IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE RELIABILITY SOCIETY SECTION

Received January 3, 2021, accepted January 17, 2021, date of publication January 21, 2021, date of current version February 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3053244

A Clustering Analysis Method With High Reliability Based on Wilcoxon-Mann-Whitney Testing

YUAN CHENG[1], WEINAN JIA[1], RONGHUA CHI[2], AND AO LI[1], (Member, IEEE)
[1]College of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China
[2]School of Computer and Information Engineering, Heilongjiang University of Science and Technology, Harbin 150022, China

Corresponding author: Yuan Cheng (changuang7@sina.com)

This work was supported in part by the Natural Science Foundation of Heilongjiang Province under Grant F2017015, in part by the Training Program For Young Innovators in Heilongjiang General Institutes of Higher Education under Grant UNPYSCT-2017079, and in part by the National Natural Science Foundation of China under Grant 62071157.

ABSTRACT As a core step in clustering analysis, distance measurement results can influence clustering accuracy. Existing measurement methods are mostly based on cluster feature information. However, these cluster features may be insufficient and result in losing data information for clusters containing a number of objects. To improve measurement accuracy, we make full use of the distribution characteristics of objects in clusters, i.e., we use descriptive statistics and the Wilcoxon-Mann-Whitney rank sum test in nonparametric statistics to measure distances during clustering. Furthermore, we propose a two-stage clustering algorithm to improve clustering analysis performance. In terms of avoiding preliminarily assuming the number of clusters, with the proposed distance measurement method, the clustering algorithm can discover clusters with arbitrary shapes and improve clustering accuracy. Experiments with multiple datasets compared with other clustering algorithms illustrate the accuracy and efficiency of the proposed clustering algorithm.

INDEX TERMS Clustering analysis, distance measurement, nonparametric statistics, Wilcoxon-Mann-Whitney rank sum test.

I. INTRODUCTION

As a basic data mining strategy, clustering analysis is significant for discovering the characteristics of data aggregation, which is an unsupervised process [1]–[3]. When the data distribution is unknown, the clustering method is effective at obtaining the inherent distribution of data [4]–[6]. To ensure the reliability of acquiring data aggregation features, it is necessary to ensure the reliability of clustering. Clustering reliability is reflected in the method finding clusters of any shape and that the number of clusters generated is not limited by input parameters. The accuracy of the clustering results also affect the reliability of the analysis of the inherent distribution of the data, that is, the more accurate the clustering results are, the higher the reliability of the data aggregation feature discovery is [7]. Although there are different ways to obtain data groups, such as the partitioning clustering method, hierarchical clustering method, density-based clustering method, grid-based method and so on, the implicit concepts of these clustering methods are similar: they are based on the distances between objects through multiple iterations to ensure the clustering quality. K-means [8], [9] adjusts the clusters and their mean values based on the distances between objects and clusters mean values in each iteration. DBSCAN [10], [11] aggregates the objects that are directly density-reachable from the core object in each iteration to generate new clusters. Therefore, the primary difference is how to divide objects into clusters during the clustering process. In this unsupervised analysis, the main basis of assigning an object to a cluster is the distance measurement, including distance between objects, distance between the object and the cluster, and distance between clusters. K-means divides objects into clusters based on the distances between objects and clusters. The judgment of directly density-reachable objects in DBSCAN is also based on the distances between objects. In addition, cluster merging is determined by the distances between clusters in agglomerative hierarchical clustering. It is obvious that the accuracy of distance measurement is of great

The associate editor coordinating the review of this manuscript and approving it for publication was Zhaojun Li.

importance in effective clustering. That is, the accuracy of distance measurement is the key to impacting the reliability of obtaining data aggregation features.

The existing distance measurement between objects can be divided into multiple methods according to the attribute types and the application scenes, such as Euclidean distance, Manhattan distance, Minkowski distance, Jaccard coefficient, cosine measure and so on [12]–[16]. In addition, the distances involving clusters are mostly measured based on the information reflecting cluster features. For instance, K-means and K-medoids [17], [18] choose the mean value or a representative object as the feature of a cluster, and they assign an object to a cluster whose feature is closest to it. DBSCAN and OPTIS [19] consider the core object as the cluster feature and decide whether to assign an object into a cluster according to whether it is density-reachable to the core object [20]–[24].

The objects in each cluster are the prime factors truly reflecting the cluster features, and thus the distance between clusters can be calculated through the distances between objects in the clusters, such as the minimum distance, the maximum distance, and the average distance. However, a large number of objects may affect the efficiency of distance measurement. Therefore, to improve the computation speed and scalability, Birch [25] used zero moments, first moments and second moments to generate a three-dimensional vector, which is represented as a cluster feature to summarize cluster information and to compute the distance between clusters for hierarchical clustering. However, it is not sufficient to simply describe the cluster information by using the representative objects or the statistics. The existing cluster features represent the aggregation features of clusters containing a number of objects, and there is a loss of information reflecting the data characteristics of clusters to a certain extent. Then, the distance measurement would contain a certain deviation and thus affect the accuracy of clustering results.

Obviously, the distance measurement is a core step in clustering. An effective information extraction that represents the data features of clusters is the key to ensuring measurement accuracy, so it is also important to ensure the accuracy of clustering. Therefore, researchers have attempted to extract effective information about cluster features. Reference [26] extracted specific adjacent objects of centroids to summarize clustering information. A group of representative objects was used, but the adjacent objects were insufficient to reflect the general data features. Reference [27] defined a core set to measure distances using the Birch concept. Although they chose a number of objects as representative cluster information, this was also insufficient and resulted in information loss. The distribution of data in clusters can reflect the general cluster data features. Reference [28] obtained the distribution features of clusters based on a probability density function. However, this method presupposed the data distribution. It is difficult to make a clear assumption about data distribution because of the sparse knowledge about the overall information. Incorrect assumptions can result in inaccurate distance measurements and clustering results.

Nonparametric statistical methods [29]–[33] can be used to estimate distribution structures based on direct data information rather than a hypothesis of the specific form of an overall distribution. The Wilcoxon-Mann-Whitney (W-M-W) rank sum test method [34]–[38] is a nonparametric statistical method used to judge whether any two sets come from the same population. In the clustering process, if two sets represented by two clusters are from the same population, they can be grouped into one cluster. Therefore, through this method, we can reserve the original cluster information features, analyze the dissimilarity between clusters directly based on the distribution features of their data, and then determine whether to merge them into one cluster without a hypothesis of the overall distribution form.

To resolve the above problems, we try to make full use of the characteristics of data, so we use W-M-W rank sum test to measure the differences between clusters, which could lay a foundation for obtaining more accurate clustering results. In addition, we also propose an improved hierarchical clustering method to increase clustering effectiveness. This method has minimal requirements for domain knowledge when determining input parameters. It could also help to discover clusters with arbitrary shapes and improve clustering accuracy. Experiments on multiple datasets are used to verify the validity of the proposed algorithm, which is a key to ensuring the reliability of obtaining data aggregation features. An experiment on a real dataset illustrates the practicability of the proposed method and further proves that this method can facilitate the reliability of obtaining the inherent distribution of data.

II. DISTANCE MEASUREMENT BASED ON NONPARAMETRIC STATISTICS

In data mining, especially in cluster analysis, distance measurement is the core step in data analysis. Its accuracy can directly affect the validity of data analysis results. There are multiple metric methods based on different data types, such as Euclidean distance, Manhattan distance, Minkowski distance, Jaccard coefficient, cosine measure and so on. In clustering analysis, objects are divided into clusters, or two similar sets are grouped into one cluster during the clustering processes. These operations are based on distance measurements, which include distances between objects and clusters as well as distances between clusters. Obviously, these distances involve clusters. To ensure the objectivity and accuracy of measurements, it is necessary to consider the distribution features of objects in the clusters with little loss of data information. To achieve this purpose, this paper will measure distances based on the distribution characteristics of data in clusters.

