A Clustering Analysis Method Based on Wilcoxon-Mann-Whitney Testing

Yuan Cheng[1], Weinan Jia[1] and Ronghua Chi[2]
{changuang7@sina.com[1], jiaweinanhd@163.com[1], chironghua@126.com[2]}

College of Computer Science and Technology, Harbin University of Science and Technology[1],
College of Computer and Information Engineering, Heilongjiang University of Science and Technology[2]

Abstract. As the core step of clustering analysis, the results of distance measurements will influence the clustering accuracy. The existing measurements are mostly based on the information about cluster features. However, the cluster features may be not sufficient enough and would result in losing data information about clusters containing a number of objects. To improve the measurement accuracy, we make full use of the distribution characteristics of objects in clusters, so we use the descriptive statistics and the Wilcoxon-Mann-Whitney rank sum test in nonparametric statistics to measure distances during clustering. Furthermore, a two-stage clustering is proposed to improve the performance of clustering analysis, from the aspects ofavoiding assuming the number of clusterspreliminarily, discovering clusters of arbitrary shapes andimproving clustering accuracy. The experiments on multiple datasets compared with other clustering algorithms illustrate the accuracy and efficiency of the proposed clustering algorithm.

Keywords: Clustering analysis, distance measurement, nonparametric statistics, Wilcoxon-Mann-Whitney rank sum test.

1 Introduction

As abasic technology of data mining, clustering analysis is significant in discovering the characteristics of data aggregation [1]. Clustering analysis is an unsupervised method, whether the partitioning methods, hierarchical methods, density-based methods, or grid-based methods, their implicit clustering ideas are similar, that is based on the distances between objects, through the iterations to ensure the clustering quality. K-Means [2] adjusts the clusters the

EAI MOBIMEDIA 2020, August 27-28, Harbin, People's Republic of China

DOI 10.4108/eai.27-8-2020.2296732

objects belonging to in each iteration; DBSCAN [3] aggregates the objects that could directly density reachable from the core object in each iteration to generate new clusters; so their difference is the way to divide the objects into clusters during the clustering process. In this unsupervised analysis process, 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 in DBSCAN is also based on the distances between objects; the clusters mergence is decided by the distances between clusters in agglomerative hierarchical clustering. It is obvious that the accuracy of distance measurement is the important basis for effective clustering.

The existing distance measurement between objects could 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 [4-6]. And the distance involving clusters are mostly measured based on the information reflecting cluster features. For instance, K-Means and K-medoids [7] choose the mean value or a representative object as the feature of a cluster; they assign an object into the cluster whose feature is closest to it; DBSCAN and OPTIS [8] 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 [9].

Actually, the objects in each cluster are the main factors truly reflecting cluster features, therefore the distance between clusters could be calculated through the distance between objects in the clusters, such as the minimum distance, the maximum distance, and the average distance. However the larger number of objects may affect the efficiency of distance measurement. Therefore, to improve the computation speed and scalability, Birch [10] uses zero moment, first moment and second moment to generate a three-dimensional vector, which is represented as the cluster feature to summarize cluster information and to compute the distance between clusters for hierarchical clustering. But it is not sufficient to describe the clustering information by using the representative objects or the statistics. The clusters contain a number of objects, and the existing cluster features represent the aggregation features of clusters. They would loss information reflecting the data characteristics of clusters to a certain extent. And distance measurement would be with a certain deviation, and thus affect the accuracy of clustering results.

Obviously, the distance measurement is a core step in clustering. The effective information extraction to representing the data features of clusters is the key to ensure the metric accuracy. And it is also the key to ensure the accuracy of clustering. Therefore, researchers are trying to extract effective information about cluster features. [11] extracted some adjacent objects of centroids to summarize the cluster information. It used a group of representative objects, but the adjacent objects are not enough to reflect the general data features. [12] defined a coreset to measure distances with the idea of Birch. Although they choose a number of objects as the representative information of clusters, it is also insufficient so as to result in the information loss. The distribution of data in the clusters could reflect the general cluster data features. [13] has obtained the distribution features of clusters based on probability density function. However, this method needs to presuppose the data distribution. It is difficult to make a clearer assumption about the data distribution, due to the little knowledge about the overall information. And the incorrect assumption would result in the inaccuracy of distance measure and even the clustering results.

Nonparametric statistical methods [14,15] can be used to estimate the distribution structures based on the data information directly, rather than based on a hypothesis about the specific form of the overall distribution. The Wilcoxon-Mann-Whitney (W-M-W) rank sum test method [16,17] is one of nonparametric statistical methods. It is always used to judge whether any two sets come from a same population. In the clustering process, if two sets represented by two clusters are considered from the same population, they could be grouped into one cluster. Therefore, through this method we could 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 the hypothesis about the overall distribution form.

