A study on Two-Stage Mixed Attribute Data Clustering Based on Density Peaks

Shihua Liu, Hao Zhang, and Xianghua Liu
Department of Information Technology, Wenzhou Polytechnic, China

Abstract: *A Two-stage clustering framework and a clustering algorithm for mixed attribute data based on density peaks and Goodall distance are proposed. Firstly, the subset of numerical attributes of the dataset is clustered, and then the result is mapped into one-dimensional categorical attribute and added to the subset of categorical attribute data. Finally, the new dataset is clustered by the density peaks clustering algorithm to obtain the final result. Experiments on three commonly used UCI datasets show that this algorithm can effectively realize mixed attribute clustering and produce better clustering results than the traditional K-prototypes algorithm do. The clustering accuracy on the Acute, Heart and Credit datasets are 17%, 24%, and 21% higher on average than that of the K-prototypes, respectively.*

Keywords: *Mixed data clustering, density peaks, k-prototypes algorithm, validity index.*

Received July 4, 2019; accepted September 27, 2020
https://doi.org/10.34028/iajit/18/5/2

1. Introduction

Cluster analysis is a research hotspot in the field of data mining and machine learning, it is widely used in many fields such as text analysis [24]. In the era of big data, there is a large amount of data generated every day. Most of these data have many attributes such as numerical and categorical values. Therefore, clustering research on mixed attribute data has become one of the areas of interest for researchers.

In 2014, Rodriguez and Laio [22] published the paper "Clustering by fast search and find of density peaks" in Science hereinafter referred to as the Density Peaks Clustering (DPC) algorithm. The method has high efficiency and promising results and requires fewer parameters. It can discover the number of clusters, cluster data in different shapes, and automatically identify outliers. However, few studies directly apply it to mixed attribute data clustering and categorical attribute data clustering.

This paper first analyses the state of research regarding mixed attribute data clustering, and then proposes a Two-stage clustering framework for mixed attributes based on the analysis of the philosophy of clustering: In the first stage, the subset of numerical attributes is clustered, the result is added to the categorical subset as a one-dimensional categorical attribute. In the second stage, the categorical attribute clustering algorithm is used to cluster the new subset to obtain the final result. Based on this clustering framework, a Two-stage clustering algorithm called K-Means and Density Peaks based Clustering (KMDPC) is proposed. The K-means algorithm guided by the Sil (Silhouette) index and the DPC algorithm based on the improved Goodall similarity [7] are used in the corresponding stage. Experiments on UCI real data sets show that the clustering framework and the corresponding algorithm are effective and can get better results.

The structure of this paper is as follows: section 2 introduces the related works of mixed attribute data clustering along with the key points of Density Peaks Clustering and Goodall similarity measure. Section 3 describes in detail the Two-stage clustering framework and its KMDPC implementation. Section 4 shows the simulation results. And section 5 summarizes the contribution of the paper and gives the conclusion.

2. The Related Works

There are many solutions to mixed attribute clustering, such as attribute conversion method, clustering ensemble method, prototype-based methods, hierarchical clustering method, density clustering method, and so on.

The attribute conversion method is to convert different types of attribute data into a certain type, and then use the corresponding clustering method for analysis. Its typical representative is the SpectralCAT algorithm proposed by David and Averbuch [2].

The clustering ensemble method was first proposed by Strehl and Ghosh [25], and then became one of the mainstream methods for mixed attribute clustering. Zhao *et al*. [32] proposed a mixed attribute clustering algorithm Cluster Ensemble-based Mixed attribute Clustering (CEMC) based on clustering ensemble. He *et al*. [9] proposed a mixed attribute clustering algorithm Cluster Ensemble Based Mixed Data Clustering (CEBMDC) based on clustering ensemble and the Squeezer algorithm [10]. The algorithm uses the

Squeezer algorithm for categorical attribute clustering and the final clustering ensemble. Li *et al*. [17] proposed a clustering ensemble based mixed attribute data incremental clustering algorithm to avoid the instability and randomness problems caused by a single cluster. Qian and Huang [21] proposed a mixed data clustering algorithm based on dimension frequency difference and strongly connected fusion.

The K-prototypes algorithm proposed by Huang in [12] is a typical prototype-based method. It is similar to the K-means algorithm. Its principle is simple but shows high efficiency and is widely used in mixed attribute clustering. However, it is also sensitive to initial point selection and the number of clusters must be specified beforehand. Moreover, the algorithm clustering results are sensitive to the weight coefficient γ. Therefore, many researchers have improved upon this, such as the global K-prototypes algorithm proposed by Bai *et al*. [1] and the weighted fuzzy K-prototypes algorithm proposed by Ji [13]. Jia and Song [14] proposed a Weighted K-prototype Clustering Algorithm (WKPCA) based on the hybrid dissimilarity coefficient. Sun *et al*. [27] proposed a K-prototypes clustering algorithm based on density optimization, which can adaptively optimize the setting of the number of clusters and the initial clustering according to the distribution density of data objects.

A typical hierarchical clustering method is the SBAC algorithm proposed by Li and Biswas [16], which is an agglomerative hierarchical clustering algorithm based on Goodall similarity. This method is highly effective but with high computational complexity.

Huang and Li [11] proposed a mixed attribute data clustering algorithm RDBC_M based on relative density. The algorithm uses density-based clustering to perform density clustering. Based on this, the incremental clustering algorithm IncRDBC_M is proposed. Rodriguez and Laio [22] proposed the DPC algorithm. This algorithm can be classified as a density clustering algorithm. As long as the construction of the distance matrix is possible, it can be applied to data clustering of any attribute type. Liu *et al*. [18] proposed a new distance measurement method for mixed attributes to construct the distance matrix and used the DPC algorithm to calculate the mixed attribute distance. The results validated the feasibility of DPC for clustering of mixed attribute data.

