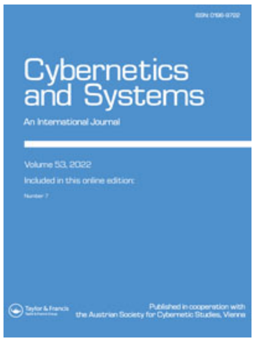


ISSN 0196-9722

Cybernetics and Systems

An International Journal

Volume 53, 2022

Included in this online edition:

Number 7

Taylor & Francis

Published in cooperation with
the Austrian Society for Cybernetic Studies, Vienna

Taylor & Francis
Taylor & Francis Group

Cybernetics and Systems

An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucbs20

A Hybrid Clustering Method Based on the Several Diverse Basic Clustering and Meta-Clustering Aggregation Technique

Bing Zhou, Bei Lu & Salman Saeidlou

To cite this article: Bing Zhou, Bei Lu & Salman Saeidlou (2022): A Hybrid Clustering Method Based on the Several Diverse Basic Clustering and Meta-Clustering Aggregation Technique, Cybernetics and Systems, DOI: 10.1080/01969722.2022.2110682

To link to this article: https://doi.org/10.1080/01969722.2022.2110682

Published online: 16 Aug 2022.

CYBERNETICS AND SYSTEMS
https://doi.org/10.1080/01969722.2022.2110682

Taylor & Francis
Taylor & Francis Group

A Hybrid Clustering Method Based on the Several Diverse Basic Clustering and Meta-Clustering Aggregation Technique

Bing Zhou[a], Bei Lu[a], and Salman Saeidlou[b]

[a]College of Information Engineering, Jiaozuo University, Jiaozuo, China; [b]School of Engineering, Technology and Design, Canterbury Christ Church University, Canterbury, UK

ABSTRACT

In hybrid clustering, several basic clustering is first generated and then for the clustering aggregation, a function is used in order to create a final clustering that is similar to all the basic clustering as much as possible. The input of this function is all basic clustering and its output is a clustering called clustering agreement. However, this claim is correct if some conditions are met. This study has provided a hybrid clustering method. This study has used the basic k-means clustering method as a basic cluster. Also, this study has increased the diversity of consensus by adopting some measures. Here, the aggregation process of the basic clusters is done by the meta-clustering technique, where the primary clusters are re-clustered to form the final clusters. The proposed hybrid clustering method has the advantages of k-means, its high speed, as well as it does not have its major weaknesses, the inability to detect non-spherical and non-uniform clusters. In the empirical studies, we have evaluated the proposed hybrid clustering method with other up-to-date and robust clustering methods on the different datasets and compared them. According to the simulation results, the proposed hybrid clustering method is stronger than other clustering methods.

KEYWORDS
Aggregation techniques; diversity of clustering; hybrid clustering; meta-clustering

1. Introduction

Nowadays, clustering plays an important role in most research fields such as engineering, medicine, biology, and data mining (Sun et al. 2018; Tan et al. 2020). Clustering is one of the fields of unsupervised learning and is an automatic process during which samples are divided into categories whose members are similar to each other, and these categories are called clusters. Therefore, a cluster is a collection of samples in which the samples are similar to each other and are not similar to the samples in other

CONTACT Bing Zhou zhoubjzu@126.com College of Information Engineering, Jiaozuo University, Jiaozuo, Henan, 454000, China; Salman Saeidlou salman.saeidlou@canterbury.ac.uk School of Engineering, Technology and Design, Canterbury Christ Church University, Canterbury, Kent, UK.

clusters (Wei et al. 2019; Trik, Pour Mozaffari, and Bidgoli 2021). Different criteria can be considered for similarity. For example, the distance criterion can be used for clustering and samples that are closer to each other can be considered as a cluster. This type of clustering is known as distance-based clustering. In simple words, the purpose is to separate groups with similar features and divide them into clusters (Yang et al. 2021; Ma et al. 2021).

Clustering methods take the data and form these groups using some kind of similarity criterion. The results obtained from these clusters/groups can be used on many applications such as image processing, pattern recognition, social network analysis, recommendation engine and information retrieval (Zhao et al. 2019). In the process of machine learning for clustering, a similarity measure based on distance plays a pivotal role in clustering decision (Ghobaei-Arani and Shahidinejad 2021). In all kinds of clustering methods, two main objectives should be considered in order to obtain the least error: one, the similarity between one data point with another point and the second, the distinction of those similar data points with other points (Forouzandeh et al. 2021; Berahmand et al. 2021). The basis for such divisions begins with our ability to scale large datasets, and this is a starting point. Another challenge in clustering is the different types of features in the data. Data can be structured, unstructured, hierarchical, and continuous (Ghobaei-Arani 2021; Shahidinejad, Ghobaei-Arani, and Esmaeili 2020). Also, it is evident that the data is not dimensionally limited and is multidimensional in nature.

Basically, a suitable distance measure can be very effective in clustering. However, the appearance of the clusters can be geometric, so this challenge must also be considered. On the other hand, the results of the clustering method should be understandable in order to solve business problems. Therefore, scalability, features, dimensions, appearance, noises, and interpretability are the things that clustering methods should consider to solve the problem (Nasiri et al. 2022; Jadidi and Dizadji 2021). In general, performing clustering using different methods have a similar architecture. This is while the differences among the clustering methods include the distance/similarity criteria, initial cluster values and how to form the final clusters. These differences have led to the development of different clustering methods over time. Basically, there are five main classes of clustering methods including Density-based Clustering (DC), Grid-based Clustering (GC), Model-based Clustering (MC), Hierarchical Clustering (HC), and Partitional Clustering (PC), as shown in Figure 1 (Wei, Li, and Zhang 2018).

Since most of the basic clustering methods emphasize on specific aspects of the data, they are efficient on specific datasets (Niu et al. 2020; Li, Qian, and Wang 2021). For this reason, there is a need for approaches that can create better results by using the combination of these methods and

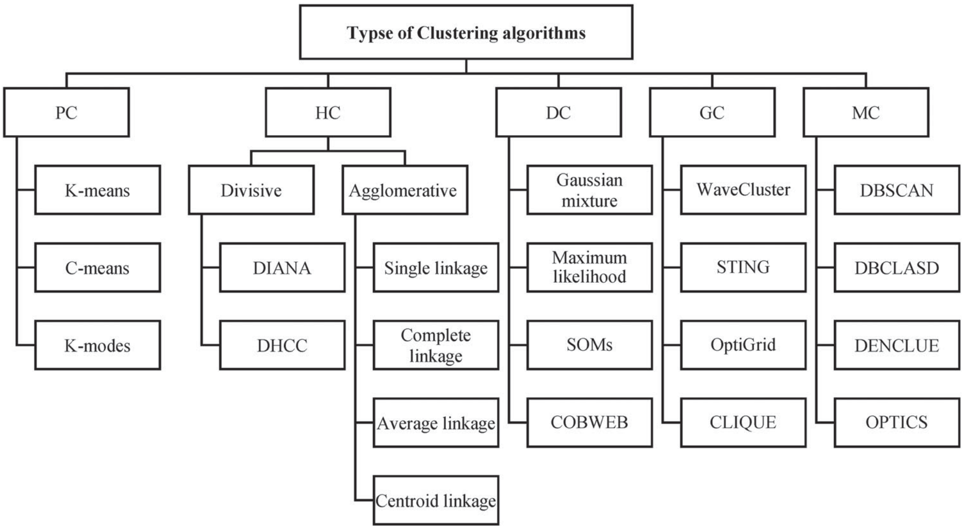


Figure 1. Taxonomy of clustering methods.

considering their strengths. Meanwhile, hybrid clustering is a new clustering method that is obtained by combining the results of different clustering methods. Accuracy, correctness, and stability are important characteristics of a hybrid clustering method compared to classical clustering methods (Zheng et al. 2021; Zhu et al. 2021; Tan et al. 2020). In fact, the main purpose of hybrid clustering is to search for better and stronger results, using the combination of information and results obtained from several primary clustering (as partitions). So far, many studies have been done on hybrid clustering. Recent research in this field has shown that data clustering can significantly benefit from the combination of several data parts. In addition, their parallelization power has a natural adaptation to the need of distributed data mining. Hybrid clustering can provide better solutions in terms of robustness, scalability, stability, and flexibility than basic clustering methods (as individual).

Basically, hybrid clustering includes two main steps: (1) producing different results from basic clustering methods and (2) combining the results obtained from basic clustering methods to produce final clusters (Zhu et al. 2021). The first is related to the creation of partitions with dispersion and diversity by different methods, and the second refers to an agreement function to combine the results (Wei et al. 2019). Usually, in the first step of hybrid clustering, a number of primary clusters are created, each of which emphasizes a specific feature of the data. Applying a clustering method on several different parts of the data or using several different clustering methods can cause dispersion and diversity in the partition results (Yang et al. 2021). After the primary partitions are formed, these results are usually combined by using an agreement function. One of the most common

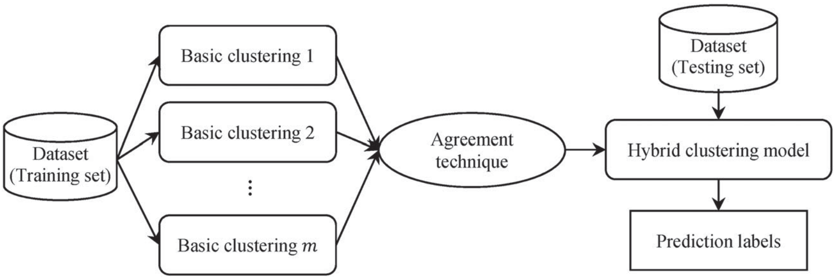


Figure 2. Hybrid clustering framework.

methods of combining the results is using the correlation matrix. A hybrid clustering framework is shown in Figure 2, where the results of several basic clustering methods are combined to achieve more stable, scalable, and quality clustering.