A. DISTANCE BETWEEN OBJECTS AND CLUSTERS BASED ON DISTRIBUTION CHARACTERISTICS

In traditional clustering methods, the distances between objects and clusters are often transformed into the distances between objects and cluster features. These features can

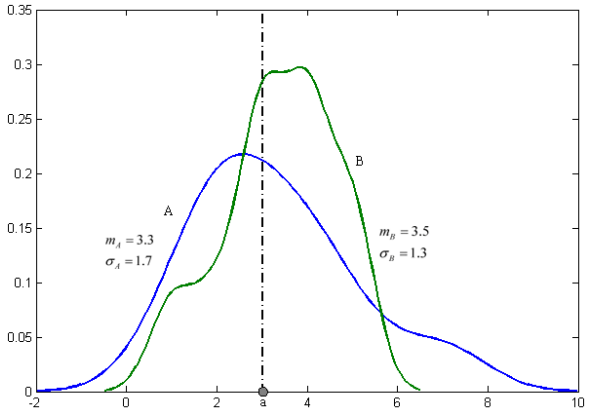


FIGURE 1. Comparing a one-dimensional object with two sets.

include cluster mean values, representative objects of clusters, core sets of clusters and so on. As mentioned above, these single descriptive features can lose the data information in the clusters to some extent. In addition, the distribution characteristics of objects in one cluster can be seen as its general cluster feature. The difference between clusters can refer to the difference between their object distribution characteristics. We thus assign an object into a cluster with a minimum distance from another object. Then, the distribution characteristics of objects in the clusters will have a greater impact on distance measurement and should be considered.

The distribution characteristics of objects in a cluster must be considered when measuring distances. We take the one-dimensional data in Fig. 1 as an example. There are two sets A and B in Fig. 1. We want to determine the closer set and assign object 'a' to that set. According to traditional clustering methods, we can compare either the difference between 'a' and the mean value of A and B or the difference between 'a' and the centroid of A and B. The former uses the mean value as the statistical feature of the set, while the latter selects a central point as the representative of the cluster. In Fig. 1, when the mean value is used to represent the set feature, the difference between the two sets is not significant.

An object a = 3 is more similar to set A. However, from the perspective of the distributions of these two sets, the dispersion degree of set B is lower, that is, its objects are more aggregated. In addition, object 'a' is more likely to belong to set B when considering distributions.

It is apparent that the mean value, as one statistical feature of a set, cannot completely reflect the distribution, while other representatives may not reflect the distribution characteristics of all objects in a cluster. Therefore, these measurement methods cannot objectively calculate the differences between objects and clusters. Thus, the use of one statistic is insufficient to represent the general characteristics of data in a set.

For a more accurate and objective result, we need to assign an object into a cluster while considering distribution characteristics of the data in the cluster. The methods that can describe the distribution features of a set include probability distribution functions and descriptive statistics. It is very time consuming to compute the probability distribution function of objects for each cluster, whereas multiple descriptive statistics can describe the statistical characteristics of a set from different perspectives; thus, statistics can be used to represent the distribution characteristics. Therefore, this paper will measure the distance between objects and clusters based on specific statistical features of a set.

If an object belongs to a cluster, it has similar characteristics to other objects in this cluster. That is, the distribution characteristics of this cluster will not change significantly after the object is assigned into it. We consider different descriptive statistics of a cluster when measuring distances between an object and a cluster and then analyze whether these statistics have changed significantly after the object is added to the cluster. We thus determine the right cluster with the smallest change of the statistics and below a threshold.

In descriptive statistical analysis, statistics such as the mean, variance, and quantile can be used to measure the average values, central tendency and location information of data in a set, respectively. These statistics describe the data information for position and dispersion of a set and actually represent the distribution characteristics of a data set.

We begin with one-dimensional data to discuss the method of determining the relationship between an object and a cluster with the above descriptive statistics. Then, we extend this method to multidimensional data. We match the object and descriptive statistical features of the cluster in each dimension and analyze the differences between the object and the cluster in an effective way.

Let o_1 be the one-dimensional object to be assigned, where the existing clusters are $C = \{C_1, C_2, \ldots, C_n\}$. The distribution feature of cluster C_i $(1 \leq i \leq n)$ can be described by a triple $DF_i = <\mu_i, \sigma_i, m_i>$, where μ_i, σ_i and m_i are the mean value, variance and median, which represent the average value, dispersion and the center position of cluster C_i, respectively. If the object is assigned to C_i, the distribution feature of C_i will be $DF_i' = <\mu_i', \sigma_i', m_i'>$. In addition, the variation of distribution feature can be calculated as Equation (1).

$$\Delta_i = \left|\mu_i' - \mu_i\right| + \left|\sigma_i' - \sigma_i\right| + \left|m_i' - m_i\right| (1 \leq i \leq n) \quad (1)$$

If o_1 belongs to cluster C_ω, its impact on the distribution feature of C_ω should be relatively small, i.e., the value of Δ_ω should be the smallest and within a certain threshold.

For instance, assume there are three clusters: $C_1 = \{4.7, 5.1, 4.8, 5.4, 5.5, 4.4, 5\}$, $C_2 = \{5.9, 5.2, 6, 5.5, 5.8, 6.1, 5.7\}$, and $C_3 = \{5.8, 6.3, 6.1, 7.1, 5.6, 6.7, 6.5\}$. The triples representing their distribution features are $DF_1 = <4.99, 0.39, 5>$, $DF_2 = <5.74, 0.29, 5.75>$ and $DF_3 = <6.3, 0.52, 6.3>$, respectively. The object to be added is $o_1 = 5.7$. The threshold of variation for the distribution feature is $\delta = 0.1$.

We can obtain the triples $DF_1' = <5.07, 0.44, 5.05>$, $DF_2' = <5.74, 0.29, 5.75>$, and $DF_3' = <6.22, 0.53, 6.2>$ if o_1 is assigned to C_1, C_2 and C_3, respectively. Their variations on the distribution feature are $\Delta_1 = 0.18$, $\Delta_2 = 0.07$

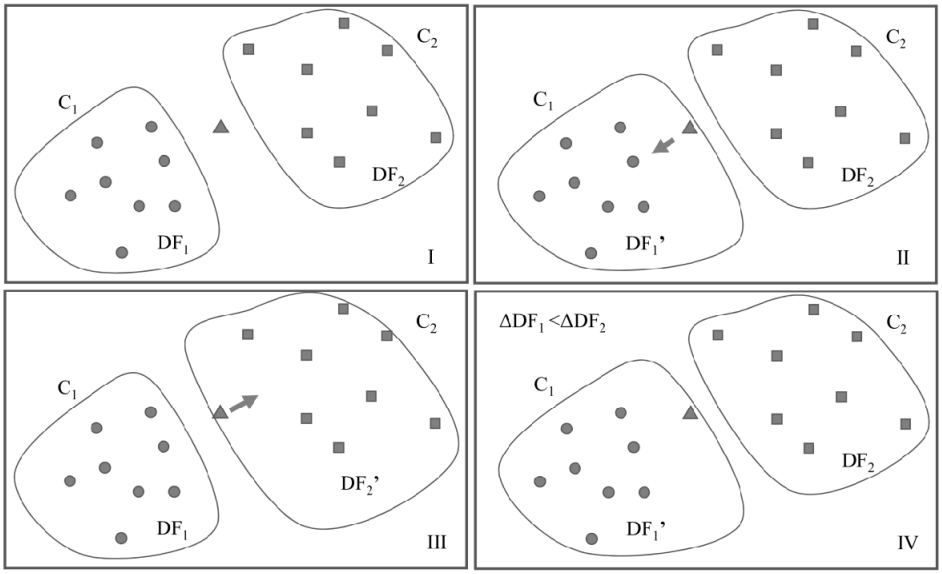


FIGURE 2. Allocating an object into a cluster based on distribution feature *DF*.

and $\Delta_3 = 0.19$, respectively, where Δ_2 has the smallest value and $\Delta_2 < \delta$. It can be concluded that o_1 is more likely to come from the same distribution with data in C_2. Then, o_1 can be assigned to C_2.

If we extend the above method to multidimensional data, we need to determine the relation between the object and the cluster distribution feature in each dimension as described above. Then, we can integrate the analysis results on each dimension to determine the cluster having the smallest variation in distribution feature after the object is added to it. This cluster is more similar to the object than the others.