Stimulated by the density peak clustering algorithm, Xie and Qu [29] proposed two new K-medoids clustering algorithms with optimized initial seeds by density peaks. Fang *et al*. [5] proposed an adaptive Core Fusion-Based Density Peak Clustering (CFDPC) for detecting clusters of any shape and density adaptively. Xu *et al*. [31] proposed a robust density peaks clustering algorithm with density-sensitive similarity (RDPC-DSS) to find accurate cluster centers on the manifold datasets. Sun *et al*. [26] presented an adaptive DPC algorithm with Fisher linear discriminant for the clustering of complex datasets, called ADPC-FLD. Du *et al*. [3] presented a novel clustering algorithm for mixed data, called DPC-MD, which improved DPC by using a new similarity criterion to deal with three types of data: numerical, categorical, and mixed data.

2.1. Density Peaks Clustering Algorithm

The DPC algorithm is based on two basic assumptions: the cluster center has a higher local density and is surrounded by points with lower local density, and the relative distance between the cluster center and the point with a higher density is larger. Therefore, the DPC algorithm constructs a decision graph to calculate the cluster center of a data set by calculating a local density ρ_i and a relative distance δ_i. The remaining data points in the data set will be assigned to the cluster to which their nearest cluster center belongs.

Let $X = \{X_1, X_2, \ldots X_n\}$ be a data set composed of N data points to be clustered. D_{ij} = dist (X_i, X_j) is defined as the distance between the data points X_i and X_j. The DPC algorithm defines a truncated distance d_c(cutoof distance). From this, the local density ρ_i and distance δ_i of each data point are defined. When x < 0, $\chi(x)$=1, otherwise 0.

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \tag{1}$$

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i} (d_{ij}) & \rho_i < \max_k(\rho_k) \\ \max_j (d_{ij}) & \rho_i = \max_k(\rho_k) \end{cases} \tag{2}$$

The distance δ_i is defined as follows: when the local density is not the maximum density, the distance corresponding to the data point X_i is the minimum distance from the point to all points with greater density. Otherwise, the maximum distance from the point to all other points is taken as the distance value.

When there are fewer data points in the data set, it is not ideal to use Equation (1) to calculate the local density ρ_i. Therefore, in [22], a Gaussian kernel function is given for data sets with fewer data points, as shown in Equation (3):

$$\rho_i = \sum\nolimits_{j \neq i} \exp(-(\frac{d_{ij}}{d_c})^2) \tag{3}$$

Based on the decision graph constructed by the local density and distance of each data point, the number and center points of clusters can be found and selected explicitly. After the cluster centers are determined, the rest of the data points are scanned one time and assigned to the same cluster of the nearest neighbors of all the points with higher local density, so that the clustering can be completed quickly.

2.2. Goodall Similarity Measure

The Goodall similarity measure can be used to calculate the distance between mixed attribute data points, or to calculate the distance of numerical attributes or categorical attributes separately. The basic idea is to

assign greater weight to values that do not often appear in attributes: For categorical attributes, an unusual eigenvalue matching has greater similarity than a common eigenvalue matching. That is to say, for the same attributes with equal values, if the matched attributes belong to the uncommon attributes in the range, then their equal contribution to the similarity of two data points will be greater. For numerical attributes, the uncommon degree of eigenvalue pairs is measured by a distance function between a pair of values and the density of data points contained between the two values. Accordingly, they give the corresponding similarity scoring models and extend them to a comprehensive similarity measure function by using chi-square transformation.

1. Distance calculation for categorical attributes

For two data points (X_i, X_j) in data set X, let their *k-th* dimension be categorical attributes. The value of the *k-th* dimension of data point X_i is represented by V_{ik}, and the distance of the data point (X_i, X_j) in the *k-th* dimension by D_{ijk}. Then, when $V_{ik} \neq V_{jk}$, $D_{ijk} = 1$. When $V_{ik} = V_{jk}$, $0 < D_{ijk} < 1$, and the calculation of the distance D_{ijk} is as below:

Firstly, in the range D(V_k) of the *k-th* dimension of the data set, the probability f_{ik} (i.e., the number of occurrences) of each $V_{ik} \in$ D*(V_k)* is calculated. Then, for a specific attribute value V_{jk}, a More Similar Feature Value Set (MSFVS) is constructed according to the size of each f_{ik} in the range, denoted MSFV(V_{jk}). The set contains all attributes with frequencies not greater than f_{jk}. The contribution to the distance d_{lk} of any pair of values *(V_{lk}, Vlk)* in the set is calculated by the following formula. Here, n is the number of data points in the data set X, and f_{lk} is the frequency (counted times) of the attribute value V_{lk} appearing in the whole range.

$$d_{lk} = p_{lk}^2 = \frac{f_{lk}(f_{lk}-1)}{n(n-1)} \tag{4}$$

Finally, the distance between two data points (X_i, X_j) in the *k-th* dimension can be calculated by the following Equation:

$$D_{ijk} = \sum_{l \in MSFV(V_{jk})} d_{lk} \tag{5}$$

2. Calculation of numerical attribute distance

For two data points (X_i, X_j) in data set X, let their m-th dimension be a numerical attribute, the m-th dimension of the data point X_i be represented by V_{im}, and the distance of the data point (X_i, X_j) in the m-th dimension be represented by D_{ijm}.

Firstly, the frequency (f_{lm}) of each value $V_{lm} \in D(V_m)$ in the range $D(V_m)$ of the m-th dimension corresponding to the data set is calculated, then, a More Similar Feature Segment Set (MSFSS) is constructed according to the gap and frequency of the two values *(V_{im}, V_{jm})* which need to be calculated. It is abbreviated to MSFS *(V_{im}, V_{jm})*. The value contained in the set corresponds to the similarity of the value pair that should be satisfied not less than the similarity of the original value pair *(V_{im}, V_{jm})*, or the gap of the value pair is not greater than $|V_{im}-V_{jm}|$. The contribution to the distance d_{ijm} of any pair of values *(V_{im}, V_{jm})* in the set is calculated by the following formula. Here n is the number of data points in the data set X.

$$d_{ijm} = \begin{cases} 2p_{im}p_{jm} = \dfrac{2f_{im}f_{jm}}{n(n-1)} & p_{im} \neq p_{jm} \\ p_{im}p_{jm} = \dfrac{f_{im}(f_{im}-1)}{n(n-1)} & p_{im} = p_{jm} \end{cases} \tag{6}$$

Finally, the distance between two data points (X_i, X_j) in the m-dimension can be calculated using the following Equation:

$$D_{ijm} = \sum_{i,j \in MSFS(V_{im},V_{jm})} d_{ijm} \tag{7}$$

3. Computing the aggregate distance of data points

After calculating the distances of the data points (X_i, X_j) in data set X by using the two methods given above, X^2 transformation is performed and then the distances of two data points are calculated by summing up the additivity of degree of freedom of X^2 distribution.