This paper will propose a new distance measurement method based on W-M-W rank sum test to resolve the above problems, and then propose an improved hierarchical clustering method to increase the clustering effectiveness. This method will need minimal requirements for domain knowledge to determine input parameters, discover clusters with arbitrary shape and improve clustering accuracy. Experiments on multiple datasets are used to verify the validity of the proposed algorithm. Finally, this paper is concluded.

2 Distance measurement based on Nonparametric Statistics

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

2.1 Distance between objects and clusters based on distribution characteristics

In the traditional clustering methods, the distances between objects and clusters are often transformed into the distances between objects and cluster features. The features include the cluster mean values, representative objects of clusters, coresets of clusters and so on. As mentioned above, these single descriptive features would lose the data information of clusters to some extent. And the distribution characteristics of objects in one cluster could be seen as its general cluster features. The difference between clusters is about the difference between their distribution characteristics of objects. While if we want to divide an object into some cluster, the distance between it and a cluster is the basis of assignment. Then the distribution characteristics of objects in this cluster would have an impact on distance measurement and should be considered.

The distribution characteristics of objects in a cluster are necessary to be considered when measuring distances. Obviously, the mean value as one statistical feature of a set could not completely reflect its distribution, while the representative also could not reflect the distribution characteristics of all objects in the cluster. Therefore these measure methods could not calculate differences between objects and clusters objectively. Sometimes the use of one statistic is insufficient to represent the general characteristics of data in a set after all.

Therefore, for a more accurate and objective result, we need to divide an object into some cluster with the consideration of distribution characteristics of data in this cluster. The methods that can describe the distribution features of a set include probability distribution functions and descriptive statistics. It is more time consuming to compute the probability distribution function of objects for each cluster. While multiple descriptive statistics could describe the statistical characteristics of a set from different perspectives; so it could be used to represent its distribution characteristics. Therefore, this paper will measure the distance between objects and clusters based on some 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. This paper will consider different descriptive statistics of a cluster when measuring distances between an objects and the cluster; and analyze that whether these statistics have changed significantly after the object is divided into it. We will determine the right cluster with the smallest change of the statistics and also below a threshold.

In descriptive statistical analysis, the statistics such as 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 about 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 multi-dimensional data. We will 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 is the one-dimensional object to be assigned, the existing clusters are $C=\{C_1,C_2,\ldots,C_n\}$. The distribution feature of cluster $C_i(1\le i\le n)$ can be described by a triple $DF_i=<\mu_i,\sigma_i,m_i>$, where μ_i,σ_i,m_i are the mean value, variance and median, which represent the average value, dispersion and the center position of cluster C_i. If the object is divided into C_i, its distribution feature would be $DF_i'=<\mu_i',\sigma_i',m_i'>$. And the variation of distribution feature could be calculated as equation (1).

$$\Delta_i=\left|\mu_i'-\mu_i\right|+\left|\sigma_i'-\sigma_i\right|+\left|m_i'-m_i\right|\ (1\le i\le n) \tag{1}$$

If o_1 belongs to the 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, 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, 0.31, 5.8> and DF_3=<6.3, 0.52, 6.3> respectively. The object to be divided is $o_1=5.7$. The threshold of variation about distribution feature is δ=0.1.

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

If we extend the above method to multi-dimensional data, we need to determine the relation between the object and the cluster distribution feature in each dimension as described above. Then integrate the analysis results on each dimension to determine the cluster having the smallest variation about distribution feature after the object is added into it. This cluster is more similar to the object than 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 $C_i (1 \le i \le n)$ can be described by a *d*-dimensional triple as equation (2).

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

During the analysis, we could calculate the distribution feature in the *k*-th $(1 \le k \le d)$ dimension of every cluster: $DF_{ik}' = < \mu_{ik}', \sigma_{ik}', m_{ik}' > \ (1 \le i \le n)$, when o_2 is assumed to be divided into each cluster. In addition, the variation about distribution feature in the *k*-th dimension could be also calculated as equation (3).

$$\Delta_{ik} = |\mu_{ik}' - \mu_{ik}| + |\sigma_{ik}' - \sigma_{ik}| + |m_{ik}' - m_{ik}| \tag{3}$$