Let o_2 be the d-dimensional object to be assigned, and $C = \{C_1, C_2, \ldots, C_n\}$ be the existing clusters. The distribution feature of cluster C_i $(1 \leq i \leq n)$ can be described by a d-dimensional triple as Equation (2).

$$\begin{aligned} DF_i &= \{DF_{i1}, \ldots, DF_{id}\} \\ &= \{< \mu_{i1}, \sigma_{i1}, m_{i1} >, \ldots, < \mu_{id}, \sigma_{id}, m_{id} >\} \end{aligned} \tag{2}$$

During the analysis, we can calculate the distribution feature in the k-th $(1 \leq k \leq d)$ dimension of each cluster: $DF'_{ik} = <\mu'_{ik}, \sigma'_{ik}, m'_{ik}>$ $(1 \leq i \leq n)$, when o_2 is assumed to be divided into each cluster. In addition, the variation for the distribution feature in the k-th dimension can also be calculated as Equation (3).

$$\Delta_{ik} = \left|\mu'_{ik} - \mu_{ik}\right| + \left|\sigma'_{ik} - \sigma_{ik}\right| + \left|m'_{ik} - m_{ik}\right| \tag{3}$$

Then, the variation in all dimensions is $\Delta_i = \sum d\, j = 1\Delta_{ij}$. Let C_ω be the cluster that o_2 is most likely to be assigned to. Its variation value Δ_ω should be the smallest and within a certain threshold.

Fig. 2 shows the process of allocating the cluster for a two-dimensional object, when there are two clusters as candidates. After comparing the variation of distribution feature *DF* for cluster C_1 and C_2 once we test that the object is allocated into these two clusters, we can assign the object to cluster C_1 because there is a smaller change of *DF*, and we consider that it is more likely to have the same distribution with objects in C_1.

The above method of assigning an object to the most similar cluster is described in the following **ocd()** algorithm. We can compute the variation of the distribution feature for each cluster with the assumption that an object is grouped into every cluster. The cluster having the minimum variation value and less than the threshold is the one most matching the object in statistical characteristics. If all the variation values are greater than the threshold, the object is more likely to be an outlier.

Then, we take data shown in Fig. 3 as an example to specify the method of assigning objects into clusters based on distribution features. The 4-dimensional object is $o_2 = (5.7, 4.4, 1.5, 0.4)$. There are three clusters: C_1, C_2 and C_3. Their distribution features, as represented by 4-dimensional triples, are $DF_1 = \{<4.99, 0.39, 5>, <3.29, 0.24, 3.3>, <1.46, 0.15, 1.4>, <0.27, 0.13, 0.2>\}$, $DF_2 = \{<5.74, 0.31, 5.8>, <2.83, 0.29, 2.9>, <4.3, 0.34,4.2>, <1.37, 0.25, 1.4>\}$, and $DF_3 = \{<6.3, 0.52, 6.3>, <3.04, 0.24,3>, <5.37, 0.48, 5.1>, <2.06, 0.22, 2>\}$, respectively. If the threshold of variation for distribution features $\delta = 0.8$, we can obtain the new triples: $DF'_1 = \{<5.07, 0.44, 5.05>, <3.425, 0.45, 3.35>, <1.46, 0.14, 1.45>, <0.29, 0.12, 0.2>\}$, $DF'_2 = \{<5.74, 0.29, 5.75>, <3.025, 0.62, 2.95>, <3.95,1.04,4.15>, <1.25,0.41,1.35>\}$, and $DF'_3 = \{<6.22, 0.53, 6.2>, <3.21,0.53, 3.1>, <4.89, 1.44, 5.1>, <1.85, 0.62, 2>\}$.

If we assume the object is assigned to these clusters, the variations with the former are $\Delta_1 = 0.665$, $\Delta_2 = 2.075$, and $\Delta_3 = 2.8$. Obviously, Δ_1 is the minimum and less than the threshold δ. Taking into account all four dimensions, the object is more likely to be from the same distribution as

Algorithm 1 ocd(o, C, δ)

Input: o, a d-dimensional object to be assigned;
$C = \{C_1, C_2, \ldots, C_k\}$, the existing cluster set;
δ, the threshold of variations about distribution features;
Output: C_ω, the cluster that o belongs to;
Steps:

1) Let $\Delta_m = \Delta_1$, $c = 0$;
2) for $i := 1$ to n:
3) calculate $DF_i = \{DF_{i1}, \ldots, DF_{id}\}$;
4) calculate DF_i' if o is assumed to be grouped into cluster C_i;
5) calculate $\Delta_i = \sum_{j=1}^{d} \Delta_{ij}$;
6) if ($\Delta_i < \Delta_m$ && $\Delta_i < \delta$) then:
7) $\Delta_m = \Delta_i$;
8) $c = i$;
9) $C_\omega = C_c$.

cluster C_1. Therefore, it can be assigned to C_1. This result is different from the above one-dimensional analysis, since object o_2 is described by the four dimensions, and its assignment is based on distribution features for all dimensions rather than one dimension.

B. DISTANCE BETWEEN CLUSTERS BASED ON RANK SUM TEST

The main purpose of measuring the distance between clusters is to merge similar clusters into one cluster. The similarity in unsupervised data analysis is based on distance measurement, while from a statistical perspective, the objects in two clusters that are similar are considered more likely to come from the same distribution. The W-M-W rank sum test method is a nonparametric statistics method. It can test whether two sets of samples are from the same population without requiring too many samples and a prehypothesis about data distribution.

This approach can provide an objective conclusion. Therefore, based on the W-M-W rank sum test method, we will determine whether to merge two clusters by testing whether the objects in the two clusters derive from the same population. If from the same population, they can be merged into one cluster; otherwise, the two clusters will exist as two separate clusters.

For any two clusters C_1 and C_2, their numbers of objects are n_{C1} and n_{C2}, respectively. The upper limit number of objects used in the rank sum test is n_δ. When $n_{C1}, n_{C2} \leq n_\delta$, all objects in these clusters can be used in the rank sum test to determine whether they are from the same distribution; otherwise, we will take n_δ samples randomly from the two clusters for the test.

Then, we take one-dimensional objects as examples to describe how to decide whether two sets need to be merged through the W-M-W rank sum test. If the objects contain multidimensional data, we need to analyze each dimension per the method. Objects in two clusters from the same population

C_1	C_2	C_3
(4.7, 3.2, 1.3, 0.2),	(4.7, 3.2, 1.3, 0.2),	(4.7, 3.2, 1.3, 0.2),
(5.1, 3.3, 1.7, 0.5),	(5.1, 3.3, 1.7, 0.5),	(5.1, 3.3, 1.7, 0.5),
(4.8, 3.1, 1.6, 0.2),	(4.8, 3.1, 1.6, 0.2),	(4.8, 3.1, 1.6, 0.2),
(5.4, 3.4, 1.5, 0.4),	(5.4, 3.4, 1.5, 0.4),	(5.4, 3.4, 1.5, 0.4),
(5.5, 3.5, 1.3, 0.2),	(5.5, 3.5, 1.3, 0.2),	(5.5, 3.5, 1.3, 0.2),
(4.4, 2.9, 1.4, 0.2),	(4.4, 2.9, 1.4, 0.2),	(4.4, 2.9, 1.4, 0.2),
(5, 3.6, 1.4, 0.2)	(5, 3.6, 1.4, 0.2)	(5, 3.6, 1.4, 0.2)

FIGURE 3. Four-dimensional clusters to be assigned.

can be identified by whether these two groups of data are from the same distribution in each dimension.

Let $C_1' = \{x_1, x_2, \ldots, x_m\}$ and $C_2' = \{y_1, y_2, \ldots, y_n\}$ be the sample sets from cluster C_1 and C_2 involved in the test, where $pdf(mean_j)$, and $(1 \leq j \leq K)$ are their object numbers, respectively. On the basis of the W-M-W rank sum test method, we test whether two sets are from the same population by using sample information without a hypothesis of data distribution. Then, we conduct a hypothesis test with the sample data. If it is validated, the null hypothesis is accepted; otherwise, the null hypothesis is rejected. Even though there is a hypothesis used in this method, it is used to make a relatively objective conclusion based on data information rather than as a basis for subsequent analysis.