For numerical attributes, Fisher X^2 is used to convert them as follows, where t_c is the number of numerical attributes in the data set, and X_c is the X^2distribution with the degree of freedom t_c:

$$(\chi_c)_{ij}^2 = -2\sum_{k=1}^{t_c} \ln(D_{ijm}) \tag{8}$$

For categorical attributes, Lancaster means X^2 transformation was implemented as follows, where t_d is the number of categorical attributes in the data set, X_d is the X^2 distribution subject to the degree of freedom t_d, and D_{ijk}' is the distance value next only to D_{ijk} in the distance of categorical attributes:

$$(\chi_d)_{ij}^2 = 2\sum_{k=1}^{t_d}\left(1 - \frac{D_{ijk}\ln(D_{ijk}) - D_{ijk}'\ln(D_{ijk}')}{D_{ijk} - D_{ijk}'}\right) \tag{9}$$

Thus, the distance of data points (X_i, X_j) can be calculated by the following Equation:

$$D_{gd}(X_i, X_j) = e^{-\frac{\chi_{ij}^2}{2}\sum_{k=0}^{t_d+t_c-1}\frac{(\frac{1}{2}\chi_{ij}^2)^k}{k!}} \tag{10}$$

Here, the chi-square distance of mixed attribute datapoints is the sum of the chi-square distances of numerical attribute and categorical attribute parts.

$$\chi_{ij}^2 = (\chi_d)_{ij}^2 + (\chi_c)_{ij}^2 \tag{11}$$

3. Two-Stage Clustering Framework

3.1. The concept and Essence of Clustering

Clustering refers to the process of dividing a series of data objects into several subsets according to certain similarity metrics. Data objects clustered in the same class should be similar to each other and not similar to objects not in the same class [8]. Mathematically, a data

set consists of n data objects, $X = \{X_1, X_2, X_3, \ldots X_n\}$, where each data object X_i ($I = \{1, \ldots, n\}$) is described by d attributes, namely $X_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T$, x_{ij} denotes the j-th attribute of the i-th data point, where $j = \{1, \ldots, d\}$. d is also known as the dimension of a data set.

Clustering analysis of a data set X aims to divide the data set into multiple subsets $X = \{C_1, C_2, \ldots, C_k\}$ (k is the number of clusters), according to the similarity between the d-dimensional attributes of each data point and the d-dimensional attributes of other data points, which makes the similarity of data points in the same subset higher, while the similarity of data points in different subsets is relatively low.

The essence of cluster analysis is to map the d-dimensional attributes of each data object in the data set to the one-dimensional categorical attribute, which is the classification result label [6].

$$\begin{Bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \cdots & & \cdots & \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{Bmatrix} \xrightarrow{f:\text{clustering}} \begin{Bmatrix} c_1 \\ c_2 \\ \cdots \\ c_n \end{Bmatrix} \quad (12)$$

As shown in formula (12), where n is the number of data points of the data set, d is the number of attribute dimensions, $C = \{c_1, c_2, \ldots, c_n\}^T$ is the clustering result, such as $c_i = c_j$, indicating that the i-th and j-th data objects have higher similarity and are classified into the same class.

3.2. Two-stage Clustering Framework

From the above analysis, it can be seen that each dimension attribute or a set of partial attributes of the data object can provide a corresponding similarity basis for the clustering of the data, and the result of the clustering is a category label which can also be used as a categorical attribute. Therefore, clustering for mixed attribute data sets can be performed using the following Two-stage clustering framework.

3.2.1. Frame Structure and Process

As shown in Figure 1 below, there are n mixed attribute data sets X of d-dimensional attribute data points, including p-dimensional numerical attributes and q-dimensional categorical attributes (q=d-p).

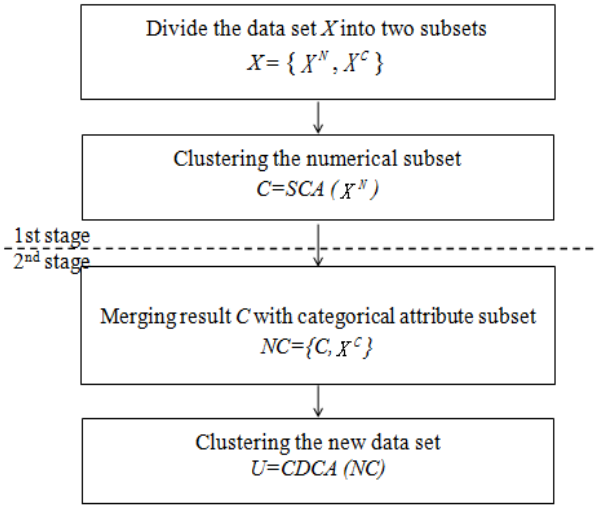


Figure 1. Two-stage clustering framework.

The first stage is as follows. Divide the d-dimensional attributes of the data set X into two subsets: only numerical attributes and only categorical attributes $X = \{X^N, X^C\}$, and each subset can be composed of one or more attributes of the original data set. A specific clustering algorithm is used for clustering the subset of numerical attributes. The clustering result can be represented by an n-dimensional column vector, $C=SCA(X^N) = \{c_1, c_2, \ldots, c_n\}^T$. The second stage is to combine the clustering result C of the first stage with the original subset of categorical attributes; a new data set $NC=\{C, X^C\}$ is constructed with q+1 dimension categorical attributes. Then the clustering result $U=CDCA(NC)$ is obtained by using the clustering algorithm for categorical attributes.