Therefore, nowadays instead of addressing the making a strong global clustering method, more attention has been paid to building frameworks that integrate several weak clusters (Zhao et al. 2019; Trik et al. 2022). In this regard, the "hybrid cluster" or "clusters aggregation", has been provided for improving the strength and quality of the clustering process (Tan et al. 2020). The k-means clustering method, which is one of the flat approaches, is known as a very fast and fairly efficient method (Yang et al. 2021; Ma et al. 2021). This method, as a weak clustering method, is one of the best basic clustering methods for contributing to consensus building in hybrid clustering. This paper addresses the existing problems by presenting valid local cluster theory. Here, the similarity between valid local clusters is estimated by applying an inter-cluster and intra-cluster similarity metric. In the next step of the method, the aggregation process of the basic clusters is done by the meta-clustering technique, where the primary clusters are re-clustered to form the final clusters. Eventually, the output of these clusters is considered along with the average credits to optimize the final agreement. The proposed hybrid clustering method has the advantages of k-means, its high speed, as well as it does not have its major weaknesses.

The main contribution of this paper is as follows:

- The aggregation process of the basic clusters with a new meta-clustering technique.
- Definition of valid local clusters by considering the data around the cluster centers in k-means.
- Generating diverse primary clusters by applying a duplicate strategy on nonappearance data in valid local clusters.
- Perform extensive experiments to demonstrate the efficacy of the proposed clustering method and give credence to our idea.

The rest of the paper is organized as follows. A brief discussion of related works in the literature is provided in Section 2. The formulation of the problem is provided in Section 3. The proposed clustering method is presented in Section 4. Experimental results are demonstrated in Section 5. Finally, Section 6 concludes the paper.

2. Related Works

So far, many studies have been presented by the research community on the development of clustering methods (Jain 2010; Hansen and Mladenović 2001; Zhang, Hsu, and Dayal 2000). The k-means method is one of the popular clustering approaches with many improved versions. For example, H-means solves the empty cluster problem in k-means (Jain 2010; Walid et al. 2021). Problems of k-means such as outliers, sensitive to noise and local optimum are considered by J-means method (Hansen and Mladenović 2001). This method can also solve the problem of degeneracy in k-means. Jiang et al. (2010) proposed K-Harmonic Means (KHM) to solve the primary clustering problem in k-means. KHM has succeeded in obtaining high-quality results by considering the harmonic mean of intervals as the objective function. However, KHM is not suitable for global optimization. In this regard, Swarm Intelligence techniques are being developed to replace KHM. The ACOKHM (Ant Colony Optimization and K-Harmonic Means) method for clustering with a global approach was presented by Bouyer and Hatamlou (2018). Although ACOKHM provides high-quality and accurate results, it has a slow convergence to the global optimum.

Hybrid clustering has become very popular as a technique to improve clustering results. The results of hybrid clustering using basic clustering methods with higher diversity and more quality are far more accurate (Bouyer and Hatamlou 2018). However, obtaining more accurate results by having more diversity in some collections has not yet been proven (Azimi and Fern 2009). Link-based Cluster Ensemble (LCE) was proposed as a hybrid clustering method by Jain (2010). LCE is an improved version of Hybrid Bipartite Graph Formulation (HBGF) in which bipartite graph is used. The authors first create a dense graph for each pair of samples and clusters and then form the final clusters using spectral clustering. Niu et al. (2020) proposed a hybrid clustering method that they developed based on the hybrid of locally reliable cluster solutions. This method is configured based on k-medoids and provides the concept of valid local clusters. Here, weighted undirected graph is used to find relationships between clusters.

Huang, Wang, and Lai (2017) proposed Locally Weighted Meta-Clustering (LWMC) to improve hybrid clustering methods. Here, the Jaccard coefficient is used to calculate the weight of connections between clusters. LWMC uses the normalized cut method to create meta-clusters, where each meta-cluster contains several clusters (Huang et al. 2020). The authors use a weighted voting-based technique to create the final clusters. Consensus clustering by partitioning similarity graph was proposed by Hamidi, Akbari, and Motameni (2019). This method uses graph pruning for clustering, where the number of clusters is automatically estimated. The authors use meta-cluster and majority vote as an aggregation function to create the final clusters. Here, the Jaccard coefficient is used to calculate the similarity. Iterative Combining Clustering Method (ICCM) was proposed by Khedairia and Khadir (2022). ICCM uses an iteratively based technique to analyze data and create primary clusters. Here a voting method is used to create a set of partitions. For this, each sample votes for its own sub-cluster so that samples with higher votes are assigned to the corresponding sub-clusters. In the meantime, the samples that do not get the highest vote are clustered in the next iterations.

The hybrid clustering is still considered as a tool as well as a research field of the theory studied. A review paper is presented by Golalipour et al. (2021) for a variety of these methods. Due to the fact that precision in clustering does not have a straightforward meaning such as classification, an alternative concept is presented for it, which states that a precise clustering is clustering which is most similar to other clusters formed on the given data, in other words, a better clustering means a more stable clustering. For a reason similar to the reason for the suitability of a diverse collection of classifiers for hybrid classification, a set of the clustering is considered as a goof set, if its basic clustering is varied (Bai, Liang, and Cao 2020). In order to generate a diverse clustering consensus, a weak clustering method must be applied to the data several times.

We use the k-means clustering method as a weak cluster for solving this problem (Abapour, Shafiesabet, and Mahboub 2021). Four sub-problems in hybrid clustering are presented as follows: (1) The problem of recognizing relatively correct labels in clustering: Unlike categorization, there is no real information about labels in clustering. (2) The problem of obtaining a variety of clustering that describe the entire data: In hybrid learning, while several poor learners are combined as strong learners, whatever the basic learners more complement each other, the hybrid learner acts better (Rezaeipanah, Nazari, and Ahmadi 2019; Rezaeipanah et al. 2021). That is, any weak clustering will cover the rest of the clustering. Therefore, for this purpose, we need to create several complementary clustering by applying k-means clustering methods. (3) The problem of determining the appropriateness between clusters: Unlike classifications in which each label is exclusively assigned to a category, the labels do not have a single meaning in clustering, and they simply represent that data has the same cluster

(Mojarad et al. 2021). The clusters with the same name in two different clustering do not imply any truth. Therefore, before doing anything in hybrid clustering, the label of different clustering should be re-labeled based on correspondence. In addition, even two clusters of the same clustering are likely to signify a real cluster. (4) The problem of combining the results of matched basic clustering: In different clustering, each sample may have different labels. So, we have to determine a final label called an agreement label. In hybrid learning, while several poor learners are combined as a strong learner, whatever the action is more effective, the hybrid learner acts better (Li, Rezaeipanah, and El Din 2022).

3. Problem Formulation

A dataset is defined as a set of data samples that each data sample itself is a numerical vector (or feature vector). The dataset is shown by X and each data sample is shown by x_i and obviously $x_i \in X$. The j-th feature of the x_i data sample is shown by x_{ij}. The size of each dataset X is shown by $|X|$. The number of features of the dataset X is shown by $|x_1|$. Let N be the number of samples and M the number of features from a dataset. Let c be the subset of data as clustered/partitioned. When the union of all subsets is equivalent to the original data set and each pair of subsets has no intersection, then each subset can be defined as a cluster. A clustering is shown by $\pi = \{\pi^1, \pi^2, ..., \pi^c\}$, where π^i represents the i-th cluster. Obviously, $\cup_{\mathrm{i}=1}^{\mathrm{c}} \pi^{\mathrm{i}} = \mathrm{X}$ and $\forall i, j \in \{1, 2, ..., c\} : \pi^i \cap \pi^j = \emptyset$. The center of each cluster π^i is shown by C^{π^i}, and its j-th feature is defined as Eq. (1)

$$C_j^{\pi^i} = \frac{\sum_{k \in \pi^i} x_{kj}}{|\pi^i|} \tag{1}$$

A valid sub-cluster from a cluster π^i is shown by r_{π^i} and is defined according to Eq. (2)

$$r_{\pi^i} = \left\{ x_k : \pi^i \middle| \sqrt{\sum_{j=1}^{|x_1|} \left| C_j^{\pi^i} - x_{kj} \right|^2} \leq \gamma \right\} \tag{2}$$

where γ is a threshold parameter. It should be noted that a sub-cluster can be considered as a cluster.

Basically, there are many similarity/distance measures in the literature to define the difference between two clusters. In this paper, we define the similarity metric between the two clusters π^i and π^j, which is shown by $sim(\pi^i, \pi^j)$, and defined as Eq. (3)

$$sim\left(\pi^i, \pi^j\right) = \begin{cases} \frac{\pi^i \cap \pi^j}{\pi^i \cup \pi^j} + \frac{\bigcup_{\mathrm{q}=1}^{9} \mathrm{T_q}\left(\pi^{\mathrm{i}},\ \pi^{\mathrm{j}}\right) - \left(\pi^{\mathrm{i}} \cup \pi^{\mathrm{j}}\right)}{\sqrt{\sum_{w=1}^{|x_1|} \left|C_w^{\pi^i} - C_w^{\pi^j}\right|^2}} & \sqrt{\sum_{w=1}^{|x_1|} \left|C_w^{\pi^i} - C_w^{\pi^j}\right|^2} \leq 4\gamma \\ 0 & Otherwise \end{cases} \tag{3}$$

where $T_q\left(\pi^i, \pi^j\right)$ is calculated using Eq. (4)

$$T_q\left(\pi^i, \pi^j\right) = \left\{ x_k : X \middle| \sqrt{\sum_{w=1}^{|x_1|} \left|p_{qw}\left(\pi^i, \pi^j\right) - x_{kw}\right|^2} \leq \gamma \right\} \tag{4}$$

where $p_q\left(\pi^i, \pi^j\right)$ is a point and w-th feature is denied as Eq. (5)

$$p_{qw}\left(\pi^i, \pi^j\right) = \frac{(q) \times C_w^{\pi^i} + (10 - q) \times C_w^{\pi^i}}{10} \tag{5}$$

Let $X = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$ be a set of n samples of the dataset X, where $x_i = \left[x_1^i, x_2^i, \ldots, x_j^i, \ldots, x_d^i\right]$ is an i-th sample with d features. Also, let $\Pi = \{\pi_1, \pi_2, \ldots, \pi_k, \ldots, \pi_m\}$ be a hybrid of m individual clustering method, where π_k is the k-th member of the hybrid. Each $\pi_k \in \Pi$ returns a set of clusters $\pi_k = \left[c_1^k, c_2^k, \ldots c_l^k, \ldots, c_{|\pi_k|}^k\right]$ (as a partition), where $|\pi_k|$ refers to the number of clusters created by π_k. For each $x_i \in X$, $\pi_k(x_i)$ represents the cluster label belonging to x_i in π_k. Here, the problem of hybrid clustering is defined as finding a new partition $\pi_* = [c_1^*, c_2^*, \ldots c_l^*, \ldots, c_K^*]$ from the consensus results of set Π, where K is the number of final clusters.