We designate the null hypothesis as the sets $x_1, x_2, \ldots, x_m \sim F(x - \mu_1)$ and $y_1, y_2, \ldots, y_n \sim F(x - \mu_2)$ have a similar distribution, without regard to data symmetry. Then, the problem for merging two sets can be transformed into the problem to be tested: H_0: $\mu_1 = \mu_2 \leftrightarrow H_1$: $\mu_1 \neq \mu_2$. This is a bilateral test problem. The null hypothesis is that the two sets have no significant difference, come from the same distribution, and can be merged. The alternative hypothesis is that the two sets have significant differences, are from different distributions and cannot be merged. During the analysis, we need to mix $x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_n$ together and assemble these $(m + n)$ numbers in ascending order. The rank of a sample is its position in this ordering sequence. In this mixed ordering sequence, let W_X be the sum of ranks (rank sum) of objects from C_1', while W_Y is the rank sum of the objects from C_2'.

We use the statistics $\min\{W_{XY}, W_{YX}\}$ for this validation problem, where W_{XY} and W_{YX} are shown as Equation (4). W_{XY} is the number of samples from C_2' whose values are greater than the values from C_1', while W_{YX} is the opposite.

$$\begin{aligned} W_{XY} &= mn + \frac{m(m+1)}{2} - W_X \\ W_{YX} &= mn + \frac{n(n+1)}{2} - W_Y \end{aligned} \tag{4}$$

If two sample sets have the same distribution, the ranks of the samples should be randomly mixed. If they have different distributions, one of the rank sums should be greater than the other. Therefore, the rank is used to calculate the statistics,

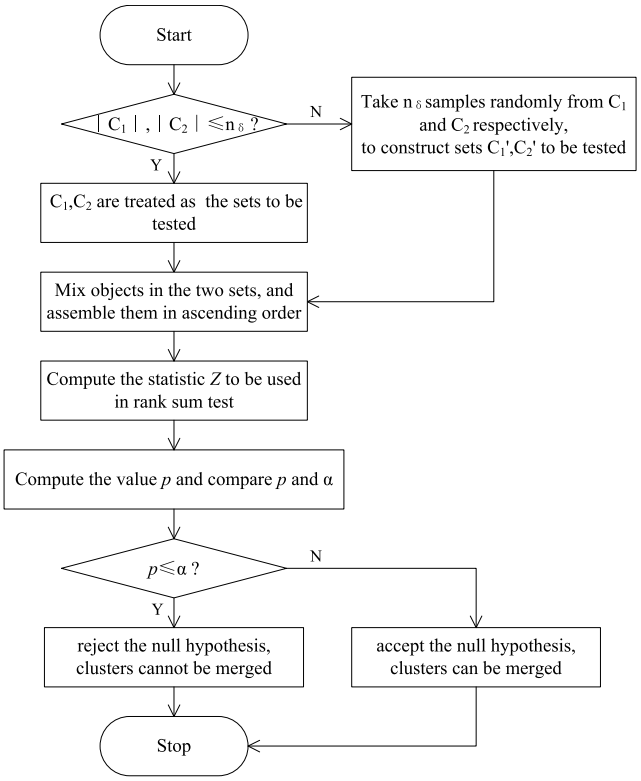


FIGURE 4. **The process of measuring distances between clusters based on W-M-W rank sum test.**

and this method can be used to analyze whether two sets are from the same population without assuming a sample distribution.

In addition, $Z = \frac{W_{XY} - mn/2}{\sqrt{mn(m+n+1)/12}} \to N(0, 1)$. Then, we can calculate the value of p with the corresponding m and n. This p-value is the minimum significance level needed to reject null hypothesis according to the test statistics calculated by the samples [39]–[41]. Then, for a given significance level α, we can obtain the analysis result of the hypothesis testing by comparing p and α. If $p > \alpha$, the null hypothesis is accepted, which indicates that there is no significant difference between the data in these two clusters, and they can be merged. If $p \leq \alpha$, the null hypothesis is rejected, that is, the data in the two clusters are more likely to come from different distributions and cannot be merged.

Fig. 4 describes the specific steps for determining whether two one-dimensional data sets have significant differences based on the above validation method. To sum up, firstly, we construct the two sets of objects to be tested. And we designate the null hypothesis that the two sets come from the same distribution. Then, mix objects in these two sets and assemble them in ascending order, and compute the statistic Z used in rank sum test. In the hypothesis testing, we compute the value p which is the minimum significance level to reject null hypothesis, and then compare p and significance level α. When $p \leq \alpha$, the null hypothesis is rejected, that is, the data in the two sets are more likely to come from different distributions.

The time complexity of the process is $O(n_\delta^2)$, where n_δ is the threshold for the number of objects in one cluster involving the rank sum test method. Even if the cluster has a large number of objects, n_δ samples can be taken randomly to constitute the data set to be tested for further analysis. The feasibility of this sampling method is based on the W-M-W rank sum test method, which is still feasible even with a small sample. Although not all of the objects are used in the analysis, the random sampling of objects will reflect the distribution characteristic to some extent. In addition, the test is based on a nonparametric statistical method. This takes full advantage of the sample data information rather than analyzing the information based on a hypothesis about data distribution. The approach tests whether two clusters are from the same distribution according to the data itself. That is, it analyzes the similarity between clusters from a statistical test perspective. This objectivity ensures the accuracy of the measuring results. Traditional distance metrics are also based on data information and calculate the distance between objects in clusters.

However, the distance values are not the final results of the measurement. They are used to analyze whether two clusters are similar and need to be merged through the comparisons of the distance values.

Therefore, whether two clusters are similar is a relatively comparative result. The distance measurement method proposed in this paper has certain advantages in accuracy and efficiency.

For multi-dimensional data, we assume the dimensions are mutually independent. Thus, multi-dimensional data need to be analyzed for each dimension as above. Once there is a significant difference to be tested in one dimension, this indicates that the data are from different populations on this dimension. It is difficult to illustrate that objects in two clusters have similar features because they are already different in one dimension. Thus, it can be determined that objects in two clusters have significant differences. There is no need to merge these clusters. This process is described in the **ccd**() algorithm. Its time complexity is $O(dn_\delta{}^2)$.

Taking C_1 and C_2 in Fig. 3 and another cluster C^* as an example, we can illustrate the method for determining whether two clusters need to be merged based on the W-M-W rank sum test method. Fig. 5 describes the analysis process.

Clusters C_1 and C_2 in Fig. 5 need to be tested for each of four dimensions. Each value of p is less than the significance level α.

The figure illustrates that these two clusters have a significant difference in all of four dimensions. It can thus be determined that C_1 and C_2 are from two different populations and cannot be merged. However, in the test cluster C_1 and C^* in four dimensions, the p-values are all greater than the significant level α. That is, there is no significant difference between C_1 and C^*, and they can be merged into one cluster.

Obviously, our proposed method can obtain a more objective result than traditional distance metrics because it directly determines whether to merge two clusters based on the

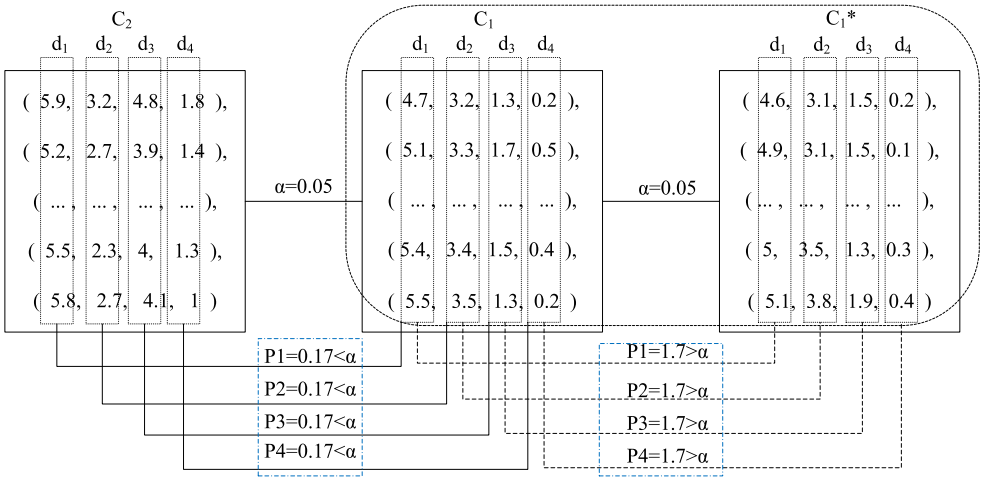


FIGURE 5. Example of measuring distances between clusters based on W-M-W rank sum test.