3.2.2. Key Issues to be Addressed

In the process of realizing the clustering framework, the following problems need to be solved:

1. Selection of clustering algorithm. For different subsets, a clustering algorithm can be selected according to the characteristics or shape of the data distribution. For example, K-means, Fuzzy C-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Expectation-Maximum (EM) etc., to deal with numerical datasets, and k-modes, Squeezer etc. to deal with categorical datasets. Determination of the number of clusters k. There are many methods to automatically determine the number of clusters, but in the Two-stage clustering framework, the selection of methods need to be compared experimentally.
2. Attributes' weight determination. Different attribute's weights lead to different clustering results, how to determine the weight of attributes need to be well studied. The simple method is to treat all members equally, and information entropy can also be used.

3.3. Implementation of the Two-Stage Clustering Algorithm

To verify the effectiveness of the framework described above, a simple implementation of Two-stage clustering algorithm is proposed. The algorithm uses the classical K-means algorithm to cluster the subset of numerical attributes and uses the Sil index to automatically determine the number of clusters [28]. In the second stage, the improved Goodall distance is used to calculate the distance between data points, and then the DPC algorithm is used for clustering. Therefore, this algorithm is abbreviated to KMDPC.

In the KMDPC algorithm, the numerical attributes are normalized and clustered using K-means guided by Sil index. The clustering result in the form of a categorical attribute is added to the original categorical attribute subset to form a new categorical attribute dataset. The new dataset is then clustered by DPC algorithm. The algorithm is described as follows:

Algorithm 1: KMDPC clustering algorithm

Input: mixed attribute dataset $X = \{X^N, X^C\}$ (X^N is the subset of numeric attributes, X^C is the subset of categorical attributes)
Output: Clustering result label vector U
Algorithm steps:
Step 1: Use the K-means algorithm guided by Sil index to perform cluster analysis on the numerical attribute subset X^N to determine the optimal number of clusters k of the numerical attribute subset.
Step 2: Use the K-means algorithm to cluster the numerical attribute subset X^N: C=K-Means(X^N, k);
Step 3: Combine the result C into the categorical attribute subset X^C to obtain a new data set NC={C，X^C};
Step 4: Calculate the distance between data points by using the improved Goodall distance metric algorithm for the new data set NC, and then the DPC algorithm is used to get the final clustering result: U=DPC(NC);

The algorithm uses the Goodall distance to represent the distance between data points of categorical attributes, the numerical attributes in Equation (11) can be partially removed, i.e.,

$$\chi_{ij}^2 = (\chi_d)_{ij}^2 \tag{13}$$

Then let t_c=0 in Equation (10). For the problem of determining the number of clusters *k*, the K-means algorithm guided by the Sil index is used in the numerical attribute clustering stage. In the final DPC clustering stage, the decision graph is used to determine. The selection of the cluster center will be discussed in the experimental analysis section. The weight of all categorical attributes is treated as the same for the sake of simplicity.

3.4. Algorithm Efficiency Analysis

It can be seen from the above that the time complexity of the KMDPC algorithm mainly comes from three parts. The first part is the clustering of the subset of numerical attributes. The execution cost of the K-means algorithm is O(*tkn*), where *t* is the number of iterations of the algorithm, which is usually very small, i.e., $t<<n$. *k* is the number of clusters determined by the K-means algorithm guided by the Sil index and *n* is the number of data points in the data set, in which the determination of *k* needs the K-means algorithm to be run *M* times, where *M* is the maximum number of possible clusters that is set manually. The second part is the Goodall distance calculation, the complexity of which is $O(n^2 logn)$. The third part is the clustering overhead of the DPC algorithm, whose complexity is O(*nlogn*).

4. Experimental Analysis

To evaluate the validity of the KMDPC algorithm and the Two-stage clustering framework, three commonly used UCI datasets with mixed attributes were used in the clustering analysis experiments, which are Acute Inflammation, Heart Disease and Credit Approval datasets.

4.1. Dataset Introduction

The three mixed attribute datasets in UCI were originally used for classification research; therefore, each data set has a classification label attribute which provides convenience for calculating the classification accuracy. The basic information of the three datasets is shown in Table 1 below.

Table 1. Brief information of the UCI datasets

Abbr.	Data Set	Instances	Numerical attribute	Categorical attribute	Label
Acute	Acute Inflammations	120	1	5	2
Credit	Credit Approval	653	6	9	1
Heart	Heart Disease	270	6	7	1

In the table, the first column is the abbreviations of the three datasets, the second column is the full name of the datasets, the third column is the number of instances contained in each dataset, and the fourth, fifth and sixth columns represent the numbers of numerical attributes, categorical attributes and class labels contained in the datasets, respectively. The Credit Approval data set contains information on 690 users who applied for a bank credit card. There are 37 records contain missing data, so the number of instances without missing data is 653. These three mixed attribute datasets are widely used in the research of mixed attribute clustering.

4.2. Experimental Setup and Result Analysis

Almost every new clustering algorithm for mixed attribute data takes K-Prototypes algorithm as the comparison benchmark. In this paper, the KMDPC algorithm and the K-prototypes algorithm are used to cluster the three aforementioned data sets.

Romano *et al*. [23] studied the applicable scenarios of external validity indicators and found that Adjusted Rand Index (ARI) was more suitable for data sets with

large scale and uniform cluster size distribution, and Normalized Mutual Information (NMI) was more suitable for unbalanced data sets with small clusters. Therefore, to verify the effectiveness of the algorithm, in addition to calculating the Clustering Accuracy (CA) like all clustering studies, this paper also uses ARI, NMI and Rand Index (RI) to compare the performance of the two algorithms on three different data sets. In order to study their stability, we also compare the worst, best and average clustering effects of the two algorithms and analyze their turbulence. In the aspect of determining the number of clusters automatically, we compare the effects of several internal validity indicators, and give the reason for using Sil index. Xu *et al*. [30] used DEPSO algorithm to compare the performance of eight famous and widely used validity indicators and reached the experimental conclusion that Sil index was better than other indicators.