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

We can compute the variations about distribution feature for each cluster with the assumption of the 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 **Figure 1** for instance 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: *C₁*, *C₂* and *C₃*. Their distribution features represented by 4-dimensional triples are as equation (4).

$$\begin{aligned} DF_1 &= \{< 4.99,\ 0.39,\ 5 >, < 3.29,\ 0.24,\ 3.3 >, < 1.46,\ 0.15,\ 1.4 >, < 0.27,\ 0.13,\ 0.2 >\} \\ DF_2 &= \{< 5.74,\ 0.31,\ 5.8 >, < 2.83,\ 0.29,\ 2.9 >, < 4.3,\ 0.34,\ 4.2 >, < 1.37,\ 0.25,\ 1.4 >\} \\ DF_3 &= \{< 6.3,\ 0.52,\ 6.3 >, < 3.04,\ 0.24, 3 >, < 5.37,\ 0.48,\ 5.1 >, < 2.06,\ 0.22,\ 2 >\} \end{aligned} \tag{4}$$

The threshold of variation about distribution features is *δ* = 0.8. Then we can obtain the new triples as shown in equation (5).

$$\begin{aligned} DF_1' &= \{< 5.07,\ 0.44,\ 5.05 >, < 3.425,\ 0.45,\ 3.35 >, < 1.46,\ 0.14,\ 1.45 >, < 0.29,\ 0.12,\ 0.2 >\} \\ DF_2' &= \{< 5.74,\ 0.29,\ 5.75 >, < 3.025,\ 0.62,\ 2.95 >, < 3.95, 1.04,\ 4.15 >, < 1.25,\ 0.41, 1.35 >\} \\ DF_3' &= \{< 6.22,\ 0.53,\ 6.2 >, < 3.21,\ 0.53,\ 3.1 >, < 4.89,\ 1.44,\ 5.1 >, < 1.85,\ 0.62,\ 2 >\} \end{aligned} \tag{5}$$

If we assume the object is divided into these clusters respectively. The variations with the former are $\Delta_1 = 0.665$, $\Delta_2 = 2.075$, $\Delta_3 = 2.8$, respectively. Obviously, Δ_1 is minimum and less than the threshold *δ*. Taking into account all the four dimensions, the object is more likely to be from the same distribution with data in cluster *C₁*. So it could be divided into *C₁*. This result is different from the above 1-dimension analysis, since object o_2 is described by the four dimensions, and its assignment is based on distribution features on all dimensions, rather than one dimension.

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)

Fig. 1. 4-dimensional clusters to be assigned

2.2 Distance between clusters based on rank sum test

The main purpose of measuring 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 are similar could be considered that they are more likely from the same distribution. The W-M-W rank sum test method is one of nonparametric statistics method. It can test whether two samples are from the same population without too much samples and a pre-hypothesis about data distribution.

That is, it could give a conclusion objectively. Therefore, based on the W-M-W rank sum test method, we will determine whether to merge two clusters through testing whether the objects in these two clusters are coming from the same population. If from the same population, they can be merged into one cluster; otherwise, the two clusters still exist as two separate clusters.

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

Then we will take one dimensional objects as example and describe the method of deciding whether two sets need to be merged through W-M-W rank sum test. If the objects are multi dimensional data, we need to analyze in each dimension as the method; its basic idea is that objects in two clustersare from the same population indicate these two groups of data are from the same distributionin 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, m and n be their object number respectively. On the basis of W-M-W rank sum test method, we want to test whether two sets are from the same population by using sample information without the hypothesis of data distribution. Then we will conduct a hypothesis test with the sample data. If it is validated, the null hypothesis will be accepted; otherwise the null hypothesis will be rejected. Even though there is a hypothesis in this method, it is used to make a relatively objective conclusion based on data information, rather than to be a basis for subsequent analysis.

We will make the null hypothesis that 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 the similar distribution, without regarding to data symmetry. Then the problem about merging two sets could be transformed into the problem to be test: $H_0: \mu_1 = \mu_2 \leftrightarrow H_1: \mu_1 \neq \mu_2$. 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 having significant differences are from different distributions and could not 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 make 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 be 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 (6). W_{XY} is the number of samples from C_2' whose values are greater than the values from C_1', while W_{YX} is on the contrary.

$$\begin{aligned} W_{XY} &= mn + \tfrac{m(m+1)}{2} - W_X \\ W_{YX} &= mn + \tfrac{n(n+1)}{2} - W_Y \end{aligned} \tag{6}$$

If two sample sets have the same distribution, the ranks of the samples should be randomly mixed. While 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, and this method can be used to analyze whether two sets are from the same population without the sample distribution.