Algorithm 2 ccd(C_1, C_2, n_δ, α)

Input: C_1, C_2: d-dimensional clusters to be tested;
n_δ: the threshold of the number of objects in one cluster processed by rank sum;
α: the significance level;
Output: mb, whether exists significant difference between C_1 and C_2;
Steps:
1) $mb = 0$;
2) if ($|C_1| > n_\delta$ or $|C_2| > n_\delta$):
3) Draw n_δ samples randomly from C_1 or C_2;
4) Mix these two sets to be tested:
$C'_m = \{x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n\}$;
5) for i:= 1 to d:
6) Make the $(m + n)$ numbers of i-dimension in ascending order;
7) Calculate the statistics $U = \min\{W_{XY}, W_{YX}\}$;
8) Compute the critical value p;
10) if ($p \leq \alpha$):
11) $mb = 1$; //C_1, C_2 have significant difference
12) break;
13) end for

distribution characteristics of the data rather than based on the comparison of distance values. In fact, these data are derived from a UCI dataset—Iris—and the data in C_1 and C^* are from the same class, while data in C_1 and C_2 are from different categories, thus illustrating the accuracy and validity of our method.

III. A DATA DISTRIBUTION FEATURE-ORIENTED HIERARCHICAL CLUSTERING ANALYSIS METHOD

Combined with the above distance measurement method, we propose a two-step hierarchical clustering algorithm to avoid assuming the number of clusters preliminarily, discover clusters of arbitrary shapes and improve clustering accuracy. The above distance measurement methods are the key for proposing such a clustering algorithm. In this hierarchical clustering algorithm, the distance metrics proposed in Section 2 are used to assign objects to the proper clusters and determine whether to merge clusters.

In the first step, K-means is used to generate a number of clusters as the initial clusters by dividing objects in the original data set. We don't need to determine the final cluster number in this step. We can set the parameter 'k' of K-means to a larger value, and then obtain multiple initial clusters. There may exist similar distribution features among these initial clusters. Therefore, in the second step, we use a hierarchical clustering to merge similar clusters generated in the first step, and assign their objects into the same cluster. During this step, we determine whether two clusters are similar and need to be merged based on Algorithm 2. This operation continues until all clusters are determined to have significant differences between each other, where the data in different clusters are likely to come from different populations. Then, the clustering process finishes. This two-step hierarchical clustering algorithm is described as follows.

The time complexity of obtaining the initial cluster is $O(tkdn)$, where t is the number of iterations, k is the number of initial clusters, d is the dimension of the data, and n is number of objects. Based on the above analysis, the time complexity of the merging step is $O(t'k^2dn_\delta^2)$, where ρ is the number of iterations for the merging step. Therefore, this two-step hierarchical clustering algorithm based on nonparametric statistics has a time complexity of $O(tkdn + t'k^2dn_\delta^2)$, where k, $n_\delta \ll n$. Obviously, the proposed algorithm is effective. The final number of clusters is generated based on the distribution features of the data rather than a presumed value. In addition, the accuracy of the distance metric proposed in Section 2 can facilitate ensuring the accuracy of the results generated by the proposed unsupervised clustering algorithm.

Algorithm 3 NPSC(D, k, δ, n_δ, α)

Input: D $= \{x_1, x_2, \ldots, x_n\}$, dataset;
k, the number of the generated initial clusters;
δ, the threshold of variations about distribution features;
n_δ, the threshold of the number of objects in one cluster processed by rank sum;
α, the significance level;
Output: $C = \{C_1, C_2, \ldots, C_{K'}\}$, the clustering result;
Steps:
(1) Generate initial clusters
1) choose k objects from dataset D to be the initial cluster centers, then we can obtain $C^t = \{C_1, C_2, \ldots, C_k\}$;
2) repeat
3) for $i := 1$ to n do
4) $C_\omega = \mathbf{ocd}(x_i, C, \delta)$;
5) assign object x_i into cluster C_ω;
6) end for
7) update the distribution features of k clusters as Equation (5):

$$DF_j = \{DF_{j1}, \ldots, DF_{jd}\} = \{< \mu_{j1}, \sigma_{j1}, m_{j1} >, \ldots, < \mu_{jd}, \sigma_{jd}, m_{jd} >\} \quad 1 \le j \le k \tag{5}$$

8) compute the objective function: $E = \sum_{j=1}^{k} DF_j$;
9) until the objective function E converges.
(2) Merge similar ones in initial clusters $\{C_1, C_2, \ldots, C_k\}$
1) Let $K' = k$;
2) repeat
3) for $i := 1$ to K' do
4) for $j := i + 1$ to K' do
5) $mb = \mathbf{ccd}(C_i, C_i, n_\delta, \alpha)$;
6) if $(mb = 1)$ then:
7) C_i and C_j have significant difference, do not merge them;
8) else if $(mb = 0)$ then:
9) C_i and C_i do not have significant difference, merge them into one cluster;
10) end for
11) Let K' be the number of clusters after merging operations;
12) until there exist significant differences between any two clusters.

IV. EXPERIMENTS

We select three two-dimensional data sets and several UCI data sets to verify the validity of the proposed clustering algorithm based on the W-M-W rank sum test method. In addition, we compare this approach with the K-means, DBSCAN, Birch, UPGMA [26], and Fast [27] algorithms to assess the run time and accuracy of the clustering results. The results will illustrate the effectiveness and practicality of our algorithm.

For the datasets with marked categories, we use the external indices Purity and Entropy [42] to evaluate the accuracy of the clustering results. Let $C = \{C_1, \ldots, C_{K'}\}$ be the clustering result, and $P = \{P_1, \ldots, P_l\}$ represents the given categories of data, where K' is the number of generated clusters and l is the number of original categories. Then, Purity and Entropy can be calculated as:

$$\text{Purity:} \quad Purity = \sum_{i=1}^{K'} \frac{1}{N} \max_j (n_i^j),$$

$$\text{Entropy:} \quad Entropy = \sum_{i=1}^{K'} \frac{n_i}{N} \left(-\frac{1}{\log l} \sum_{j=1}^{i} \frac{n_i^j}{n_i} \log \frac{n_i^j}{n_i}\right),$$

where N is the number of objects in the dataset, n_i^j is the number of objects divided into the i-th cluster which belong to the j-th category in the original dataset, and n_i is the number of objects divided into i-th cluster. The higher the purity is, the more accurate the clustering result is. The lower the entropy is, the more accurate the clustering result is. Ideally, Entropy $= 0.0$ and Purity $= 1.0$.

A. TWO-DIMENSIONAL DATASETS

We select three two-dimensional graphic data sets—Aggregation, Spiral and Flame—to verify that the proposed method can discover clusters with arbitrary shapes. These datasets contain similar spatial data within the same category (clusters), not only simple spherical clusters. They can also be visualized. Therefore, these datasets can be used to validate the capacity of discovering clusters with arbitrary shapes. Fig. 6 shows the visualized clustering results of three two-dimensional datasets obtained by our proposed NPSC algorithm. It can be seen that NPSC can identify the clusters of data, that is, it can discover clusters with arbitrary shapes.

This result is mainly attributable to the distance measurement method used in the proposed algorithm. It determines the similarities between clusters based on the distribution features of the data rather than traditional distance metrics. This method can merge similar clusters according to the characteristics of the data based on the nonparametric statistical hypothesis test method without a hypothesis of data distributions. In addition, this approach can be used in the second step of the proposed clustering algorithm. A number of closely similar clusters are generated in the first step of the clustering process. Then, clusters with similar distribution features can be discovered based on our proposed distance measurement method. In addition, the similar clusters are merged into on clusters.