4.2.1. Feasibility Experimental Results

According to the research in [22], the parameters for density calculation in the DPC algorithm in the second stage of KMDPC clustering are taken as p=1.5%. According to the research in [12], the important parameter γ of the K-prototypes algorithm is $1/2\sigma$ (σ represents the average standard deviation of the numerical attributes). Since the dataset provides real class labels, we can use CA, NMI, RI, ARI to indicate the clustering results. The bigger the values of these indicators are, the better the clustering result is [20, 23].

Since there are two decision attributes in the Acute dataset, and each attribute is binary, it is converted into a four-category decision attribute to process and calculate the clustering accuracy.

The experimentation is implemented using matlab R2015a. The K-prototypes algorithm runs 100 times to get the average value, while KMDPC runs 20 times to get the average value because the number of clusters must be chosen manually. The clustering results are shown in Table 2 below. The first and second rows in the table respectively represent the CA results of K-Prototypes and KMDPC algorithm on the three data sets, the bold font indicates the best result. The NMI results of the two clustering algorithms are shown in the third and fourth rows. The RI and ARI results are presented in the next four rows.

Table 2. Results of K-prototypes and KMDPC.

Clustering results on data sets		Acute	Heart	Credit
CA	k-prototypes	0.7121	0.5926	0.5528
	KMDPC	**0.8333**	**0.7346**	**0.6670**
NMI	k-prototypes	0.6071	0.0202	0.0303
	KMDPC	**0.7602**	**0.2077**	**0.0967**
RI	k-prototypes	0.8063	0.5153	0.5048
	KMDPC	**0.8839**	**0.6086**	**0.5558**
ARI	k-prototypes	0.5158	0.0303	0.0026
	KMDPC	**0.7074**	**0.2171**	**0.1114**

Accordingly, we can see that the performance of KMDPC is much better than the traditional K-prototypes algorithm in terms of clustering accuracy, standardized mutual information and other external indicators. In the Acute dataset, the accuracy of KMDPC is 17% higher than that of K-prototypes, the NMI, RI, ARI of KMDPC are 25.2%, 9.6%, and 37.1% higher than those of K-prototypes. In the Heart and Credit datasets, the results are similar as in Acute, the four indicators of KMDPC are much higher than those of K-prototypes.

To analyze the stability of the two algorithms, the optimal, worst and average values of the results are listed and compared as follows.

Table 3. Clustering results on acute dataset.

Clustering results of different algorithms		worst	average	optimal
CA	k-prototypes	0.5083	0.7121	0.7833
	KMDPC	**0.6750**	**0.8333**	**0.8417**
NMI	k-prototypes	0.4582	0.6071	0.6757
	KMDPC	**0.6051**	**0.7602**	**0.7684**
RI	k-prototypes	0.7136	0.8063	0.8468
	KMDPC	**0.7775**	**0.8839**	**0.8895**
ARI	k-prototypes	0.3608	0.5158	0.6036
	KMDPC	**0.4392**	**0.7074**	**0.7215**

As shown in Table 3, the optimal, worst and average values of KMDPC on Acute are all bigger than those of K-prototypes. The CA value of KMDPC changes from 0.6750 to 0.8417, while the value of K-prototypes is [0.5083, 0.7833], The later range of variation is larger than that of former. The NMI, RI and ARI values are similar as CA. That is to say, the KMDPC is more stable than K-Prototypes algorithm.

In Heart and Credit datasets as illustrated in Tables 4 and 5, the range of variation of KMDPC is larger than that of K-prototypes, but its own range of variation is relatively small, for example, the CA on Credit dataset is between -4.3% to +1.9%, which means the KMDPC algorithm is more stable. In addition, the worst values of all four indicators of KMDPC are larger than the optimal values of K-prototypes.

Table 4. Clustering results on heart dataset.

Clustering results of different algorithms		worst	average	optimal
CA	k-prototypes	0.5889	0.5926	0.5926
	KMDPC	**0.7333**	**0.7346**	**0.7444**
NMI	k-prototypes	0.0182	0.0202	0.0204
	KMDPC	**0.2051**	**0.2077**	**0.2238**
RI	k-prototypes	0.5140	0.5153	0.5154
	KMDPC	**0.6074**	**0.6086**	**0.6186**
ARI	k-prototypes	0.0276	0.0303	0.0303
	KMDPC	**0.2146**	**0.2171**	**0.2359**

Table 5. Clustering results on credit dataset.

Clustering results of different algorithms		worst	average	optimal
CA	k-prototypes	0.5513	0.5528	0.5528
	KMDPC	**0.6386**	**0.6670**	**0.6799**
NMI	k-prototypes	0.0257	0.0303	0.0304
	KMDPC	**0.0713**	**0.0967**	**0.1084**
RI	k-prototypes	0.5045	0.5048	0.5048

ARI	**KMDPC**	**0.5377**	**0.5558**	**0.5641**
	k-prototypes	0.0019	0.0026	0.0026
	KMDPC	**0.0751**	**0.1114**	**0.1281**

The results in Acute dataset are also presented as a histogram in Figure 2 below, with positive and negative errors as error lines. The histogram is based on the average value, where the worst value is marked as negative error, and the best value is marked as positive error. The results for the Heart and Credit datasets are also directly represented by the histogram with errors as shown in Figures 3 and 4. As can be seen from these figures, the KMDPC algorithm is superior to the k-prototypes algorithm on all datasets, in terms of the worst, average, and optimal values.

As can be seen from Figure 2, for the Acute dataset, the average of the KMDPC algorithm is near to the optimum, which also indicates that the KMDPC algorithm more likely to get the best results. From Figures 3 and 4, it can be seen that the errors of the two algorithms on the Heart and Credit data sets are relatively small, and the algorithm can obtain more stable clustering results. This is related to the fact that there are a large number of numerical attributes for the two datasets and the cluster number is only 2. It can also be seen from the two figures that the worst value obtained by the KMDPC algorithm is better than the optimal value of the K-prototypes algorithm, which also highlights the superiority of it.