In addion, $Z = \frac{W_{XY} - mn/2}{\sqrt{mn(m+n+1)/12}} \to N(0,1)$. Then we can calculate the values of p with the corresponding m and n. This p-value is minimum significance level to reject null hypothesis according to the test statistics calculated by the samples[18,19]. Then for a given significance level α, we can obtain the analysis result of hypothesis testing through comparing p and α. If

$p > \alpha$, the null hypothesis is accepted, which indicates that there is no significant difference between data in these two clusters and they could be merged; While if $p \leq \alpha$, the null hypothesis is rejected, that is, data in the two clusters are more likely to come from different distributions and they could not be merged.

Figure 2 describes the specific steps of determining whether two one-dimensional sets have significant differences based on the above validation method.

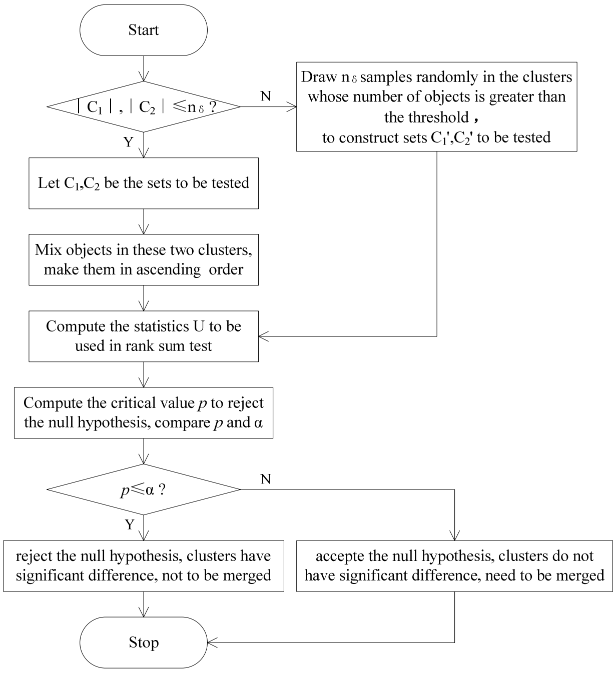
Start
∣C1∣, ∣C2∣ ≤nδ ?
N
Y
Draw nδ samples randomly in the clusters whose number of objects is greater than the threshold，to construct sets C1',C2' to be tested
Let C1,C2 be the sets to be tested
Mix objects in these two clusters, make them in ascending order
Compute the statistics U to be used in rank sum test
Compute the critical value p to reject the null hypothesis, compare p and α
p≤α ?
N
Y
reject the null hypothesis, clusters have significant difference, not to be merged
accepte the null hypothesis, clusters do not have significant difference, need to be merged
Stop

Fig.2. The process of measuring distances between clusters based on W-M-W rank sum test

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 could be drawn 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 to be analyzed, the random samples of objects will reflect the distribution characteristic to some extent. In addition, the test is based on nonparametric statistical method; it would take full advantage of sample data information, rather than analyze based on a hypothesis about data distribution. It tests whether two clusters are from the same distribution according to data itself. That is, it analyzes the similarity between clusters from the statistical test perspective. Its objectivity will ensure the accuracy of the measuring results. Although traditional distance metric are also based on data information and calculate distance between objects in the clusters. However their values of distances are not the final results of the measurement. They will be used to analyze whether two clusters are similar and are needed to be merged through the comparisons about the distance values. Therefore whether two clusters are similar is a relatively comparative result. Therefore, the distance measurement method proposed in this paper has certain advantages in the accuracy and efficiency.

Multi-dimensional data need to be analyzed on each dimension as above. Once there is a significant difference to be tested in one dimension, it indicates that data are from different populations on this dimension. It is difficult to illustrate the objects in two clusters have similar features, because they are already different in one dimension. Then it can be determined that objects in two clusters have significant differences; and there is no need to merge these clusters. The time complexity of this process is $O(dn_{\delta}^{2})$.

Taking C_1 and C_2 in **Figure 1** and another cluster C* as an example, we will illustrate the method determining whether two clusters need to be merged based on W-M-W rank sum test method. **Figure 3** describes its analysis process.