These characteristics make the proposed clustering method better suited to discover nonspherical clusters.

B. UCI DATASETS

Then, we cluster the UCI datasets shown in Table 1 and compare the results with other clustering algorithms to verify the effectiveness and accuracy of our proposed algorithm.

Fig. 7 compares the accuracy of the clustering results based on the evaluation indices Purity and Entropy. Obviously, NSPC obtains relatively higher Purity values and lower Entropy values than the other algorithms. Although there

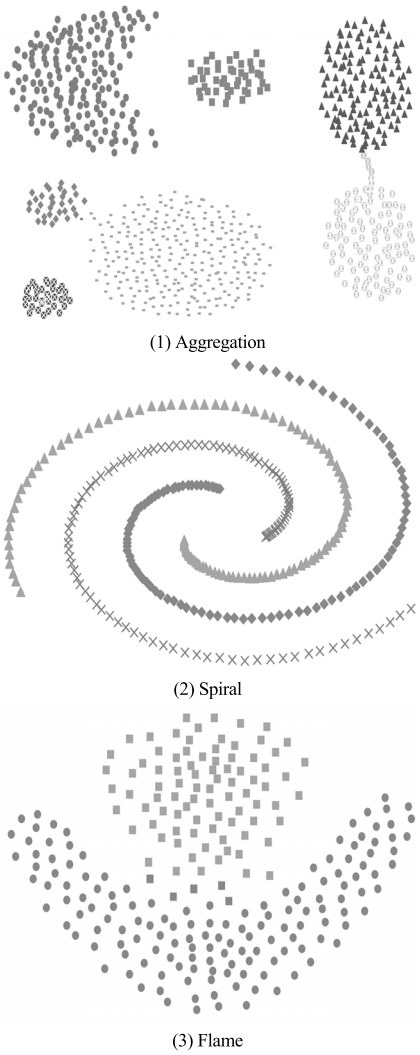


FIGURE 6. Clustering results for two-dimensional datasets obtained by NPSC.

TABLE 1. UCI datasets.

Dataset	Object number	Attribute number	Category number
Abalone	4177	8	16
Ecoli	336	8	8
Iirs	150	4	3
Letter	20000	16	26
Yeast	1484	8	10

are multiple choices for the input parameters of comparative algorithms, which may help to obtain more accurate clustering results, our proposed algorithm still has advantages. Because our algorithm doesn't need to set parameters that directly affect the accuracy of clustering result. The clustering result mainly depends on the data distribution. So compared with the other algorithms, our algorithm could obtain relatively steady and accurate results, without depending on the parameters directly. This indicates that NSPC can be more likely to obtain more accurate clustering results.

This is mainly due to the proposed distance measurement method, where the unsupervised clustering analysis determines the generation of clusters based on the results of distance measurement. The proposed method does not assign objects into clusters based on the relatively comparative distances, such as in K-means or Birch. It also does not depend on the neighborhood radius parameter to determine the density of clusters as in DBSCAN.

NSPC does not need to entirely depend on the numerical distance measurement results as traditional methods do. It is based on a nonparametric statistical hypothesis testing method and determines whether clusters are similar according to the distribution features of the data.

UPGMA and Fast have improved the shortcomings inherent in traditional clustering methods. However, UPGMA still extracts cluster features based on neighboring objects, that is, it also depends on the distances to some extent. While NSPC draws samples in each cluster randomly during the similarity analysis between clusters, these samples reflect the distribution features of clusters to some extent. Fast uses a probability density function to obtain the distribution features of clusters. However, this method needs to make assumptions about the data distributions, and these assumption are more likely to not match the real data distributions. Therefore, its results obtain a lower Purity value and a higher Entropy value compared with NSPC. These results illustrate that our proposed distance metric can obtain more accurate clustering results.

Fig. 8 compares the run time results between these clustering algorithms on the UCI datasets. K-means exhibits high efficiency due to its linear time complexity. NSPC has a run time close to Birch, which is also a hierarchical clustering method. In addition, NSPC has a relatively high efficiency compared to the other algorithms. UPGMA needs to obtain neighboring objects and then calculate clustering features, and Fast needs to calculate probability density distribution functions for clusters. These operations require considerable computation time.

Clearly, our proposed algorithm can not only obtain relatively accurate clustering results but also provide high efficiency.

This is due to the use of our proposed distance metric, which is based on the distribution features of data during clustering.

This approach also relies on our proposed two-step clustering process, which can help to ensure the accuracy and effectiveness of the clustering. Because we use several descriptive statistics to represent data distribution features in a cluster when measuring distance between an object and a cluster, we analyze the distribution feature variations once the object is assigned to a cluster. Therefore, we can obtain a more objective similarity result between the object and a cluster. We use the W-M-W rank sum test method to measure distances between clusters. This avoids inaccuracies when determining whether to merge clusters according to less objective comparison values in traditional metrics. This approach also ensures the efficiency of the clustering process by not using

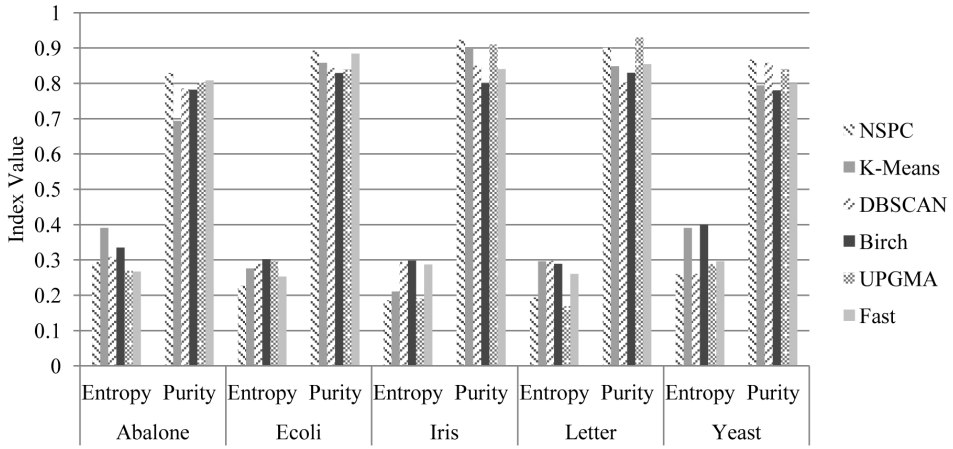


FIGURE 7. Comparison of accuracy for clustering results on UCI datasets.

TABLE 2. Telephone traffic dataset.

DCH	DSCH	Time	Longitude	Latitude
26.23	18.11	2013/2/4 1:00	126.6833	45.75097
15.26	12.54	2013/2/4 2:00	126.6741	45.74918
…	…	…		…

all objects in the clusters. The two-step process assures that the number of generated clusters will not depend on a preassumed value. This determines when the clustering process is terminated through the analysis of data distribution based on a nonparametric statistical hypothesis test. The final number of clusters does not need to be relative to the initial number set by the parameter.

C. REAL DATASET

We further verify the effectiveness of our proposed algorithm on a real dataset in a communication field. The dataset in Table 2 contains the data of user call volume in an hour covered by every base station in a city.

Based on this telephone traffic dataset, we first compare the clustering accuracy with other clustering algorithms. Second, based on the clustering results, we analyze the user behavior patterns for every day in the regions covered by each base station. That is, different clusters have different data distribution features, which are reflected as different user calling behaviors in this city. In addition, we can further analyze the regional functions of the city through different user data.

By comparing our approach with the results obtained by the other methods, we can verify the effectiveness of our proposed algorithm. Through the user behavior patterns and regional function discovery based on the clustering results, we can illustrate the practicability of the method.