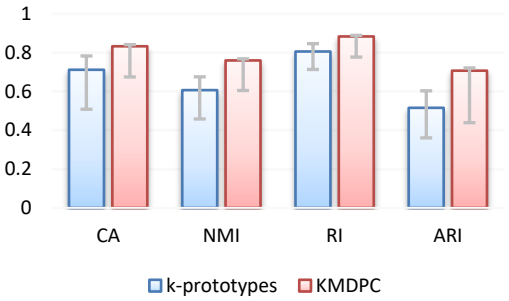


Figure 2. Clustering results for the acute dataset.

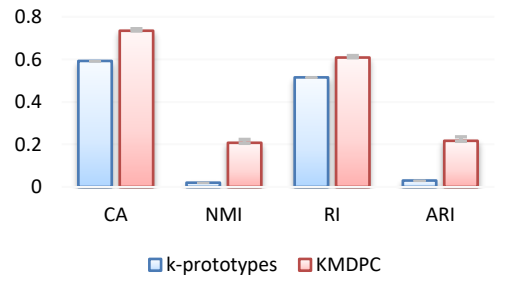


Figure 3. Clustering results for the heart dataset.

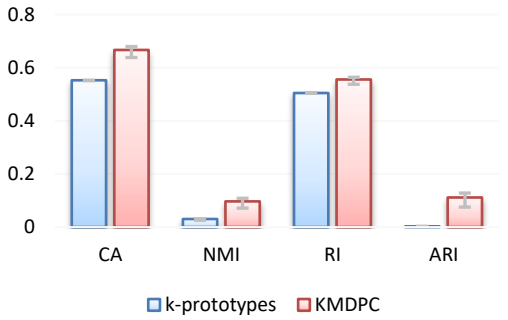


Figure 4. Clustering results for the Credit dataset.

4.2.2. Experimental Analysis of the Determination of the Number of Clusters

Since the K-means algorithm is sensitive to the initial center point, it will exert a certain impact on the stability of the clustering result of the KMDPC algorithm. To determine the number of clusters k in the first stage and its influence on the clustering results, this paper first uses the Calinski-Harabasz Index (CHI), Davies-Bouldin Index (DBI) and Sil index to analyze the k value. Then, the k value is set to between 2 to $\sqrt{n}$ for clustering, and the influence of different k values on the clustering results of the KMDPC algorithm is analyzed via clustering accuracy. Finally, using the decision graph of the DPC algorithm, the final number of clusters can be determined.

1. Indicator guidance method to determine the k value

For each of the three data sets, three indicators are used to determine the best k value. The Acute data set has 120 data points, and k takes values from 2 to 11, while the Heart data set has 270 data points, and k takes values from 2 to 16. The Credit dataset has 653 valid data points, and k takes values from 2 to 26. The k value results determined by the indicator method are shown in Table 6 below.

Table 6. Best k according to the validate indexes.

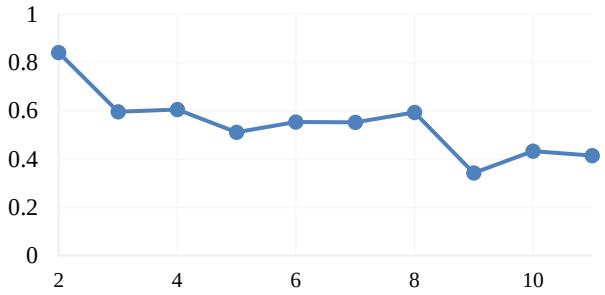
Validity indicator determines k value	Acute	Heart	Credit
CHI	11	2	4
DBI	2	2	uncertain
Sil	2	2	2

The above experimental results show that different validity indicators may lead to different optimal k values. Some indicators such as (such as DBI) are not suitable in some datasets (such as in Credit). Specific algorithm applications can be determined jointly by multiple indicators voting. In the first stage, the KMDPC algorithm under the Two-stage clustering framework proposed in this paper uses the Sil to guide the determination of the number of clusters.

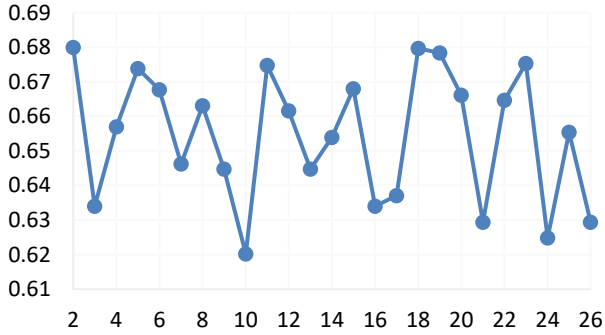
2. The effect of the k value on the KMDPC algorithm

To further analyze the effect of the choice of k on the final clustering accuracy, we manually set k in the first

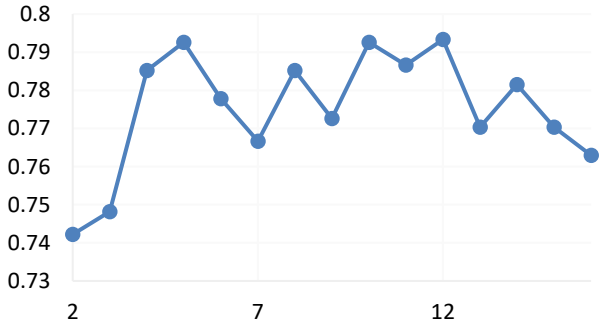
stage according to the range described in the previous section and run the KMDPC algorithm to calculate the clustering accuracy. Each k value corresponds to the average value obtained by running the algorithm five times.



a) CAs according to different k values for the Acute dataset.



b) CAs according to different k values for the Credit dataset.



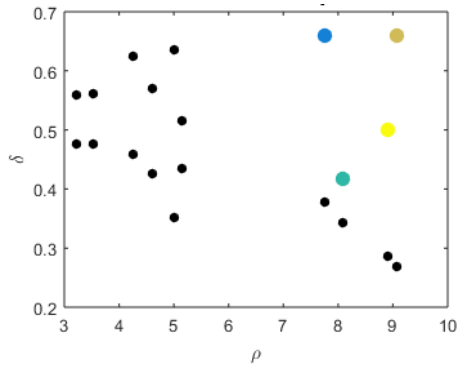
c) CAs according to different k values for the Heart dataset.