A weighting graph corresponding to a consensus of the clustering is shown by Π with $G(\Pi)$ and is defined as $G(\Pi) = [V(\Pi), E(\Pi)]$. The vertex set of this graph is also the valid subsets of all consensus's clusters, namely, $V(\Pi) = \left\{r_{\pi_1^1}, \ldots, r_{\pi_1^{c_1}}, r_{\pi_2^1}, \ldots, r_{\pi_2^{c_2}}, \ldots r_{\pi_B^1}, \ldots, r_{\pi_B^{c_B}}\right\}$. The weight of the edges between the vertices of this graph or the cluster-cluster connections is considered as the similarity value, as shown in Eq. (6)

$$E(v_1, v_2) = sim\left(v_j, v_i\right) \tag{6}$$

Basically, the k-means clustering method is considered as an unsupervised learning method, where it is used to process unlabeled data. The purpose of this clustering is to find the best group in the data and k determines the number of clusters. The data is placed in clusters based on the degree of similarity. In such a way that the data with the most similarity are placed in one group and have the least similarity with other groups. Here, k specifies the number of clusters and means the averaging. Clusters have a number of characteristics. The first feature: all the data in a cluster must be most similar to each other. The second feature: the data in different clusters should have the greatest difference. The time complexity of the k-means method is $O(N.k.I)$, so that I is the number of iterations.

The pseudocode for k-means-based hybrid clustering is shown in Algorithm 1. In this pseudocode, the original dataset is saved as TX and then an improved version of k-means is called sequentially to find and store the clustering results.

Algorithm 1. The hybrid clustering method based on k-means method.

```
$\Pi = \emptyset$;
$TX = X$;
For $i = 1$ to $B$ do
    $\pi_i =$ modified k-means $(TX, c_i)$;
    $TX = TX - \bigcup_{\mathrm{j}=1}^{\mathrm{c_i}} \mathrm{r}_{\pi_{\mathrm{i}}^{\mathrm{j}}}$;
    $\Pi = \Pi \cup \{\pi_i\}$;
End
```

Since the difference between basic clustering is a prerequisite for the effectiveness of the cluster group, in the following, how to obtain several k-means clustering with different valid local labels will be discussed. For the first time, we define an optimization problem for generating basic clustering as Eqs. (7) and (8)

$$\underset{\pi}{min}\left[z(\pi) = \sum_{h=1}^{N}\sum_{i=1}^{T}\theta_h(X_i)\ \leftthreetimes(\pi_{\mathrm{h}}\ (\mathrm{X_i}))\mathrm{d}(\mathrm{X_i}\ ,\ \ \mathrm{V}_{\pi\mathrm{h}}\ (\mathrm{X_i}))\Big)\right] \tag{7}$$

$$\sum_{h=1}^{T}\theta_h(X_i)\ \leftthreetimes_{\mathrm{h}}(X_i) = 1\ ,\ \ 1 \le i \le N \tag{8}$$

where $\theta_h(X_i)$ is a Boolean variable that if is equal to 1, X_i will partly play a role in the production of the basic cluster h. $\theta_h(X_i)$ is provided to control this issue that how many times do each sample play.

Here, constraint is required to each of the samples is applied only once simultaneously to produce basic clustering that is provided by cluster centers in clustering. The purpose of minimizing the objective function Z is to create cluster centers in each basic cluster to indicate that samples are in the valid local and possible spaces. We suggest an incremental learning method for solving the optimization problem. This method gradually produces the productive basic clustering by trying to optimize an incremental problem in each step. The incremental problem is as Eq. (9). Given that Π has gained the first basic cluster $g(0 < g < T)$.

$$Min\ Z(\ \Pi' \ \cup \{\pi_{g+1}\} \tag{9}$$

In addition, θ_{h+1i} is estimated through Eq (10)

$$\theta_{h+1i} = \begin{cases} 1, & \sum_{h=I}^{g} \leftthreetimes_{\mathrm{h}}(X_i) = 0, \\ 0, & \qquad otherwise \end{cases} \tag{10}$$

where $1 \le i \le N$.

Given this constraint, we see that samples which are obtained by cluster centers and not shown in Π play an important role in basic clustering $g + 1$. The incremental learning method is as follows: We first set $h = 1$, $\theta_h(X_i) = \ 1$ for $1 \le i \le N$ and $S = X$. At each step, we select k samples as the primary cluster centers from S randomly and use k-means with limitation for its cluster. In the clustering method, the cluster centers are limited, which can be seen only in relation to their neighborhoods in Eq. (9). This will cause the final cluster centers obtained to show samples in local spaces to be valid. After executing k-means, we will update $S = S - S'$, where S' is a set of samples that have valid local labels in the k_h basic clustering. Additionally, we will update $h = h + 1$, if $x_i \in S$, then $\theta_h(X_i) = 1$, otherwise, for $1 \le i \le N$, it will be 0.

The above method repeats until the number of samples in S is less than k_h^2. Updating the cluster centers at each step through the iteration mechanism leads to the production of the final cluster centers. It can guarantee data description by multiple clustering. On the other hand, the importance of satisfying the final conditions should be determined. Many researchers argued that the maximum number of clustering in the set S of samples should be less than $\sqrt{|S|}$ (Zheng et al. 2021; Zhu et al. 2021; Tan et al. 2020). Thus, while the number of samples in S is less than k_h^2, we assume that S cannot be divided into k_h clusters. Finally, if these conditions are met, the repetition can be stopped.

Algorithm 2. Pseudocode of the MKM scheme.

```
Input: $X,\ k,\ \varepsilon$
Output: $\Pi,\ V$
$\Pi = \emptyset, V = \emptyset, S = X, h = 0;$
$\theta_h(X_i) = 1,\ for\ 1 \le i \le N;$
Randomly select $k_h$ primary cluster centers as $v_h$ on $S$;
While $F < F'$ do
    $F' = F;$
        Given $v_h$, $\hat{\pi}_h$ is updated by $argmin_{l=1\ldots k_h} d(X_i,\ v_{h_l})$
    For $X_i \in\ S;$
```

```
            Given $\pi_h$, $\hat{v}_h$ is updated by $\hat{v}_h = {}^{\sum_{X_i \in D} X_i} \Big/ {}_{|D|}$,
            where $D = \left|\left\{\pi\ (X_i) = l\ \bigwedge\ X_i \in B(v_{hl}), X_i \in S\ \right\}\right|$
            For $1 \leq l \leq k_h$;
                $F = \sum_{l=1\ \ \pi_{\mathrm{h}}(X_i)=l\ \ ,X_i\ \in\ S}^{k_h} \sum d(X_i, v_{hl})^2$
                $S' = \left\{\lambda_{\mathrm{h}}(X_i) = 1,\ X_i \in S\ \right\}$;
            End For
            For $i = 1$ to $N$ do
                If $X_i \in S'$ then
                    $\theta_{h+1}(X_i) = 0$;
                else
                    $\theta_{h+1}(X_i) = \theta_h(X_i)$;
                End If
            End For
            $\Pi = \Pi \cup \{\pi_h\}$;
            $V = V \cup v_h$;
            $S = S - S'$;
        End For
End While
```

The incremental method is called the Modified k-means (MKM) clustering method, which is formally described in Algorithm 2. The time complexity of MKM, $O(Nt\ Tk_h)$, where T is the number of partitions generated. The outputs of the algorithm have been the clustering set $\Pi = \{\pi_h,\ 1 \leq h \leq T\}$ and also the set of cluster centers, which is equal to $V = \{v_h,\ 1 \leq h \leq T\}$. In order to simplify the basic clustering generation process, we determine a number of clusters in each basic clustering as k, $k_h = k$, $1\ \leq h \leq T$. We continue the following example in Figure 3. Here, we obtain a set of data as $\varepsilon = 0.8$ on the dataset as well as 10 cluster bases. Part (d) shows the partition lines of these basic clusters generated by the

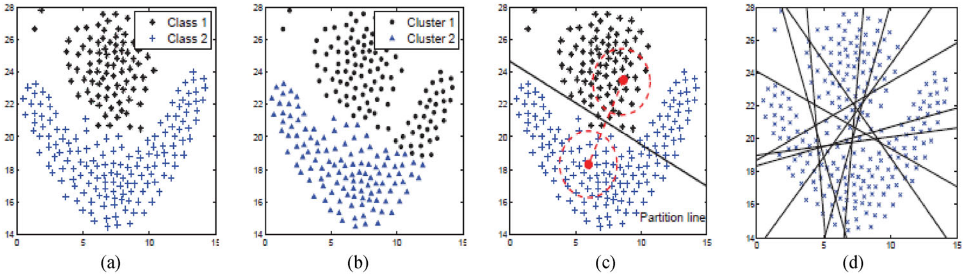


Figure 3. An example of MKM: (a) real class labels, (b) clustering from k-means, (c) local hypothesis of the clusters, and (d) multiple partitions by MKM.