The original data label information is difficult to obtain in most practical application domains, especially in a communication field that consistently generates data. Therefore, we use the relative evaluation indices Dunn [43] and DB [44] to analyze the clustering accuracy on this real dataset [45]. These two indices can be calculated as:

(1) $Dunn = \min_{i=1,\ldots,n_c}\{\min_{j=i+1,\ldots,n_c}\left(\frac{d(c_i,c_j)}{\max_{k=1,\ldots,n_c} diam(c_k)}\right)\}$, where $diam = \sum_{i=1,\ldots,n}\sum_{o_p,o_q\in C} d(o_p,o_q)/n$

(2) $DB = \frac{1}{n_c}\sum_{i=1}^{n_c}\max_{j\neq i}\left(\frac{\overline{C_i}+\overline{C_j}}{d(C_i,C_j)}\right)$

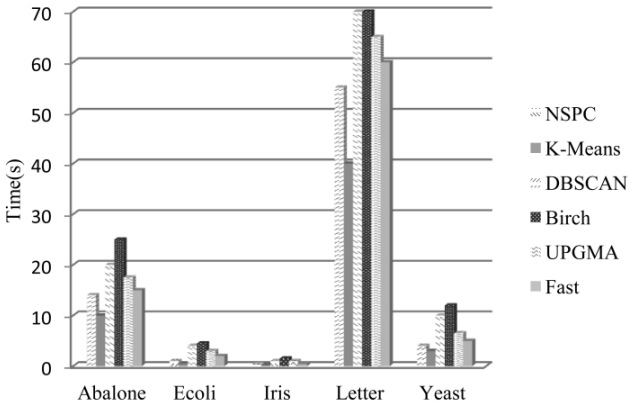


FIGURE 8. Comparison of run time for clustering results on UCI datasets.

Both datasets measure compactness within clusters and the separation between clusters. We can evaluate the relative validity of the clustering results through a comparison of their index values. The greater the Dunn value is, the more accurate the clustering result is, while DB performs in an opposite manner.

Fig. 9 shows a comparison of the clustering results on this real dataset based on different clustering algorithms.

This figure shows that with the use of distribution features, the clustering results generated by our proposed algorithm have larger index Dunn values and smaller DB values, i.e., our results are relatively more accurate.

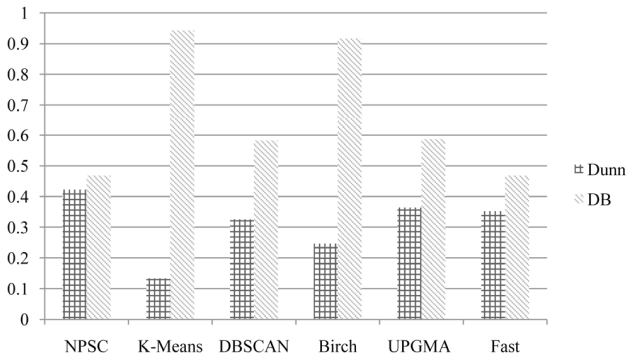


FIGURE 9. Comparison of clustering accuracy on traffic dataset.

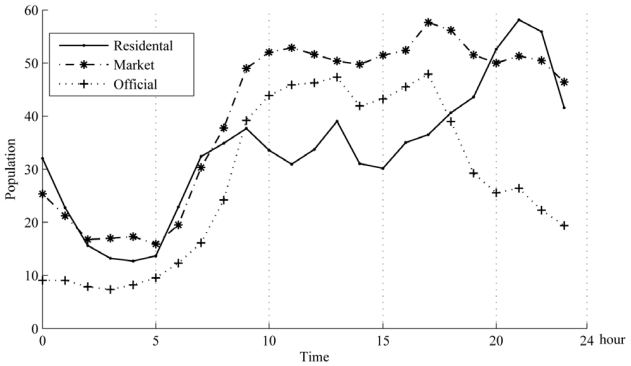


FIGURE 10. The distribution characteristics of clusters generated on a real dataset.

This result occurs because the attributes of the real dataset are numerical, which are ideal for obtaining distribution features. Our proposed clustering algorithm takes advantage of the data distribution features and obtains a relatively more accurate clustering result.

We can also describe the distribution diagram for the clustering result based on our proposed algorithm, which is shown in Fig. 10. The figure illustrates that with the application domain knowledge, the generated significant different clusters correspond to different types of regional functions in the city, i.e., the official region, residential region and market region. These regions have unique distribution features that give them different functions. For example, the telephone traffic always increases and maintains a certain volume in the official region during working hours because of the nature of the work. Telephone traffic has relatively lower values in the daytime and higher values after working hours in the residential areas. We can thus develop service mechanisms for the base stations in these different regions based on the clustering results.

From the above experiments, we find that our proposed clustering algorithm can generate clusters with arbitrary shapes and is applicable to datasets of many types, including UCI datasets and telephone traffic data. These results illustrate that the proposed method has a higher reliability when used to discover inherent data distribution characteristics.

V. CONCLUSION

To ensure the reliability of obtaining inherent data distribution, a distance metric is proposed based on descriptive statistics and nonparametric statistical methods. The distance measurement method based on nonparametric statistics could take full advantage of the data distribution features, obtain clusters in a more straightforward and objective way compared with traditional distance metrics. In addition, a two-step hierarchical clustering algorithm is also proposed. The proposed clustering algorithm can avoid the presumed initial number of clusters with its two-step characteristics; the final number of clusters does not need to be relative with the initial number set by the input parameter. It can also discover clusters with arbitrary shapes and obtain more accurate results due to the distance metrics: it determines the similarities between clusters on the basis of data distribution features.

Therefore, the proposed distance measurement method can provide stronger support for unsupervised clustering analysis. In addition, the clustering algorithm can be used to analyze data with unknown category information. They both facilitate ensuring the reliability of obtaining data aggregation features, which is illustrated by experiments on different datasets.

REFERENCES

[1] M. Verma, M. Srivastava, N. Chack, A. K. Diswar, and N. Gupta, "A comparative study of various clustering algorithms in data mining," *Int. J. Eng. Res. Ind. Appl.*, vol. 2, no. 3, pp. 1379–1384, May 2012.

[2] Y. Lin, Y. Tu, and Z. Dou, "An improved neural network pruning technology for automatic modulation classification in edge devices," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5703–5706, May 2020.

[3] Y. Lin, Y. Tu, Z. Dou, L. Chen, and S. Mao, "Contour stella image and deep learning for signal recognition in the physical layer," *IEEE Trans. Cognit. Commun. Netw.*, early access, Sep. 18, 2020, doi: 10.1109/TCCN.2020.3024610.

[4] H. Li, J. Liu, Z. Yang, R. W. Liu, K. Wu, and Y. Wan, "Adaptively constrained dynamic time warping for time series classification and clustering," *Inf. Sci.*, vol. 534, pp. 97–116, Sep. 2020.

[5] H. Li, J. Liu, R. Liu, N. Xiong, K. Wu, and T.-H. Kim, "A dimensionality reduction-based multi-step clustering method for robust vessel trajectory analysis," *Sensors*, vol. 17, no. 8, p. 1792, Aug. 2017.

[6] Y. Huang, Y. Li, Z. Zhang, and R. W. Liu, "GPU-accelerated compression and visualization of large-scale vessel trajectories in maritime IoT industries," *IEEE Internet Things J.*, vol. 7, no. 11, pp. 10794–10812, Nov. 2020.

[7] Y. Lin, Y. Li, X. Yin, and Z. Dou, "Multisensor fault diagnosis modeling based on the evidence theory," *IEEE Trans. Rel.*, vol. 67, no. 2, pp. 513–521, Jun. 2018.

[8] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th BSMSP*, Berkeley, CA, USA, 1967, vol. 1, no. 14, pp. 281–297.

[9] C. W. Tsai, C.-F. Lai, M.-C. Chiang, and L. T. Yang, "Data mining for Internet of Things: A survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 77–97, 1st Quart., 2014.

[10] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, Portland, OR, USA, 1996, pp. 226–231.

[11] S. F. Galán, "Comparative evaluation of region query strategies for DBSCAN clustering," *Inf. Sci.*, vol. 502, pp. 76–90, Oct. 2019.

[12] S. Pandit and S. Gupta, "A comparative study on distance measuring approaches for clustering," *Int. J. Res. Comput. Sci.*, vol. 2, no. 1, pp. 29–31, Dec. 2011.