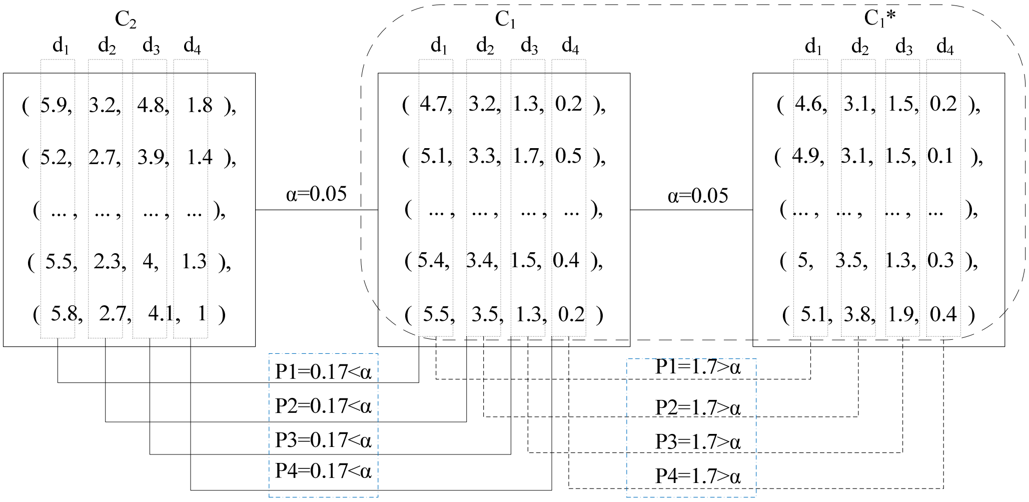

Fig. 3. Example of measuring distances between clusters based on W-M-W rank sum test

Clusters C_1 and C_2 in **Figure 3** are needed to be tested for each of four dimensions. Each value of p is less than the significance level α. It illustrates that these two clusters have a significant difference in all of four dimensions. Then it can be determined that C_1 and C_2 are from two different populations and they cannot be merged. While 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 betweenC_1 and C^*, and they can be merged into one cluster.

Obviously, our proposed method would obtain a more objective result than the traditional distance metrics, because it directly determines whether to merge two clusters based on the distribution characteristics of data, rather than based on the comparison of distance values. In fact, these data are from Iris dataset, and data in C_1 and C^* are from the same class, while data in C_1 and C_2 are from different categories. It illustrates the accuracy and validity of our method.

3 A data distribution feature oriented hierarchical clustering analysis method

Combined with the distance measurement method proposed above, this paper will propose a two-steps hierarchical clustering algorithm, so as to avoid assuming the number of clusters preliminarily, discover clusters of arbitrary shapes and improve clustering accuracy. The above distance measurement methods are the point of proposing such a clustering algorithm. In this hierarchical clustering algorithm, the distance metrics proposed in Section 2 are used to divide objects to the proper clusters and determine whether to merge clusters.

Firstly, in the first step, the idea of K-Means is used to generate a number of clusters as the initial clusters through dividing objects in the original data set. The generated number of clusters 'k' will be set to a larger value. Then the more similar objects would be divided into the same cluster. Then the idea of hierarchical clustering will be used to merge similar clusters in these initial clusters. The number of initial clusters is set to be a larger one, there will exist similar ones about distribution features among these initial clusters. Therefore in the second step we will merge the similar ones into one cluster so as to divide their objects into the same cluster. During this process, we will determine whether two clusters are similar and need to be merged. This operation will continue until all clusters are tested to have significant difference between each other, when data in different clusters are likely to come from different populations. Then the clustering process could stop. This two-steps hierarchical clustering algorithm is described as follows.

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: $C = \{C_1, \ldots, C_{K'}\}$, the clustering result;

```
Steps:
(1) Generate initial clusters
    1) choose $k$ objects from dataset $D$ to be the initial cluster centers, then obtain $C^t = \{C_1, \ldots, C_k\}$;
    2) repeat
    3) for $i := 1$ to $n$ do
    4) $C_\omega = ocd(x_i, C, \delta)$;
    5) divide object $x_i$ into cluster $C_\omega$;
    6) end for
    7) update the distribution features of $k$ clusters as equation (7):
       $DF_j = \{DF_{j1}, \ldots, DF_{jd}\} = \{< \mu_{j1}, \sigma_{j1}, m_{j1} >, \ldots, < \mu_{jd}, \sigma_{jd}, m_{jd} >\} \quad 1 \le j \le k \qquad (7)$
    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) $mb = ccd(C_i, C_j, n_\delta, \alpha)$;
    6)if ( $mb = 1$ ) then
    7) $C_i$, $C_j$ have significant difference, do not merge them;
    8) else if ( $mb = 0$ ) then
    9) $C_i$, $C_j$ 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.
```

The time complexity of obtaining initial cluster is $O(tkdn)$, where t is the iterations, k is the number of initial clusters, d is the dimensions of 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 t' is the iterations for the merge step. Therefore, this proposed two-steps hierarchical clustering algorithm based on nonparametric statistics has the time complexity of $O(tkdn + t'k^2dn_\delta{}^2)$, where $k, n_\delta \ll n$. Obviously, the proposed algorithm could be effective. The final number of clusters is generated based on the distribution features of data, rather than a pre-assumed value. In addition, the accuracy of distance metric proposed in Section 2 could ensure the accuracy of results generated by the proposed unsupervised clustering algorithm.

