( c_k \), and delete the vertex from the set of candidates. Update \( p = p + 1 \) and constitute \( G_p \) from \( c_k \) and \( G_{p-1} \); that is, \( G_p = \{c_k\} \cup G_{p-1} \). If \( k < K \), repeat Step 7.

**Step 8:** Output \( K \) initial cluster centers.

Regarding Algorithm 1, the reason why the first initial cluster center is selected from the valley point set instead of the set of candidates can be explained by an example. Randomly generate a data set \( X = \{x_i | x_i \in R^2, i = 1, 2, ..., 15\} \) with \( K = 3 \) clusters, where each cluster contains 5 data points. Obtain the undirected unweighted graph \( G \) with 15 vertices \( V = \{a, b, c, ..., o\} \), as shown in Fig.1, where \( \triangle \), * , and + represent three clusters, respectively, and the gray line indicates the existence of an edge between two vertices. Calculate the clustering coefficients of the 15 vertices \( V = \{a, b, c, ..., o\} \) according to (6); these are noted beside the vertices. Determine the set of valley points \( \{a, g, j, l, m, n, o\} \) (blue circles) and set of peak points \( \{f, k\} \) (red circles) according to Definitions 1 and 2, respectively. Regarding the locations of the valley points and peak points, most valley points are located near the outer margins of the data set, whereas most peak points are located near the junctions between two clusters. Suppose that vertex \( a \) is selected from the set of valley points and vertex \( f \) is selected from the set of peak points as the first initial cluster center; the average distance between the remaining vertices in the set of candidates and vertex \( a \) is greater than the average distance between the remaining vertices and vertex \( f \). Furthermore, the peak points tend to be close to other peak points; for good clustering results, excessively close initial cluster centers should be avoided because that might lead to unfavorable clustering results. Therefore, the first initial cluster center is selected from the set of valley points in Algorithm 1.
3. Time Complexity
The time complexity of the proposed algorithm was analyzed as follows.

In Step 2, the time complexity for representing the data set $X$ as an undirected unweighted graph is $n(n-1)/2=O(n^2)$. Computation of the clustering coefficients of all vertices $V$ requires $O(n^2r)$ in Step 3, where $r$ is the average of vertex degrees. From Steps 4 to 7, the initial cluster centers are selected from the set of candidates, in which $m$ elements have a time complexity of $m(m-1)/2=O(m^2)$. Because $n >> m, n >> r$, the entire time complexity of the proposed algorithm is $O(n^3)$.

4. Experiments and Results
The proposed algorithm on the Iris, Wine, Balance-scale, and Soybean-small data sets from UCI data sets are evaluated which are shown in Table 1. We compared the results from the k-means method for the following initialization methods: the random initialization method, the maximin method, and the proposed method. In order to test the performance of our algorithm, five evaluation indexes were used to obtain the clustering validity index: Mirkin metric Index (MI), adjusted Rand index (ARI), accuracy (AC), Rand index (RI), and Hurber’s $\Gamma$ index (HI). Higher clustering validity index values indicate better clustering except MI; As for MI, the opposite is true. For the maximin method and random initialization method, the clustering results are the average results of 10 runs. Fig. 2 shows the values of the five evaluation indexes obtained from the four data sets by using the three initialization methods.

Fig. 2(a) and Fig. 2(c) show that, for the Wine and Balance-scale data sets, the proposed algorithm clearly outperformed the random initialization method and maximin method according to the five evaluation indexes. For the Iris data set, as shown in Fig. 2(b), the proposed algorithm outperformed the random initialization method and was comparable to the maximin method. For the Soybean-small data set, as shown in Fig. 2(d), the proposed algorithm was the same as or more effective than the random initialization method in terms of the evaluation indexes. However, the performance of the proposed algorithm was worse than that of the maximin method for RI, MI, and HI, yet it clearly outperformed the maximin method for ARI. For individual data sets, the performance of the proposed algorithm was slightly worse than that of the maximin method; overall, however, the proposed algorithm outperformed the random initialization method and maximin method for these five evaluation indexes.

![Figure 1. Graph G with 15 vertices](image-url)
Table 1. Data set description

| Data set     | # Attributes(D) | # Points(N) | # Classes(K) |
|--------------|-----------------|-------------|--------------|
| Wine         | 13              | 178         | 3            |
| Iris         | 4               | 150         | 3            |
| Balance-scale| 4               | 625         | 3            |
| Soybean-small| 35              | 47          | 4            |

Figure 2. Values of the five evaluation indexes on the (a) Wine, (b) Iris, (c) Balance-scale, and (d) Soybean-small data sets with the proposed algorithm, random initialization method, and maximin method

5. Conclusion
A novel algorithm for the choice of initial cluster centers based on a complex network is proposed in this paper. In this algorithm, valley points and peak points are selected from the vertices, and for constructing the sets of candidates. The vertices that are selected from the final set of candidates are the initial cluster centers. We compared the clustering results with k-means method for our algorithm, random initialization method, and maximin method on the Balance-scale, Wine, Iris, and Soybean-small data sets. With respect to the evaluation results, the performance of the k-means method with our algorithm is the best among the compared methods.

Acknowledgments
This work was financially supported by college scientific research development fund.
References
[1] N. Abdelhamid, A. Ayesh, F. Thabtah, Phishing detection based associative classification data mining, Expert Systems with Applications, 41 (2014) 5948-5959.
[2] Q. Xu, Q. Zhang, J. Liu, B. Luo, Efficient synthetical clustering validity indexes for hierarchical clustering, Expert Systems with Applications, 151 (2020) 113367.
[3] M.E. Celebi, H.A. Kingravi, P.A. Vela, A comparative study of efficient initialization methods for the k-means clustering algorithm, Expert Systems with Applications, 40 (2013) 200-210.
[4] J. Yang, Y. Ma, X. Zhang, S. Li, Y. Zhang, An initialization method based on hybrid distance for k-means algorithm, Neural Computation, 29 (2017) 3094-3117.
[5] S.J. Redmond, C. Heneghan, A method for initialising the K-means clustering algorithm using kd-trees, Pattern Recognition Letters, 28 (2007) 965-973.
[6] Y. Sun, Q. Zhu, Z. Chen, An iterative initial-points refinement algorithm for categorical data clustering, Pattern Recognition Letters, 23 (2002) 875-884.
[7] L. Bai, J. Liang, C. Dang, An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data, Knowledge-Based Systems, 24 (2011) 785-795.
[8] J. Lee, D.-W. Kim, An effective initialization method for genetic algorithm-based robot path planning using a directed acyclic graph, Information Sciences, 332 (2016) 1-18.
[9] K.M. Kumar, A.R.M. Reddy, An efficient k-means clustering filtering algorithm using density based initial cluster centers, Information Sciences, 418 (2017) 286-301.
[10] V.P. Murugesan, P. Murugesan, A new initialization and performance measure for the rough k-means clustering, Soft Computing, 24 (2020) 11605-11619.
[11] A. Hatamlou, Black hole: A new heuristic optimization approach for data clustering, Information Sciences, 222 (2013) 175-184.
[12] S.F. Muldoon, E.W. Bridgeford, D.S. Bassett, Small-world propensity and weighted brain networks, Scientific reports, 6 (2016) 1-13.
[13] A.R. Backes, O.M. Bruno, Shape classification using complex network and multi-scale fractal dimension, Pattern Recognition Letters, 31 (2010) 44-51.