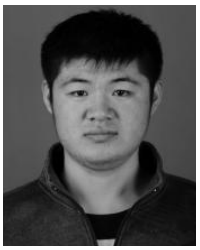
[13] D. G. Ferrari and L. N. de Castro, "Clustering algorithm selection by meta-learning systems: A new distance-based problem characterization and ranking combination methods," *Inf. Sci.*, vol. 301, pp. 181–194, Apr. 2015.

[14] A. KumarPatidar, J. Agrawal, and N. Mishra, "Analysis of different similarity measure functions and their impacts on shared nearest neighbor clustering approach," *Int. J. Comput. Appl.*, vol. 40, no. 16, pp. 1–5, Feb. 2012.

[15] M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 217–227, Jan. 2014.
[16] Z. Kang, H. Xu, B. Wang, H. Zhu, and Z. Xu, "Clustering with similarity preserving," *Neurocomputing*, vol. 365, pp. 211–218, Nov. 2019.
[17] W. S. Sarle, "Finding groups in data: An introduction to cluster analysis," *J. Amer. Stat. Assoc.*, vol. 86, no. 415, pp. 830–833, Sep. 1991.
[18] M. Mohibullah, M. Z. Hossain, and M. Hasan, "Comparison of Euclidean distance function and Manhattan distance function using K-mediods," *Int. J. Comput. Sci. Inf. Secur.*, vol. 13, no. 10, pp. 61–71, Oct. 2015.
[19] S. T. Mai, I. Assent, and A. Le, "Anytime OPTICS: An efficient approach for hierarchical density-based clustering," in *Proc. 21st Int. Conf. DSAA*, Dallas, TX, USA, 2016, pp. 164–179.
[20] P. Sharma and K. A. Ramya, "Review on density based clustering algorithms for very large datasets," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 3, no. 12, pp. 398–403, 2013.
[21] R. J. G. B. Campello, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdiscip. Rev. Data Mining Knowl. Discovery*, vol. 10, no. 2, p. e1343, Oct. 2020.
[22] E. Bae, J. Bailey, and G. Dong, "A clustering comparison measure using density profiles and its application to the discovery of alternate clusterings," *Data Mining Knowl. Discovery*, vol. 21, no. 3, pp. 427–471, Jan. 2010.
[23] P. Bhattacharjee and P. Mitra, "A survey of density based clustering algorithms," *Frontiers Comput. Sci.*, vol. 15, no. 1, pp. 1–27, Feb. 2021.
[24] G. Sehgal and D. K. Garg, "Comparison of various clustering algorithms," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 3, pp. 3074–3076, 2014.
[25] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large database," *ACM SIGMOD Rec.*, no. 415, Jun. 1996, pp. 103–114.
[26] A. Bouguettaya, Q. Yu, X. Liu, X. Zhou, and A. Song, "Efficient agglomerative hierarchical clustering," *Expert Syst. Appl.*, vol. 42, no. 5, pp. 2785–2797, Apr. 2015.
[27] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 1–14, Jan. 2013.
[28] C. Boutsidis and M. Magdon-Ismail, "Deterministic feature selection for K-means clustering," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 6099–6110, Sep. 2013.
[29] I. Antoniano-Villalobos, E. Borgonovo, and X. Lu, "Nonparametric estimation of probabilistic sensitivity measures," *Statist. Comput.*, vol. 30, no. 2, pp. 447–467, Mar. 2020.
[30] W. E. Wansouwé, C. C. Kokonendji, and D. T. Kolyang, "Nonparametric estimation for probability mass function with Disake," *ARIMA J.*, vol. 19, no. 415, pp. 1–23, 2015.
[31] J.-N. Hwang, S.-R. Lay, and A. Lippman, "Nonparametric multivariate density estimation: A comparative study," *IEEE Trans. Signal Process.*, vol. 42, no. 10, pp. 2795–2810, Oct. 1994.
[32] F. Comte and N. Marie, "Nonparametric estimation in fractional SDE," *Stat. Inference Stochastic Processes*, vol. 22, no. 3, pp. 359–382, Oct. 2019.
[33] B. Trawiński, M. Smętek, Z. Telec, and T. Lasota, "Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms," *Int. J. Appl. Math. Comput. Sci.*, vol. 22, no. 4, pp. 867–881, Dec. 2012.
[34] G. I. Salov, "Controllability of a new nonparametric statistical criterion alternative to the Wilcoxon—Mann—Whitney test," *Numer. Anal. Appl.*, vol. 12, no. 3, pp. 263–269, Jul. 2019.
[35] R. Bergmann, J. Ludbrook, and W. Spooren, "Different outcomes of the Wilcoxon—Mann—Whitney test from different statistics packages," *Amer. Stat.*, vol. 54, no. 1, pp. 72–77, Feb. 2000.
[36] F. Dexter, "Wilcoxon-Mann-Whitney test used for data that are not normally distributed," *Anesthesia Analgesia*, vol. 117, no. 3, pp. 537–538, Sep. 2013.
[37] D. W. Zimmerman and B. D. Zumbo, "Mann-Whitney test and student *t* test under simple bounded transformations," *J. Gen. Psychol.*, vol. 117, no. 4, pp. 425–436, 2017.
[38] M. A. Conde, F. García, M. J. Rodríguez-Conde, M. Alier, and A. García-Holgado, "Perceived openness of learning management systems by students and teachers in education and technology courses," *Comput. Hum. Behav.*, vol. 31, pp. 517–526, Feb. 2014.
[39] D. R. Anderson, K. P. Burnham, and W. L. Thompson, "Null hypothesis testing: Problems, prevalence, and an alternative," *J. Wildlife Manage.*, vol. 64, no. 4, pp. 912–923, Oct. 2000.
[40] J. W. Schneider, "Null hypothesis significance tests. A mix-up of two different theories: The basis for widespread confusion and numerous misinterpretations," *Scientometrics*, vol. 102, no. 1, pp. 411–432, Jan. 2015.
[41] B. Efron, "Large-scale simultaneous hypothesis testing: The choice of a null hypothesis," *J. Amer. Stat. Assoc.*, vol. 99, no. 465, pp. 96–104, 2004.
[42] G. Kou, Y. Peng, and G. Wang, "Evaluation of clustering algorithms for financial risk analysis using MCDM methods," *Inf. Sci.*, vol. 275, pp. 1–12, Aug. 2014.
[43] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *J. Cybern.*, vol. 4, no. 1, pp. 95–104, Jan. 1974.
[44] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.
[45] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1650–1654, Dec. 2002.

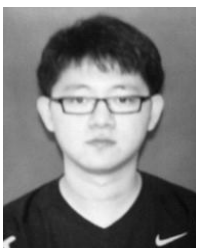
YUAN CHENG was born in Harbin, China, in 1985. She received the M.S. degree in computer science and technology from the Harbin Institute of Technology, China, in 2009, and the Ph.D. degree in computer applied technology from the Harbin University of Engineering, China, in 2014.

Since 2014, she has been a Lecturer with the College of Computer Science and Technology, Harbin University of Science and Technology. Her research interests include data mining, machine learning, and uncertainty analysis.



WEINAN JIA was born in Handan, China, in 1995. He received the bachelor's degree in Internet of Things engineering from Xuchang University, Henan, China, in 2019. He is currently pursuing the degree in computer science and technology from the Harbin University of Science and Technology.

He is also working as an Assistant with the Artificial Intelligence Laboratory, Harbin University of Science and Technology. His research interest includes mobile population.



RONGHUA CHI was born in Mudanjiang, China, in 1981. He received the Ph.D. degree in computer science and technology from the Harbin University of Engineering, China, in 2018.

Since 2018, he has been a Lecturer with the College of Computer Science and Technology, Heilongjiang University of Science and Technology. His research interests include big data analysis, data mining, machine learning, and uncertainty analysis.

AO LI (Member, IEEE) received the B.S. and Ph.D. degrees from Harbin Engineering University, Harbin, China, in 2009 and 2014, respectively.

From 2017 to 2018, he was a Research Assistant with Wright State University. He currently works with the Harbin University of Science and Technology, Harbin. His current research interests include sparse representation, pattern recognition, machine learning, and their applications to computer vision.