Figure 5. CAs according to different k values for the three datasets.

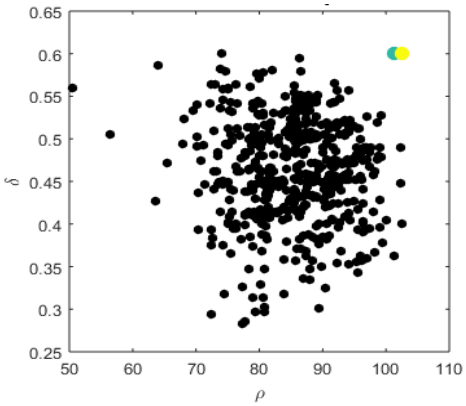
As can be seen from the above figure, the value of k is essential in this algorithm and has some influence on the clustering accuracy. The average clustering accuracy is 0.8417 when k=2 for the Acute dataset, and 0.3423 when k=9. For the Acute and Credit datasets, k=2 can obtain quasi-optimal clustering results. For the Heart dataset, the three validity indicators in the first section all show the best value of k to be 2. From the graph in Figure 5-c), we can see that the clustering accuracy reaches its lowest at this time, but the overall accuracy is higher than 0.74. Moreover, the clustering effect is better than the average clustering accuracy of the K-prototypes algorithm. It shows that it is feasible to use the Sil index to guide k value determination.

3. Manual determination of the number of clusters

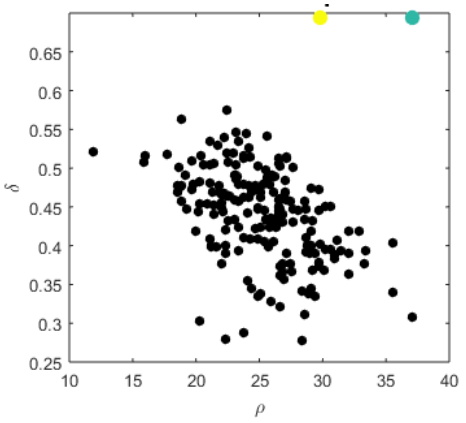
The final number of clusters for the KMDPC algorithm can be determined manually by using the decision graph of the DPC algorithm. Figure 6 below is a decision diagram for the algorithm running on three data sets and the correspondingly determined cluster centers. As can be seen from the figure, the Acute dataset has four central points, and the Heart and Credit datasets each has two central points.



a) Decision graph for the acute dataset.



b) Decision graph for the credit dataset.



c) Decision graph for the heart dataset.

Figure 6. Decision graphs and the center points for the three datasets.

5. Conclusions

In this paper, a Two-stage clustering framework is proposed, in which the numerical attributes are clustered in the first stage, and then the results are combined into the categorical attributes and clustered in the second stage. The implementation of the KMDPC algorithm is proposed. This paper studied the key problems in the implementation, such as the selection of algorithm, the determination of the number of clusters k, and the

distance calculation of the categorical attribute dataset. This algorithm solves several important problems as follows:

1. K-means algorithm guided by SIL index is used to solve the problem of automatic determination of the number of numerical attribute clustering;
2. The improved Goodall similarity is used to calculate the distance between the data instances of the categorical attribute, so the DPC algorithm can be successfully applied to cluster the categorical attribute data;
3. The Two-stage clustering framework is adopted to solve the clustering problem of mixed attribute data.

Experiments show that the algorithm is easy to understand and has promising clustering performance.

As can be seen from the existing experiments, the KMDPC algorithm proposed in this paper is superior to the K-prototypes. The clustering accuracy on the Acute, Heart and Credit datasets are 17%, 24% and 21% higher on average than those of the K-prototypes algorithms, respectively.

The Two-stage clustering framework has simple principles and promises flexible applications, and can solve the clustering problem of mixed attribute data with the help of existing clustering algorithms. It is still worth further study and discussion concerning the selection of algorithms in the two stages, the automatic determination of the number of clusters, and the determination of weights of attributes in the subset of categorical attributes. The Adaptive Density Peaks Clustering Based on K-Nearest Neighbors (K-NN) [15, 19] and the data-driven thought [4] can be introduced in further researches.

Acknowledgements

This paper was supported by Zhejiang Natural Science Foundation Qingshan Lake Science and Technology City Joint Fund (Grant No. LQY19F020001), and the Major Scientific Research Projects of Wenzhou Polytechnic (Grant No. WZYSDCY2018002).

References

[1] Bai T., Ji J., He J., and Zhou C., "New Clustering Method of Mixed-Attribute Data," *Journal of Jilin University (Engineering and Technology Edition*, vol. 43, no.1, pp.130-134, 2013.

[2] David G. and Averbuch A., "Spectralcat: Categorical Spectral Clustering of Numerical and Nominal Data," *Pattern Recognition*, vol. 45, no. 1, pp. 416-433, 2012.

[3] Du M., Ding S., and Xue Y., "A Novel Density Peaks Clustering Algorithm for Mixed Data," *Pattern Recognition Letters*, vol. 97, pp. 46-53, 2017.

[4] Du T., Qu S., and Wang Q., "A Data-Driven Parameter Adaptive Clustering Algorithm Based on Density Peak," *Complexity*, pp.1-14, 2018.

[5] Fang F., Qiu L., and Yuan S., "Adaptive Core Fusion-Based Density Peak Clustering for Complex Data with Arbitrary Shapes and Densities," *Pattern Recognition*, vol. 107, no. 3, pp. 107452, 2020.

[6] Gan G., Ma C., and Wu J., *Data Clustering: Theory, Algorithms, and Applications*, Siam, 2007.

[7] Goodall D., "A New Similarity Index Based on Probability" *Biometrics*, vol. 22, no. 4, pp. 882-907, 1966.

[8] Han J., Kamber M., and Pei J., *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2012.