MKM scheme. We observe that these basic clusters are somewhat different, which is useful for the cluster group.

Note that the number of T basic cluster depends on the parameter ε. When the amount of ε decreases, the T value must be increased, because a small amount of ε indicates that each basic cluster contains a number of local modifications. Therefore, while the ε is set to a smaller value, we need a more basic clustering to describe the whole data. The setting of ε depends on the needs of users, so that users can set the parameter to control the basic cluster number based on their needs.

4. Proposed Clustering Method

This study has provided a hybrid clustering method. This study has used the basic k-means clustering method as a basic cluster. Also, this study has increased the diversity of aggregation by adopting some measures. Here, the aggregation process of the basic clusters is done by the meta-clustering technique, where the primary clusters are re-clustered to form the final clusters. The proposed hybrid clustering method has the advantages of k-means, its high speed, as well as it does not have its major weaknesses.

In general, the labels in the dataset represent classes, but the labels in clustering only represent groups. Therefore, the labels in the clustering cannot be used for comparisons and cluster analysis. In this regard, it is necessary to align labels in clustering. Additionally, since the k-means method can only detect spherical and uniform clusters, two of the same clustering can represent a clustering. Hence, analysis of the relationship between clusters through similar clustering in needed. Now, there are inconsistent measures among the clusters proposed in the research literature (Yang et al. 2021; Ma et al. 2021). An example of this can be seen in chain clustering, where the intersection between clusters is determined by the distance between the farthest/closest sample between two clusters (Zhao et al. 2019). This method is sensitive to noise because it depends on a few specific samples to determine the final clusters. On the other hand, the distance between centers in center-based clustering approaches is defined as the absence of correlation. This method does not have the ability to effectively identify the border between clusters, but it has high computational efficiency and is resistant to noise.

In general, the similarity between two clusters in different partitions can be estimated based on the number of samples belonging to those clusters. This strategy cannot reflect samples with wrong labels in the cluster. However, some of these samples can have a high impact on the similarity calculation. Also, two clusters from the same partition share no sample, which is the reason for the inability of this metric to calculate similarity.

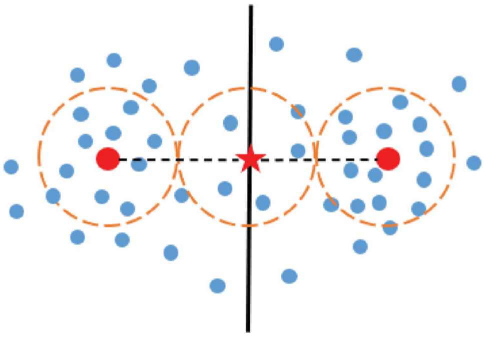


Figure 4. Hidden cluster between clusters.

Although there is good practice coordination between measures, they are not suitable for hybrid clustering. As mentioned, the labels of the created base partitions are different from the valid local labels. In other words, the validity of labels of each cluster may be low or high. Hence, the calculation of differences between clusters should be considered based on local labels. However, due to the use of MKM to generate initial partitions, the overlap between local labels should be relatively small. In this regard, we use an indirect overlap technique to calculate the similarity between clusters.

If c_{h_l} and c_{g_i} are two clusters, V_{h_l} and V_{g_i} are their cluster centers and $(V_{h_l} + V_{g_i})/2$ is the middle point of two centers. We assume there is a hidden cluster c_z whose cluster center is $(V_{h_l} + V_{g_i})/2$ for hidden for the cluster. Let the probability of samples being in valid local locations be greater with the density of samples. If there is a hidden cluster and the distance between V_{h_l} and V_{g_i} is not greater than $4 \times \varepsilon$, valid local spaces from the clusters c_{h_l} and c_{g_i} are overlapping with the hidden clusters c_z, as shown in Figure 4. In this case, the valid local spaces c_{h_l} and c_{g_i} are indirectly overlapping with the hidden cluster. For clusters c_{h_l} and c_{g_i}, we consider these parameters to estimate the similarity between clusters.

The distance between cluster centers is estimated based on the probability of a hidden cluster between them. As we know, whatever $d(V_{h_l}, V_{g_i})$ is smaller, the valid local spaces between them and c_z will more overlap. In this respect, it is a fact that their similarity is inversely related to $d(V_{h_l}, V_{g_i})$. Also, k-means is a clustering approach with a linear mechanism and can identify the border of two clusters through a line between their centers. If the range around them is among several samples, they can be clearly identified. We use the following example in Figure 5.

It is clearly seen that clusters B and C have centers with larger distances compared to clusters A and B. Meanwhile, it is easier to determine the border between clusters A and B. Hence, the distance between the centers of clusters A and B may be increased considering the clarity of the boundary

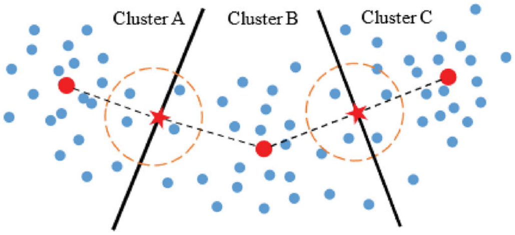


Figure 5. Similarity between clusters.

identification. According to this hypothesis, let the similarity between two clusters be estimated through a hidden cluster. Formally, similarity is measured as Eq (11)

$$\delta(c_{h_l}, c_{g_i}) = \begin{cases} \dfrac{\left| B\left(\dfrac{V_{h_l} + V_{g_i}}{2}\right) \right|}{d(V_{h_l}, V_{g_i})} & d(V_{h_l}, V_{g_i}) \leq 4 \times \varepsilon \\ 0 & Otherwise \end{cases} \tag{11}$$

Given the defined similarity matrix, we use an undirected weight graph (e.g., $G =< A, W >$) to describe the relationships between clusters. In this graph, A refers to the set of nodes that represent the cluster labels in Π. On the other hand, W in G refers to the weight of edges, which expresses the similarity between clusters. Hence, the similarity of both clusters is the concept of the weight of the edges between them, for example, $x,\ y\ \in\ A$, $w_{Xy} = \delta(c_x, c_y)$, and whatever there is similarity between them. By calculating the weighted graph, the relationships between the clusters can be mapped to the normal graph discharge challenge, which is as Eq. (12)

$$\underset{\Omega}{min}\left[Q(\Omega) = \frac{1}{k}\sum_{l=1}^{K} \frac{\sum_{x\in A_l\ ,\ y\in A-A_l} w_{xy}}{\sum_{x\in A_l\ ,\ z\in A} w_{xz}} \right] \tag{12}$$

where $\Omega = A_l,\ \forall l = 1, 2, ..., k$ is a partition of nodes in G and A_l is one of the subsets of A.

Our goal is to measure this partition using the minimization of the objective function Q. This is achieved by creating a partition that has high similarity between nodes in similar subsets and low similarity with nodes in other subsets. To solve this problem and create partition A, the normalized spectral clustering method has been used, where nodes in similar subsets represent a cluster. Hence, if $L(c_x)$ is the label of the subset which c_x belongs to it, then we will have $L(C_x) = l$, if $C_x\ \in A_l$. If $1 < l < k$ and $x \in A$, the time complexity of the making of the cluster relationship is $O\left(N(T.k_h)^2\right)$.

The use of hybrid clustering leads to the mapping of the clustering problem from the sample level to the cluster level. Assume that PC is a set containing all primary clusters created from all basic methods. Taking each cluster as a sample, the clustering process is applied again, where this time the clusters are clustered. This technique can create meta-clusters, where each meta-cluster contains several clusters. Meta-clusters have more knowledge about the data than clusters because they combine the latent knowledge from different clustering methods. Here, the clusters' clustering method is done using k-means. Let the similarity of two samples from the available dataset be $s(x_i, x_j)$. Anyway, in meta-clusters the concept of similarity is extended from the sample level to the cluster level. We define the similarity measure of clusters in a meta-cluster through Eq. (13)

$$\Psi(mc_\alpha, mc_\beta) = \frac{1}{|mc_\alpha|.|mc_\beta|} \sum_{v=1}^{|mc_\alpha|} \sum_{w=1}^{|mc_\alpha|} \left[\frac{\sum_{i=1}^{|c_v|} \sum_{j=1}^{|c_w|} \Gamma(x_i, x_j)}{|c_v|.|c_w|} \right] \quad \forall x_i \in c_v,\ x_j \in c_w \tag{13}$$

where mc_α and mc_β are two meta-clusters, and $\Psi(mc_\alpha, mc_\beta)$ refers to the average similarity between them. Also, $|mc_\alpha|$ and $|mc_\beta|$ are the number of clusters in mc_α and mc_β, respectively. Moreover, $|c_v|$ and $|c_w|$ describe the number of samples in c_v and c_w, respectively.

We create the final clusters by considering meta-clusters, where each instance of the dataset is assigned to a meta-cluster with maximum similarity. Meanwhile, the number of suitable clusters can be recognized by merging the initial clusters and applying a threshold value. Therefore, k is determined as the number of optimal clusters by merging the initial clusters until the threshold θ is reached, as defined in Eq. (14). In other words, clusters are merged until the similarity of each existing pair of clusters is greater than θ.

$$if\ \ \sigma(c_a, c_b) \geq\ \theta \Longrightarrow \begin{cases} hence\ merged & True \\ not\ merged & False \end{cases}, \forall a, b \in P\ C \tag{14}$$

where c_a and c_b are two clusters of the PC. Also, $\sigma(c_a, c_b)$ defines to the average similarity between c_a and c_b.