4 Experiments

We will choose three two-dimensional data sets and several UCI data sets, to verify the validity of the proposed clustering algorithm based on W-M-W rank sum test method. And we will compare with the following clustering algorithms: K-Means, DBSCAN, Birch, UPGMA [11], and Fast [12] about run time and accuracy of clustering results. The results will illustrate the effectiveness and practicality of our proposed algorithm.

For the data set having marked categories, we will use external indices Purity and Entropy [20] to clustering to evaluate the accuracy of clustering results. Let $C = \{C_1, \ldots, C_{K'}\}$ be the clustering result, and $P = \{P_1, \ldots, P_l\}$ represent 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:

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

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

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, n_i is the number of objects divided into *i*-th cluster. The higher the purity is, the more accurate the clustering result is; and the lower the entropy is, the more accurate the clustering result is. Ideally, *Entropy* = 0.0 and *Purity* = 1.0.

4.1 Two-dimensional datasets

We will choose three two-dimensional graphic data sets: Aggregation, Spiral and Flame to verify the proposed method could discover clusters of arbitrary shapes. These datasets contain the similar spatial data within the same category (clusters), not only simple spherical clusters. They could also be visualized. Therefore these datasets could be used to validate the capacity of discovering clusters of arbitrary shapes. **Figure 4** shows the visualized clustering results of three two-dimensional datasets obtained by our proposed algorithm NPSC. It can be seen that NPSC could identify the clusters of data, that is, it could discover clusters of arbitrary shape better.

This is mainly due to the distance measurement method used in the proposed algorithm. It determines the similarities between clusters on the basis of distribution features of data, rather than simply based on the traditional distance metrics. This method could merge similar clusters according to the characteristics of data based on the nonparametric statistical hypothesis test method without the hypothesis of data distributions. And it is used in the second step of the proposed clustering algorithm. A number of closely similar clusters will be generated in the first step of the clustering process.Then clusters having similar distribution features discovered based on our proposed distance measurement method. And the similar clusters will be merged into on clusters. These characteristics make the proposed clustering method more suitable todiscover non-spherical clusters.

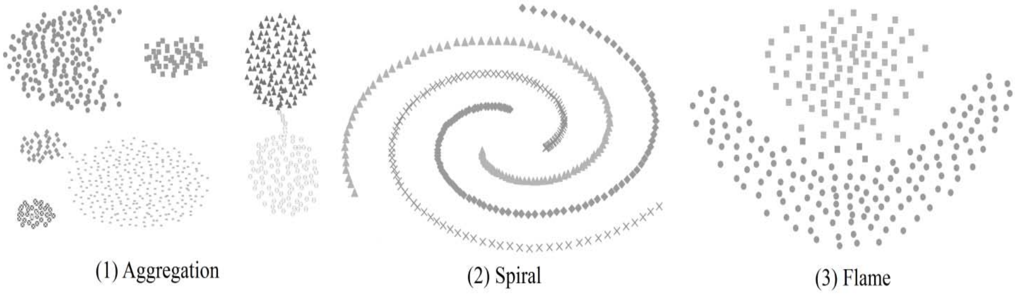

Fig. 4. Clustering results of two-dimensional datasets obtained by NPSC

4.2 UCI datasets

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

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

Figure 5 describes the comparison about the accuracy of clustering results based on evaluation indices Purity and Entropy. Obviously, NSPC obtains relatively greater Purity values and lower of Entropy values than other algorithms. It indicates that NSPC could obtain more accurate clustering results.