[9] He Z., Xu X., and Deng S., "Squeezer: An Efficient Algorithm for Clustering Categorical Data," *Journal of Computer Science and Technology*, vol. 17, no. 5, pp. 611-624, 2002.

[10] He Z., Xu X., and Deng S., "Clustering Mixed Numeric and Categorical Data: A Cluster Ensemble Approach," *High Technology Letters*, vol. 9, no. 4, pp. 1-14, 2005.

[11] Huang D. and Li X., "Incremental Relative Density-Based Clustering Algorithm for Mixture Datasets," *Control and Decision*, vol. 28, no. 6, pp. 815-822, 2013.

[12] Huang Z., "Clustering Large Data Sets with Mixed Numeric and Categorical Values," in *Proceedings of the 1st pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 21-34, 1997.

[13] Ji J., Research on Algorithms for the Data with Multidimensional Mixed Attributes, Theses, Jilin University, 2013.

[14] Jia Z. and Song L., "Weighted k-Prototypes Clustering Algorithm Based on the Hybrid Dissimilarity Coefficient," *Mathematical Problems in Engineering*, vol. 2020, pp. 1-13, 2020.

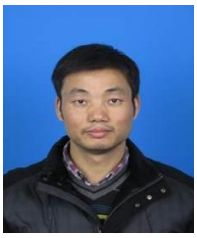
[15] Jiang D., Zang W., Sun R., Wang Z., and Liu X., "Adaptive Density Peaks Clustering Based on K-Nearest Neighbor and Gini Coefficient," *IEEE Access*, vol. 8, pp. 113900-113917, 2020.

[16] Li T., Chen Y., Zhang J., and Qin S., "Incremental Clustering Algorithm of Mixed Numerical and Categorical Data Based on Clustering Ensemble," *Control and Decision*, vol. 27, no. 4, pp. 603-608, 2012.

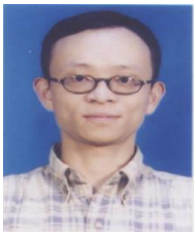
[17] Li C. and Biswas G., "Unsupervised Learning with Mixed Numeric and Nominal Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 4, pp. 673-690, 2002.

[18] Liu S., Zhou B., Decai H., and Shen L., "Clustering Mixed Data by Fast Search and Find of Density Peaks," *Mathematical Problems in Engineering*, vol. 2017, pp.1-7, 2017.

[19] Liu Y., Ma Z., and Yu F., "Adaptive Density Peak Clustering Based on K-Nearest Neighbors with Aggregating Strategy," *Knowledge-Based Systems*, vol.133, pp. 208-220, 2017.
[20] Piao S., Chaomurilige., and Yu J., "Cluster Validity Indexes for FCM Clustering Algorithm," *Pattern Recognition and Artificial Intelligence*, vol. 28, no. 5, pp. 452-461, 2015.
[21] Qian C. and Huang D., "Clustering Algorithm for Mixed Data Based on Dimensional Frequency Dissimilarity and Strongly Connected Fusion," *Pattern Recognition and Artificial Intelligence*, vol. 29, no. 1, pp. 82-89, 2016.
[22] Rodriguez A. and Laio A., "Clustering by Fast Search and Find of Density Peaks," *Science*, vol. 344, no. 6191, pp. 1492-1496, 2014.
[23] Romano S., Vinh N., Bailey J., and Verspoor K., "Adjusting for Chance Clustering Comparison Measures," *Journal of Machine Learning Research*, vol. 17, pp. 1-32, 2015.
[24] Shah S. and Amjad M., "Preceding Document Clustering by Graph Mining Based Maximal Frequent Termsets Preservation," *The International Arab Journal of Information Technology*, vol. 16, no. 3, pp. 364-370, 2019.
[25] Strehl A. and Ghosh J., "Cluster Ensembles-A Knowledge Reuse Framework for Combining Multiple Partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583-617, 2003.
[26] Sun L, Liu R., Xu J., and Zhang S., "An Adaptive Density Peaks Clustering Method with Fisher Linear Discriminant," *IEEE Access*, vol. 7, pp. 72936-72955, 2019.
[27] Sun Z., Su H., and Liang Y., "Improved K-Prototypes Clustering Algorithm," *Computer Engineering and Applications*, vol. 56, no. 21, pp. 54-59, 2020.
[28] Wang K., Li J., Zhang J., and Guo L., "Experimental Comparison of Clusters Number Estimation for Cluster Analysis," *Computer Engineering*, vol. 34, no. 9, pp.198-199, 2008.
[29] Xie J. and Qu Y., "K-medoids Clustering Algorithms with Optimized Initial Seeds by Density Peaks," *Computer Science and Exploration*, vol. 10, no. 2, pp. 230-247, 2016.
[30] Xu R., Xu J., and Wunsch D., "A Comparison Study of Validity Indices on Swarm-Intelligence-Based Clustering," *IEEE Transactions on Systems Man and Cybernetics Part B Cybernetics*, vol. 42, no. 4, pp. 1243-1256, 2012.
[31] Xu X., Ding S., Wang L., and Wang Y., "A Robust Density Peaks Clustering Algorithm with Density-Sensitive Similarity," *Knowledge-Based Systems*, vol. 200, pp.106028, 2020.
[32] Zhao Y., Li B., Li X., and Liu W., "Cluster Ensemble Method for Databases with Mixed Numeric and Categorical Values," *Journal of Tsinghua University (Science and Technology)*, vol. 46, no. 10, pp. 1673-1676, 2006.



Shihua Liu received his M.S. degree from Zhejiang University of Technology in Hangzhou, china in 2008, and received the Ph.D. degree in control science and engineering from Zhejiang University of Technology in 2018, He is currently an Associate professor in Wenzhou Polytechnic, his research interests include Machine Learning, Data Mining and Information Security.



Hao Zhang is currently an Associate professor in Information Technology Department of Wenzhou Polytechnic, his research interests include Data Mining, Digital forensics and Information Security.

Xianghua Liu received her M.S. degree from Huazhong University of Science and Technology in Software Engineering. She is a lecturer in Wenzhou Polytechnic, her research interests include Web application development and security, Data Mining.