5. Experimental Results

This section is related to the evaluation of the proposed clustering method based on four synthetic datasets and five real datasets. Here, the efficiency of the proposed method is evaluated through the analysis of different validation methods and runtime. The evaluation of the proposed method is compared with some state-of-the-art methods such as COllaborative-Single Link (CO-SL) (Fred and Jain 2005), COllaborative-Average Link (CO-AL)

(Fred and Jain 2005), Combined Similarity Measure-Single Link (CSM-SL) (Iam-On et al. 2011), Combined Similarity Measure-Average Link (CSM-AL) (Iam-On et al. 2011), Weighted Triple Quality-Single Link (WTQ-SL) (Iam-On et al. 2011), Weighted Triple Quality-Average Link (WTQ-AL) (Iam-On et al. 2011), Weighted Connection Triple-Single Link (WCT-SL) (Iam-On et al. 2011), Weighted Connection Triple-Average Link (WCT-AL) (Iam-On et al. 2011), Meta-Clustering Algorithm (MCLA) (Strehl and Ghosh 2002), HyperGraph Partitioning Algorithm (HGPA) (Strehl and Ghosh 2002), Cluster-based Similarity Partitioning Algorithm (CSPA) (Strehl and Ghosh 2002), Selective Voting (SV) (Zhou and Tang 2006), Selective Weighted Voting (SWV) (Zhou and Tang 2006), Iterative Voting Consensus (IVC) (Nguyen and Caruana 2007), Expectation–Maximization (EM) (Topchy, Jain, and Punch 2005), Normalized Spectral Clustering (NSC) (Ng, Jordan, and Weiss 2001), Density Based Spatial Clustering of Applications with Noise (DBCAN) (Ester et al. 1996), and Clustering by Fast Search and Find of Density Peaks (CFSFDP) (Rodriguez and Laio 2014).

5.1. Experiment Settings

A number of settings for these compared methods are listed below to ensure that the comparisons are in uniform environment. The number of clusters per basic cluster is equal to the actual number of classes in each of the desired datasets. k-means is also used as a productive of basic clustering. There are two methods for basic clustering: (1) Multiple implementations of the k-means $T,$ each with a random amount of cluster centers. Let N refer to the number of samples. A set of the group T is set based on the dataset scale. If $N \leq 500,$ then $T = 25,$ if $500 \leq N < 1,000,$ then $T = 45$ and if $N \geq 1,000,$ then $T = 15.$ (2) Implement the proposed method. The method requires an input parameter $\varepsilon,$ which is adapted to the size of the T group, as required. Here, the size of the group T is essentially consistent with the first plan.

We implemented all these methods with MATLAB 2019a simulator for experiments. The simulations are based on a synthetic fog environment on the Dell Latitude Laptop with Intel$^{®}$ Atom$^{\text{TM}}$ processor N550 (Core i7 at 3.3 GHz) and 16 GB of RAM. Meanwhile, the proposed method has some parameters as input whose values are adjusted using Taguchi approach (Yang et al. 2021).

5.2. Evaluation Criteria

Given the availability of real labels from the original dataset, we use two common measures based on unsupervised learning to estimate the similarity between the results and the correct division of the dataset of different

Table 1. The default table to compare two partitions.

C P	p_1	p_2	$p_{k'}$	Sums
c_1	n_{11}	n_{12}	$n_{1k'}$	b_1
c_2	n_{21}	n_{22}	$n_{2k'}$	b_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
c_k	n_{k1}	n_{k2}	$n_{kk'}$	b_k
Sums	d_1	d_2	$d_{k'}$	–

methods. Given a dataset X and two partitions of these samples, namely $C = \{c_1, c_2, \ldots, c_k\}$ (clustering result) and $P = \{p_1, p_2, \ldots, p_{k'}\}$ (real partition), the values associated with C and P can be provided in a contingency table (Table 1), so that n_{ij} indicants the number of same nodes in the clusters c_i and p_j : $n_{ij} = |c_i \cap p_j|$.

Normalized Mutual Information (NMI) and Adjust Rand Index (ARI) are evaluation metrics in experiments. The details of these criteria are described below.

5.2.1. NMI

Let $\pi_\alpha = \left[c_1^\alpha, c_2^\alpha, \ldots, c_{|\pi_\alpha|}^\alpha\right]$ and $\pi_\beta = \left[c_1^\beta, c_2^\beta, \ldots, c_{|\pi_\beta|}^\beta\right]$ be the results of two basic clustering methods as two partitions with $|\pi_\alpha|$ and $|\pi_\beta|$ clusters, respectively. Accordingly, the $\mathrm{NMI}(\pi_\alpha, \pi_\beta)$ defines the diversity value for these partitions (Li, Qian, and Wang 2021), as shown in Eq. (15)

$$\mathrm{NMI}(\pi_\alpha, \pi_\beta) = \frac{2\sum_{i=1}^{|\pi_\alpha|}\sum_{j=1}^{|\pi_\beta|} n_{ij}\log\left(\frac{n.n_{ij}}{n_{i\alpha}.n_{\beta j}}\right)}{\sum_{i=1}^{|\pi_\alpha|} n_{i\alpha}\ \log\left(\frac{n_{i\alpha}}{n}\right) + \sum_{j=1}^{|\pi_\beta|} n_{\beta j}\ \log\left(\frac{n_{\beta j}}{n}\right)} \tag{15}$$

where n is the number of samples, n_{ij} is the same number of samples in c_i^α and c_j^β, $n_{i\alpha}$ is the number of samples in c_i^α, and $n_{\beta j}$ is the number of samples in c_j^β.

5.2.2. ARI

This measure is often used in cluster validation and can indicate agreement between two partitions (Niu et al. 2020). The ARI is calculated based on the Rand Index, as defined in Eq. (16)

$$\mathrm{ARI}(\pi_\alpha, \pi_\beta) = \frac{\sum_{i=1}^{|\pi_\alpha|}\sum_{j=1}^{|\pi_\beta|}\binom{n_{ij}}{2} - \left[\sum_{i=1}^{|\pi_\alpha|}\binom{n_{i\alpha}}{2}\sum_{j=1}^{|\pi_\beta|}\binom{n_{\beta j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_{i=1}^{|\pi_\alpha|}\binom{n_{i\alpha}}{2} + \sum_{j=1}^{|\pi_\beta|}\binom{n_{\beta j}}{2}\right] - \left[\sum_{i=1}^{|\pi_\alpha|}\binom{n_{i\alpha}}{2}\sum_{j=1}^{|\pi_\beta|}\binom{n_{\beta j}}{2}\right] / \binom{n}{2}} \tag{16}$$

Table 2. Description of the datasets used.

Type dataset	Datasets	Number of samples	Number of features	Number of clusters
Artificial dataset	Imbalance	2,250	2	2
	Aggregation	788	2	7
	Banana	2,000	2	2
	Ring	1,500	2	3
Real dataset	Wine	178	13	3
	Iris	150	4	3
	Digits	5,620	63	10
	WBCD	569	30	2
	KDD-CUP99	1,048,576	39	2

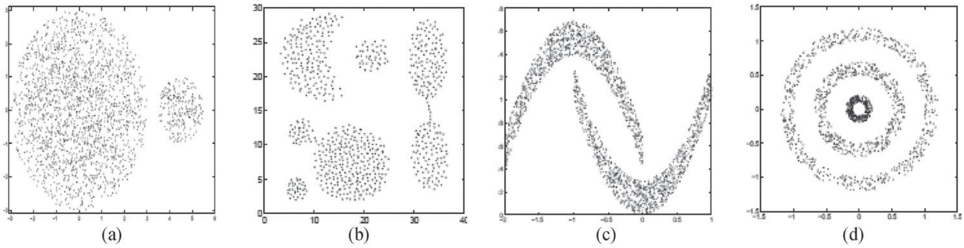


Figure 6. Distribution of four synthetic datasets: (a) imbalance, (b) aggregation, (c) banana, and (d) ring.

where n is the number of samples, n_{ij} is the same number of samples in c_i^α and c_j^β, $n_{i\alpha}$ is the number of samples in c_i^α, and $n_{\beta j}$ is the number of samples in c_j^β.

5.3. Datasets

Experimental evaluations were performed on nine datasets. More information describing the datasets used in the experiments can be found in Table 2. The cluster distribution of this synthetic 2D dataset is shown in Figure 6. The real datasets are derived from the UCI machine learning repository (Golrou et al. 2018; Movahhed Neya, Saberi, and Rezaie 2022).

5.4. Compared Methods

The proposed hybrid clustering method is evaluated in comparison with a wide range of clustering methods. Most of the clustering methods used for comparison are state-of-the-art and hybrid clustering methods. These methods include CO-average as a dual similarity approach that performs clustering through shared similarity matrix (Fred and Jain 2005). Similarity matrices based on CSM, WTQ, and WCT also belong to dual similarity approaches and are considered for comparison (Iam-On et al. 2011). Here, CO, CSM, WTQ, and WCT are analyzed through Single-Link (SL) and Average-Link (AL) hierarchical clustering methods to calculate the final results.