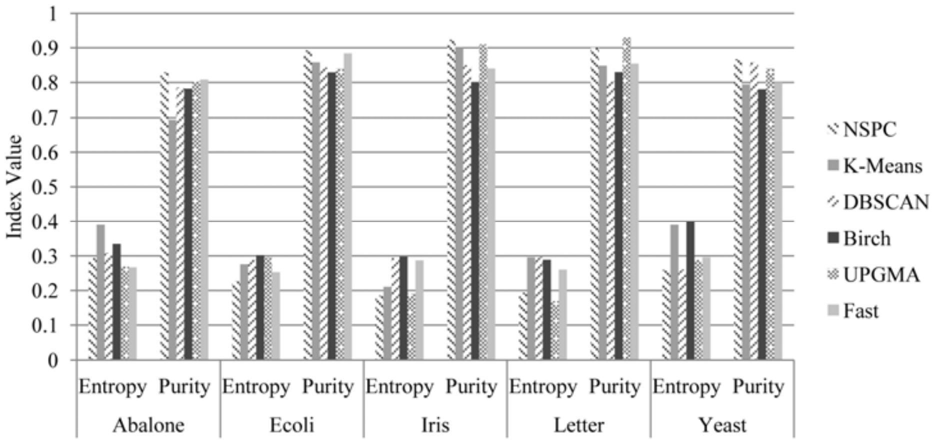

Fig. 5. The comparison about accuracy of clustering results on UCI datasets.

This is mainly due to the proposed distance measurement method, because 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 as K-Means or Birch. It also does not dependent on the parameter about neighborhood radius to determine the density of clusters as DBSCAN. NSPC could not entirely depend on the numerical distance measurement results as traditional methods do. It is based on nonparametric statistical hypothesis testing method, and determines whether clusters are similar according to the distribution features of data. UPGMA and Fast have improved the problems 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 analysis of similarity between clusters, these samples would reflect the distribution features of clusters to some extent. Fast uses the probability density function to obtain the distribution features of clusters. However this method needs to make assumptions about the distributions of data, and this assumption are more likely not to match the real data distributions. Therefore its results obtain lower Purity value and greater Entropy value compared with NSPC. It illustrates that our proposed distance metric helps to get more accurate clustering results.

Figure 6 shows the run time comparative results between these clustering algorithms on the UCI datasets. K-Means has high efficiency due to its linear time complexity. NSPC has the run time close to Birch which is also a hierarchical clustering method. And it has a relatively high efficiency compared to the other algorithms. Because UPGMA needs to obtain neighboring objects and then calculate clustering features. Fast needs to calculate probability density distribution functions for clusters. These would take some time.

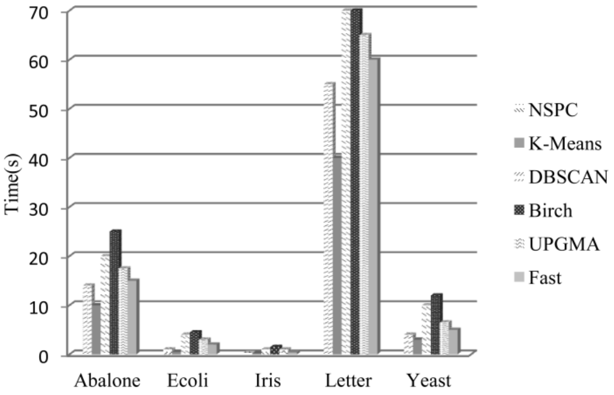

Fig. 6. The comparison about run time of clustering on UCI datasets.

Obviously, our proposed algorithm could not only obtain relatively accurate clustering results, but also have high efficiency. Firstly, it is due to the use of our proposed distance metric based on the distribution features of data during clustering. Secondly, it relies on our proposed two-steps clustering process. The former could help to ensure the accuracy and effectiveness of clustering. Because we use several descriptive statistics to represent distribution features of data in a cluster when measuring distances between an object and a cluster. It analyzes the distribution feature variations once the object is divided into a cluster. So it could get a more objective similarity result between the object and a cluster. We use W-M-W rank sum test method to measure distances between clusters. It could avoid the inaccuracy problems when determining whether to merge clusters according to a less objective comparison value in the traditional metrics. It could also ensure the efficiency of clustering process through not using all objects in the clusters. The latter uses a two-steps process, so that the number of generated clusters would not depend on a pre-assumed value. It could determine when the clustering process is terminated through the analysis of data distribution based on the nonparametric statistical hypothesis test.The final number of clusters does not be relative with the initial number set by the parameter.

Conclusion

This study aims at the purpose of distance measurement in unsupervised clustering: to generate new and accurate clusters. So a distance metric is proposed based on descriptive statistics and nonparametric statistical methods. In addition, a two-steps hierarchical clustering algorithm is also proposed. The distance measurement method based on nonparametric statistics could take full advantage of the distribution features of data. It could obtain clusters in a more straightforward and more objective way compared with the traditional distance metrics. The hierarchical clustering algorithm could avoid the pre-assumed initial number of clusters with its two-steps characteristics; the final number of clusters does not be relative with the initial number set by the input parameter. It could also discover clusters of arbitrary shapes and obtain more accurate results due to the distance metrics: it determines the similarities between clusters on the basis of distribution features of data. Therefore the proposed distance measurement method could provide a stronger support for unsupervised clustering analysis.