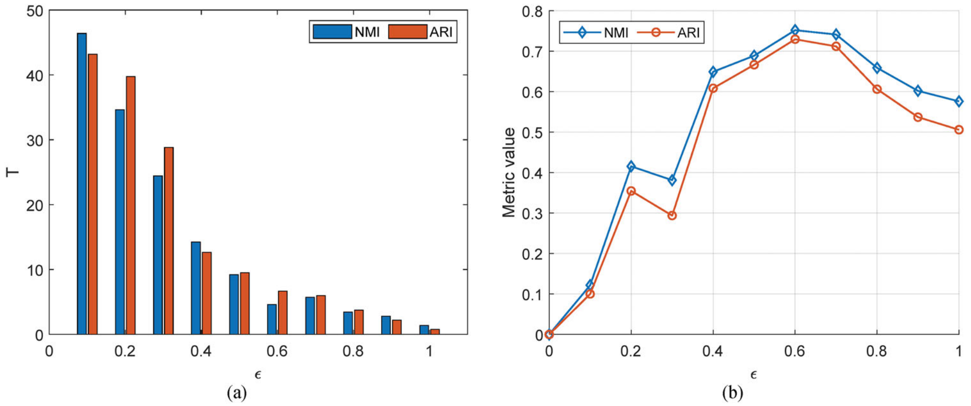


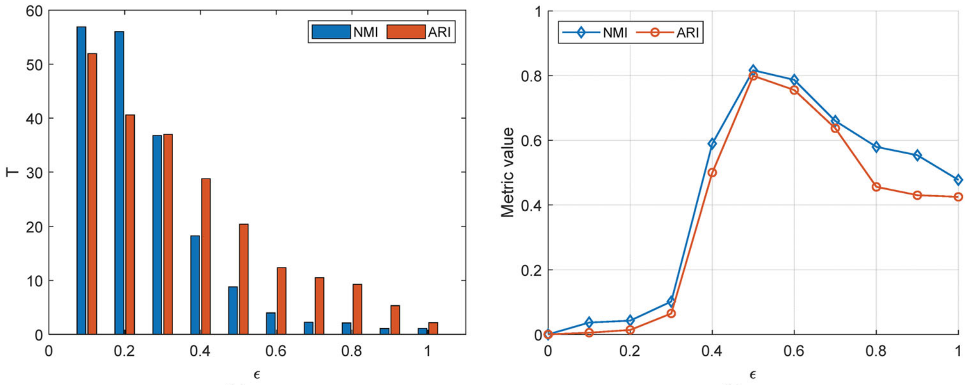
Figure 7. Analysis of the ε parameter of the proposed method on the Wine dataset. (a) number of basic clusters generated and (b) quality of clustering results.

Also, HGPA, MCLA, and CSPA are hybrid clustering methods presented by Strehl and Ghosh (2002). These methods are also considered for comparison and evaluation of the proposed method. In addition, we use SV and SWV as analysis-based weighted clustering methods for comparison work (Zhou and Tang 2006). Here, two feature-based clustering methods including IVC and EM are also used for comparison. IVC is presented by Nguyen and Caruana (2007) and EM by Topchy, Jain, and Punch (2005).

We also evaluated the proposed method in comparison with some Strong clustering approaches. Here, DBCAN, NSC, and CFSFDP were used for comparison. NSC is presented by Ng, Jordan, and Weiss (2001), DBCAN by Ester et al. (1996), and CFSFDP by Rodriguez and Laio (2014).

5.5. Results and Discussions

This section analyzes the results of the proposed method in comparison with existing clustering methods. First, the parameter ε is analyzed as an effective input parameter for the proposed method. In general, setting the parameter ε is an important challenge in the proposed method. We discussed that the selection of this parameter depends on the number of basic clustering considered by the users. Then, we examined the effect of the parameter ε on the performance of proposed method with performing relevant tests. For example, this problem has been analyzed on Wine and Iris datasets. As shown in part (a) in Figures 7 and 8, the number of basic clusters generated by the MKM scheme decreases with increasing ε. However, as shown in part (b) in these figures, the quality of the clustering results does not increase, hence the value of ε should be slightly increased. On the other hand, the results clearly show that the number of basic clustering methods considered is not suitable. In other words, the number of methods considered to produce high-quality final



(a) (b)

Figure 8. Analysis of the ε parameter of the proposed method on the Iris dataset. (a) number of basic clusters generated and (b) quality of clustering results.

Table 3. Results of evaluations on synthetic datasets based on NMI metric.

	Imbalance		Aggregation		Banana		Ring		Average	SD
Methods	Random	MKM	Random	MKM	Random	MKM	Random	MKM	MKM	MKM
CO-SL	0.269	0.007	0.875	0.631	0.402	0.001	0.693	0.693	0.333	0.238
CO-AL	0.269	0.231	0.875	0.841	0.402	0.415	0.210	0.194	0.420	0.261
CSM-SL	0.269	0.314	0.872	0.765	0.402	0.412	0.220	0.189	0.420	0.258
CSM-AL	0.269	0.001	0.872	0.024	0.402	0.002	0.020	0.001	0.007	0.310
WTQ-SL	0.269	0.009	0.875	0.550	0.402	0.002	0.539	0.693	0.314	0.225
WTQ-AL	0.269	0.205	0.869	0.887	0.402	0.432	0.216	0.237	0.440	0.257
WCT-SL	0.269	0.004	0.875	0.631	0.402	0.002	0.119	0.002	0.160	0.283
WCT-AL	0.269	0.298	0.875	0.863	0.402	0.395	0.215	0.118	0.418	0.260
MCLA	0.269	0.238	0.838	0.678	0.402	0.384	0.001	0.006	0.326	0.303
HGPA	0.004	0.007	0.628	0.678	0.001	0.007	0.001	0.129	0.205	0.271
CSPA	0.160	0.160	0.736	0.678	0.391	0.438	0.377	0.122	0.350	0.206
SV	0.269	0.001	0.386	0.304	0.402	0.230	0.174	0.163	0.174	0.093
SWV	0.269	0.002	0.617	0.435	0.402	0.003	0.247	0.148	0.147	0.147
IVC	0.269	0.223	0.893	0.810	0.402	0.395	0.380	0.170	0.399	0.240
EM	0.269	0.320	0.874	0.828	0.003	0.003	0.148	0.148	0.325	0.331
Proposed method	0.817	0.997	0.981	0.981	0.693	0.999	0.763	0.995	0.993	0.106

clusters is high or low. Therefore, we must select an appropriate value of ε to control the number of basic clustering on each dataset.

In the following, the proposed method is evaluated in comparison with other hybrid clustering methods. Based on the NMI and ARI credit criteria, the performance of different clustering methods has been compared on synthetic and real datasets. Table 3 shows the results of the comparisons for the synthetic dataset based on NMI, and the results of this measure for the real dataset are reported in Table 4. These comparisons for ARI are presented in Tables 5 and 6, respectively. Here, the last two columns indicate the average and Standard Deviation (SD) of each method for this dataset based on MKM. As illustrated, the superiority of the proposed method in creating high-quality and high-accuracy clusters on synthetic data sets is clear. This issue is confirmed by observing the results of the subject clustering methods. The proposed method has succeeded in creating higher

Table 4. Results of evaluations on real datasets based on NMI metric.

	Wine		Iris		Digits		WBCD			Average	SD
Methods	Random	MKM	Random	MKM	Random	MKM	Random	MKM	Random	MKM	Random
CO-SL	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760
CO-AL	0.725	0.725	0.725	0.725	0.725	0.725	0.725	0.725	0.725	0.725	0.725
CSM-SL	0.836	0.464	0.760	0.670	0.732	0.474	0.625	0.007	0.836	0.464	0.760
CSM-AL	0.836	0.023	0.760	0.027	0.005	0.005	0.625	0.007	0.836	0.023	0.760
WTQ-SL	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760
WTQ-AL	0.725	0.725	0.725	0.725	0.725	0.725	0.725	0.725	0.725	0.725	0.725
WCT-SL	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760
WCT-AL	0.725	0.725	0.725	0.725	0.725	0.725	0.725	0.725	0.725	0.725	0.725
MCLA	0.836	0.548	0.760	0.772	0.764	0.605	0.625	0.473	0.836	0.548	0.760
HGPA	0.172	0.374	0.162	0.725	0.495	0.005	0.002	0.002	0.172	0.374	0.162
CSPA	0.779	0.582	0.682	0.725	0.787	0.594	0.300	0.444	0.779	0.582	0.682
SV	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.020
SWV	0.243	0.243	0.243	0.243	0.243	0.243	0.243	0.243	0.243	0.243	0.243
IVC	0.730	0.585	0.682	0.711	0.722	0.688	0.043	0.336	0.730	0.585	0.682
EM	0.799	0.558	0.674	0.731	0.729	0.717	0.541	0.532	0.799	0.558	0.674
Proposed method	0.888	0.856	0.816	0.806	0.903	0.861	0.614	0.668	0.888	0.856	0.816

Table 5. Results of evaluations on synthetic datasets based on ARI metric.

	Imbalance		Aggregation		Banana		Ring		Average	SD
Methods	Random	MKM	Random	MKM	Random	MKM	Random	MKM	MKM	MKM
CO-SL	0.247	0.001	0.795	0.419	0.505	0.001	0.502	0.502	0.231	0.194
CO-AL	0.247	0.181	0.795	0.752	0.505	0.511	0.132	0.110	0.388	0.255
CSM-SL	0.247	0.323	0.794	0.710	0.505	0.507	0.146	0.104	0.411	0.251
CSM-AL	0.247	0.001	0.794	0.008	0.505	0.001	0.006	0.002	0.003	0.293
WTQ-SL	0.247	0.001	0.795	0.379	0.505	0.001	0.413	0.502	0.221	0.199
WTQ-AL	0.247	0.134	0.790	0.784	0.505	0.527	0.140	0.168	0.403	0.251
WCT-SL	0.247	−0.013	0.795	0.419	0.505	0.001	0.027	0.002	0.102	0.287
WCT-AL	0.247	0.297	0.795	0.766	0.505	0.491	0.140	0.020	0.394	0.253
MCLA	0.247	0.193	0.704	0.485	0.505	0.481	0.002	0.007	0.291	0.265
HGPA	0.001	0.001	0.493	0.485	0.001	0.001	0.002	0.122	0.152	0.213
CSPA	0.051	0.051	0.551	0.485	0.494	0.546	0.318	0.086	0.292	0.195
SV	0.247	−0.004	0.310	0.231	0.505	0.117	0.086	0.063	0.102	0.150
SWV	0.247	0.001	0.372	0.300	0.505	0.005	0.182	0.032	0.084	0.124
IVC	0.247	0.167	0.821	0.775	0.505	0.491	0.325	0.119	0.388	0.221
EM	0.247	0.333	0.826	0.743	0.005	0.005	0.032	0.032	0.278	0.330
Proposed method	0.423	1.000	0.906	0.988	0.670	1.000	0.508	0.993	0.996	0.184

Table 6. Results of evaluations on real datasets based on ARI metric.

	Wine		Iris		Digits		WBCD			Average	SD
Methods	Random	MKM	Random	MKM	Random	MKM	Random	MKM	Random	MKM	Random
CO-SL	0.732	0.732	0.729	0.732	0.732	0.732	0.732	0.732		0.732	0.001
CO-AL	0.567	0.567	0.564	0.567	0.567	0.567	0.567	0.567		0.567	0.001
CSM-SL	0.849	0.356	0.729	0.655	0.616	0.298	0.732	0.004		0.328	0.082
CSM-AL	0.849	0.002	0.729	0.002	0.001	0.001	0.732	0.004		0.002	0.336
WTQ-SL	0.732	0.732	0.729	0.732	0.732	0.732	0.732	0.732		0.732	0.001
WTQ-AL	0.567	0.567	0.564	0.567	0.567	0.567	0.567	0.567		0.567	0.001
WCT-SL	0.732	0.732	0.729	0.732	0.732	0.732	0.732	0.732		0.732	0.001
WCT-AL	0.567	0.567	0.564	0.567	0.567	0.567	0.567	0.567		0.567	0.001
MCLA	0.732	0.732	0.729	0.732	0.732	0.732	0.732	0.732		0.732	0.001
HGPA	0.733	0.733	0.730	0.733	0.733	0.733	0.733	0.733		0.733	0.001
CSPA	0.849	0.849	0.846	0.849	0.849	0.849	0.849	0.849		0.849	0.001
SV	0.008	0.008	0.005	0.008	0.008	0.008	0.008	0.008		0.008	0.001
SWV	0.164	0.164	0.161	0.164	0.164	0.164	0.164	0.164		0.164	0.001
IVC	0.689	0.539	0.596	0.689	0.602	0.600	0.050	0.354		0.545	0.253
EM	0.787	0.493	0.599	0.698	0.622	0.620	0.634	0.599		0.603	0.074
Proposed method	0.913	0.870	0.773	0.758	0.844	0.786	0.680	0.771		0.796	0.086

quality clusters due to the use of valid local clustering theory as well as the use of meta-clusters. Therefore, the proposed method identifies the final clusters more effectively and increases the efficiency.

Basically, the performance of the proposed method is better than other methods in the real dataset. However, improving the accuracy of the proposed method in the real dataset is less than that of the synthetic dataset. One of the most important reasons for this is that the dimensions of real data sets are much larger than synthetic datasets. In addition, according to the results, it can be stated that most of the compared clustering methods have performed better than the MKM scheme considering the random scheme. Because any basic cluster generated by the MKM scheme only includes the local modification of clusters in a dataset. However, these existing methods do not recognize or consider local modification. Therefore, they cannot get the results of a good group in the MKM scheme. The proposed method has better performance in MKM scheme than other methods. Note that the proposed method implements only in the MKM scheme, because the MKM scheme is part of it. We observe that the proposed method in the MKM scheme works better in terms of NMI and ARI based on other methods in the randomized scheme.

In the following, the proposed method is evaluated in comparison with other strong clustering methods. The results of comparing the proposed method with three strong clustering methods (i.e., NSC, DBCAN, and CFSFDP) based on the synthetic and real datasets are reported in Tables 7 and 8, respectively. Here, the last two rows refer to the mean and SD in each clustering method. As shown in these experiments, the clustering quality provided by the proposed method is better or promising compared to other methods. As the simulation results show, the proposed method can simulate strong simulation results and realize “a few clusters equal to a strong cluster.”

In another experiment, the computational complexity of clustering methods is evaluated through runtime analysis. The efficiency of the proposed method on the KDD-CUP99 dataset was tested. We set $k = 2$ and $\varepsilon = 0.14$. The runtime of the method with a number of samples (i.e., N) is shown in Table 9. It is clearly evident that the number of T basic clustering

Table 7. Results of evaluations on synthetic datasets.

	NSC		DBSCAN		CFSFDP		Proposed method	
Synthetic datasets	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI
Imbalance	0.9989	0.9992	0.9987	0.9993	0.9990	0.9988	0.9995	0.9989
Aggregation	0.9931	0.9919	0.8076	0.8887	0.9866	0.9811	0.9860	0.9809
Banana	0.9989	0.9992	0.9987	0.9993	0.9990	0.9988	0.9995	0.9989
Ring	0.9989	0.9992	0.9987	0.9993	0.3217	0.3780	0.9995	0.9989
Average	0.9975	0.9974	0.9509	0.9717	0.8266	0.8392	0.9961	0.9944
SD	0.0025	0.0032	0.0827	0.0479	0.2915	0.2664	0.0058	0.0078

Table 8. Results of evaluations on real datasets.

	NSC		DBSCAN		CFSFDP		Proposed method	
Synthetic datasets	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI
Wine	0.9297	0.9003	0.3574	0.4438	0.7401	0.7515	0.8674	0.8529
Iris	0.7442	0.7967	0.5149	0.5891	0.7015	0.7264	0.7552	0.8029
Digits	0.7523	0.8106	0.5039	0.7150	0.7571	0.8632	0.7828	0.8580
WBCD	0.7480	0.6315	0.0465	0.0290	0.7292	0.6139	0.7687	0.6654
KDD-CUP99	0.6346	0.6276	0.2843	0.3551	0.5853	0.5907	0.6346	0.6356
Average	0.7618	0.7533	0.3414	0.4264	0.7026	0.7091	0.7617	0.7630
SD	0.0948	0.1072	0.1715	0.2337	0.0614	0.0989	0.0747	0.0943

Table 9. Performance of the proposed method based on runtime (s) on the KDD-CUP99 dataset.

Number of samples	T	CSPA	MCLA	SWV	IVC	CFSFDP	Proposed method
5,000	44	11.82	6.24	10.85	12.39	18.81	5.96
10,000	107	41.18	39.87	39.74	39.70	53.26	26.14
15,000	110	67.98	62.21	54.08	51.37	63.20	39.54
20,000	114	99.11	88.15	70.79	65.01	74.91	55.11
25,000	114	123.91	108.81	83.65	75.26	83.01	67.51
30,000	119	156.64	136.09	101.38	89.81	95.75	83.87
35,000	119	183.18	158.21	115.20	100.87	104.60	97.14
40,000	179	316.56	269.36	192.11	166.36	168.89	163.83
45,000	232	505.95	427.18	297.38	254.11	249.69	258.53
50,000	234	562.69	474.46	327.19	278.09	269.27	286.92

increases with increasing number of samples. Given the time complexity of the proposed method, the runtime with T is second order. Given the runtime of the proposed method, the runtime with T is second order. However, since that $T < N$ and T slowly increase in comparison with N growth, the cost of increasing T growth time is acceptable. As depicted, the cost of runtime of the method proposed is proportional to the number of linear samples. Therefore, the proposed method must be able to quickly obtain the final clustering in a large-scale dataset. As illustrated, the proposed method is very efficient.

6. Conclusion

Among the clustering methods, hybrid clustering is one of the popular methods with high stability and robustness, which provides the ability to discover hidden patterns with high accuracy. Hybrid clustering can adapt itself to the input dataset by using the knowledge of different methods and increase the quality of the final solution. The different quality of partitions from basic clustering methods is one of the arguments of hybrid clustering, which can achieve better results by combining them. Although k-means is a poor clustering method, it has a low computational cost, which makes it unsuitable for clustering results. Therefore, this study used the k-means clustering method as the basic cluster. Here, we presented a different definition of valid local clusters by considering the data around the cluster centers in k-means. To increase the diversity in the primary clusters, we used a duplicate strategy on nonappearance data in valid local clusters. Also, we used the inter-cluster and intra-cluster similarity measure to estimate the similarity between valid local clusters. This process has resulted in the production of a weighted graph in which the weight of the edges expresses the degree of similarity between the clusters. An aggregation function based on meta-clustering was used to create the final clusters, in which the primary clusters were re-clustered to obtain the final clusters. In general, the idea of the proposed method is to understand the concept of several weak clusters equal to a strong cluster by k-means. The results obtained from the proposed hybrid clustering method are more consistent with the real data structure. This method has reported better results compared to state-of-the-art methods on different datasets. Based on the results, the proposed method is effective for dealing with large-scale datasets. According to the concept of granular computing, how to extract the relationship between basic clustering methods and primary partitions is worth studying in future work. Also, the proposed method can appear more effective considering feature extraction/selection approaches. On the other hand, it is recommended to use techniques of increasing diversity such as bagging to select suitable basic clustering methods for future works.

Data availability

Data sharing not applicable to this manuscript as no datasets were generated or analyzed during the current study.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Training plan for young backbone teachers in Henan Province (No.2018GGJS267).

References

Abapour, N., A. Shafiesabet, and R. Mahboub. 2021. A novel security based routing method using ant colony optimization algorithms and RPL protocol in the IoT networks. *International Journal of Electrical and Computer Sciences* 3 (1):1–9.

Azimi, J., and X. Fern. 2009. Adaptive cluster ensemble selection. In *Twenty-First International Joint Conference on Artificial Intelligence*, Vol. 9, 992–7, California, USA, July 11–17.

Bai, L., J. Liang, and F. Cao. 2020. A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters. *Information Fusion* 61:36–47. doi:10.1016/j.inffus.2020.03.009.

Berahmand, K., E. Nasiri, R. Pir Mohammadiani, and Y. Li. 2021. Spectral clustering on protein-protein interaction networks via constructing affinity matrix using attributed graph embedding. *Computers in Biology and Medicine* 138:104933.

Bouyer, A, and A. Hatamlou. 2018. An efficient hybrid clustering method based on improved cuckoo optimization and modified particle swarm optimization algorithms. *Applied Soft Computing* 67:172–82. doi:10.1016/j.asoc.2018.03.011.

Ester, M., H. P. Kriegel, J. Sander, and X. Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD-96, 226–31, Portland, Oregon, USA, August 2–4.