Acknowledgments. This work is sponsored by the Natural Science Foundation of Heilongjiang Province under grant number F2017015, and Training Program For Young Innovators in Heilongjiang General Institutes of Higher Education under grant number UNPYSCT-2017079.

References

[1]Verma M, Srivastava M, Chack N, et al.: A comparative study of various clustering algorithms in data mining.International Journal of Engineering Research and Applications,Vol. 2(No. 3), pp. 1379-1384(2012)

[2]Tsai C W, Lai C F, Chiang M C, et al.: Data mining for Internet of Things: A survey. IEEE Communications Surveys and Tutorials,Vol. 16(No. 1), pp. 77-97(2014)

[3]Fahad A, Alshatri N,Tari Z, et al.: A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE transactions on emerging topics in computing,Vol. 2(No. 3), pp. 267-279(2014)

[4]Ferrari D G, De Castro L N.: Clustering algorithm selection by meta-learning systems: A new distance-based problem characterization and ranking combination methods. Information Sciences,No. 3, pp. 181-194(2015)

[5]Patidar A K, Agrawal J, Mishra N.: Analysis of different similarity measure functions and their impacts on shared nearest neighbor clustering approach. International Journal of Computer Applications,Vol. 40(No. 16), pp. 1-5 (2012)

[6]Senoussaoui M, Kenny P, Stafylakis T, et al.: A study of the cosine distance-based mean shift for telephone speech diarization.IEEE/ACM Transactions on Audio, Speech and Language Processing,Vol. 22(No. 1), pp. 217-227 (2014)

[7]Sarle W S.: Finding groups in data: An introduction to cluster analysis.Journal of the American Statistical Association,Vol. 86(No. 415), pp.830-833.(1991)

[8]AnkerstM, BreunigM, KriegelHP, et al.: OPTICS: ordering points to identify the clustering structure.ACM SIGMOD Record,No. 28, pp. 49-60.(1999)

[9]Lovely Sharma P, Ramya K A.: Review on density based clustering algorithms for very large datasets[J]. Int J Emerg Technol Adv Eng.Vol. 3(No. 12), pp. 398-403 (2013)

[10]Zhang T, Ramakrishnan F, Livny M, et al.: BIRCH: An efficient data clustering method for very large database.ACM SIGMOD Record,No. 415, pp. 103-114. (1996)

[11]Bouguettaya A, Yu Q, Liu X, et al.: Efficient agglomerative hierarchical clustering. Expert Systems with Applications,Vol. 42(No.5), pp. 2785-2797(2015)

[12]Song Q, Ni J, Wang G.: A fast clustering-based feature subset selection algorithm for high-dimensional data. IEEE transactions on knowledge and data engineering,Vol. 25(No. 1), pp. 1-14(2013)

[13]Boutsidis C, Magdon-Ismail M. Deterministic feature selection for k-means clustering. IEEE Transactions on Information Theory, Vol. 59(No. 9), pp. 6099-6110(2013)

[14]Gautier E, Kitamura Y. Nonparametric estimation in random coefficients binary choice models. Econometrica,Vol. 81(No. 2), pp. 581-607(2013)

[15]Wansouwé W E, Kokonendji C C, Kolyang D T, et al.: Nonparametric estimation for probability mass function with Disake. ARIMA Journal,No. 415, pp. 1-23(2015)

[16]Zimmerman D W, Zumbo B D.: Mann-Whitney Test and Student t Test Under Simple Bounded Transformations. The Journal of General Psychology,Vol. 117(No. 4), pp. 425-436(2017)

[17]Conde M A, García F, Rodríguez-Conde M J, et al.: Perceived openness of Learning Management Systems by students and teachers in education and technology courses. Computers in Human Behavior,Vol. 31, pp. 517-526 (2014)

[18]Anderson D R, Burnham K P, Thompson W L, et al.: Null hypothesis testing: problems, prevalence, and an alternative. The journal of wildlife management, pp. 912-923(2000)

[19]Schneider J W.: 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(2015)

[20]Kou G, Peng Y, Wang G, et al.: Evaluation of clustering algorithms for financial risk analysis using MCDM methods. Information Sciences,No.275, pp. 1-12(2014)