Forouzandeh, S., K. Berahmand, E. Nasiri, and M. Rostami. 2021. A hotel recommender system for tourists using the Artificial Bee Colony Algorithm and Fuzzy TOPSIS Model: A case study of tripadvisor. *International Journal of Information Technology & Decision Making* 20 (1):399–429. doi:10.1142/S0219622020500522.

Fred, A. L., and A. K. Jain. 2005. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (6):835–50.

Ghobaei-Arani, M. 2021. A workload clustering based resource provisioning mechanism using Biogeography based optimization technique in the cloud based systems. *Soft Computing* 25 (5):3813–30. doi:10.1007/s00500-020-05409-2.

Ghobaei-Arani, M., and A. Shahidinejad. 2021. An efficient resource provisioning approach for analyzing cloud workloads: a metaheuristic-based clustering approach. *The Journal of Supercomputing* 77 (1):711–50. doi:10.1007/s11227-020-03296-w.

Golalipour, K., E. Akbari, S. S. Hamidi, M. Lee, and R. Enayatifar. 2021. From clustering to clustering ensemble selection: A review. *Engineering Applications of Artificial Intelligence* 104:104388. doi:10.1016/j.engappai.2021.104388.

Golrou, A., A. Sheikhani, A. M. Nasrabadi, and M. R. Saebipour. 2018. Enhancement of sleep quality and stability using acoustic stimulation during slow wave sleep. *International Clinical Neuroscience Journal* 5 (4):126–34. doi:10.15171/icnj.2018.25.

Hamidi, S. S., E. Akbari, and H. Motameni. 2019. Consensus clustering algorithm based on the automatic partitioning similarity graph. *Data & Knowledge Engineering* 124:101754. doi:10.1016/j.datak.2019.101754.

Hansen, P., and N. Mladenović. 2001. J-means: A new local search heuristic for minimum sum of squares clustering. *Pattern Recognition* 34 (2):405–13. doi:10.1016/S0031-3203(99)00216-2.

Huang, D., C. D. Wang, and J. H. Lai. 2017. LWMC: A locally weighted meta-clustering algorithm for ensemble clustering. In *International Conference on Neural Information Processing*, 167–76. Cham: Springer.

Huang, D., C. D. Wang, J. S. Wu, J. H. Lai, and C. K. Kwoh. 2020. Ultra-scalable spectral clustering and ensemble clustering. *IEEE Transactions on Knowledge and Data Engineering* 32 (6):1212–26. doi:10.1109/TKDE.2019.2903410.

Iam-On, N., T. Boongoen, S. Garrett, and C. Price. 2011. A link-based approach to the cluster ensemble problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33 (12):2396–409. doi:10.1109/TPAMI.2011.84.

Jadidi, A., and M. R. Dizadji. 2021. Node clustering in binary asymmetric stochastic block model with noisy label attributes via SDP. In *2021 International Conference on Smart Applications, Communications and Networking (SmartNets)*, 1–6. New York: IEEE. doi:10.1109/SmartNets50376.2021.9555421.

Jain, A. K. 2010. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters* 31 (8):651–66. doi:10.1016/j.patrec.2009.09.011.

Jiang, H., S. Yi, J. Li, F. Yang, and X. Hu. 2010. Ant clustering algorithm with K-harmonic means clustering. *Expert Systems with Applications* 37 (12):8679–84. doi:10.1016/j.eswa.2010.06.061.

Khedairia, S., and M. T. Khadir. 2022. A multiple clustering combination approach based on iterative voting process. *Journal of King Saud University – Computer and Information Sciences* 34 (1):1370–80. doi:10.1016/j.jksuci.2019.09.013.

Li, F., Y. Qian, and J. Wang. 2021. GoT: A growing tree model for clustering ensemble. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 8349–56, California, USA, February 2–9.

Li, T., A. Rezaeipanah, and E. M. T. El Din. 2022. An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement. *Journal of King Saud University – Computer and Information Sciences* 34 (6):3828–42. doi:10.1016/j.jksuci.2022.04.010.

Ma, T., Z. Zhang, L. Guo, X. Wang, Y. Qian, and N. Al-Nabhan. 2021. Semi-supervised Selective Clustering Ensemble based on constraint information. *Neurocomputing* 462: 412–25. doi:10.1016/j.neucom.2021.07.056.

Mojarad, M., F. Sarhangnia, A. Rezaeipanah, H. Parvin, and S. Nejatian. 2021. Modeling hereditary disease behavior using an innovative similarity criterion and ensemble clustering. *Current Bioinformatics* 16 (5):749–64. doi:10.2174/1574893616999210128175715.

Movahhed Neya, N., S. Saberi, and B. Rezaie. 2022. Design of an adaptive controller to capture maximum power from a variable speed wind turbine system without any prior knowledge of system parameters. *Transactions of the Institute of Measurement and Control* 44 (3):609–19. doi:10.1177/01423312211039041.

Nasiri, E., K. Berahmand, Z. Samei, and Y. Li. 2022. Impact of centrality measures on the common neighbors in link prediction for multiplex networks. *Big Data* 10 (2):138–50.

Ng, A., M. Jordan, and Y. Weiss. 2001. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems* 14:849–56.

Nguyen, N., and R. Caruana. 2007. Consensus clusterings. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 607–12. New York: IEEE. doi:10.1109/ICDM.2007.73.

Niu, H., N. Khozouie, H. Parvin, H. Alinejad-Rokny, A. Beheshti, and M. R. Mahmoudi. 2020. An ensemble of locally reliable cluster solutions. *Applied Sciences* 10 (5):1891. doi:10.3390/app10051891.

Rezaeipanah, A., P. Amiri, H. Nazari, M. Mojarad, and H. Parvin. 2021. An energy-aware hybrid approach for wireless sensor networks using re-clustering-based multi-hop routing. *Wireless Personal Communications* 120 (4):3293–314. doi:10.1007/s11277-021-08614-w.

Rezaeipanah, A., H. Nazari, and G. Ahmadi. 2019. A hybrid approach for prolonging lifetime of wireless sensor networks using genetic algorithm and online clustering. *Journal of Computing Science and Engineering* 13 (4):163–74. doi:10.5626/JCSE.2019.13.4.163.

Rodriguez, A., and A. Laio. 2014. Clustering by fast search and find of density peaks. *Science (New York, N.Y.)* 344 (6191):1492–6. doi:10.1126/science.1242072.

Shahidinejad, A., M. Ghobaei-Arani, and L. Esmaeili. 2020. An elastic controller using Colored Petri Nets in cloud computing environment. *Cluster Computing* 23 (2):1045–71. doi:10.1007/s10586-019-02972-8.

Strehl, A., and J. Ghosh. 2002. Cluster ensembles – A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research* 3:583–617.

Sun, S., S. Wang, G. Zhang, and J. Zheng. 2018. A decomposition-clustering-ensemble learning approach for solar radiation forecasting. *Solar Energy* 163:189–99. doi:10.1016/j.solener.2018.02.006.

Tan, H., Y. Tian, L. Wang, and G. Lin. 2020. Name disambiguation using meta clusters and clustering ensemble. *Journal of Intelligent & Fuzzy Systems* 38 (2):1559–68. doi:10.3233/JIFS-179519.

Topchy, A., A. K. Jain, and W. Punch. 2005. Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (12):1866–81. doi:10.1109/TPAMI.2005.237.

Trik, M., A. M. N. G. Molk, F. Ghasemi, and P. Pouryeganeh. 2022. A hybrid selection strategy based on traffic analysis for improving performance in networks on chip. *Journal of Sensors* 2022:1–19. doi:10.1155/2022/3112170.

Trik, M., S. Pour Mozaffari, and A. M. Bidgoli. 2021. Providing an adaptive routing along with a hybrid selection strategy to increase efficiency in NoC-based neuromorphic systems. *Computational Intelligence and Neuroscience* 2021:8338903. doi:10.1155/2021/8338903.

Walid, W., M. Awais, A. Ahmed, G. Masera, and M. Martina. 2021. Real-time implementation of fast discriminative scale space tracking algorithm. *Journal of Real-Time Image Processing* 18 (6):2347–60. doi:10.1007/s11554-021-01119-6.

Wei, S., Z. Li, and C. Zhang. 2018. Combined constraint-based with metric-based in semi-supervised clustering ensemble. *International Journal of Machine Learning and Cybernetics* 9 (7):1085–100. doi:10.1007/s13042-016-0628-6.

Wei, Y., S. Sun, J. Ma, S. Wang, and K. K. Lai. 2019. A decomposition clustering ensemble learning approach for forecasting foreign exchange rates. *Journal of Management Science and Engineering* 4 (1):45–54.

Yang, W., Y. Zhang, H. Wang, P. Deng, and T. Li. 2021. Hybrid genetic model for clustering ensemble. *Knowledge-Based Systems* 231:107457. doi:10.1016/j.knosys.2021.107457.

Zhang, B., M. Hsu, and U. Dayal. 2000. K-harmonic means-a spatial clustering algorithm with boosting. In International Workshop on Temporal, Spatial, and Spatio-Temporal Data Mining, 31–45. Berlin, Heidelberg: Springer.

Zhao, Q., Y. Zhu, D. Wan, Y. Yu, and Y. Lu. 2019. Similarity analysis of small-and medium-sized watersheds based on clustering ensemble model. *Water* 12 (1):69. doi:10.3390/w12010069.

Zheng, Y., Z. Long, C. Wei, and H. Wang. 2021. Particle swarm optimization for clustering ensemble. In *2021 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, 385–91. New York: IEEE. doi:10.1109/ISKE54062.2021.9755338.

Zhou, Z. H., and W. Tang. 2006. Clusterer ensemble. *Knowledge-Based Systems* 19 (1):77–83. doi:10.1016/j.knosys.2005.11.003.

Zhu, X., B. Fei, D. Liu, and W. Bao. 2021. Adaptive clustering ensemble method based on uncertain entropy decision-making. In *2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 61–7. New York: IEEE. doi:10.1109/TrustCom53373.2021.00026.