A three-stage framework for clustering mixed data

SHI-HUA LIU [1,2], LIANG-ZHONG SHEN [3],DE-CAI HUANG [1*]
[1] College of Computer Science and Technology
Zhejiang University of Technology
288 Liuhe Road, Hangzhou, CHINA
[2] Department of Information Technology,
Wenzhou Vocational & Technical College
Chashan University town, Wenzhou, CHINA
[3] City College of Wenzhou University,
Chashan University town, Wenzhou, CHINA
Email: [1*] hdc@zjut.edu.cn; [2] chaoshua@gmail.com;

Abstract: - A three-stage framework and a simple implemented algorithm for clustering mixed data are proposed. In the framework, the mixed dataset is divided into several subsets according to the different types of attributes, and each subset is clustered using according off-the-shelf algorithms, the results are combined as a new categorical dataset and then be clustered. The final result is the answer for clustering the original mixed dataset. The experimental results show that the proposed framework and the implemented algorithm can be used to cluster the mixed dataset efficiently and it is prior to the k-prototypes.

Key-Words: - Clustering mixed data, Cluster ensemble, Clustering principle, k-prototypes, kMM algorithm

1 Introduction

As the development of the information technology, the massive data are produced every day in many fields such as Health, Education, Business, Social network and Shopping. All the data are contain the numeric and categorical attributes. How to clustering the mixed data has become a research hotspot.

The paper gives an overview of the related works of clustering the mixed data, and categorizes them into four classes. When studying the off-the-shelf approaches, we find that every clustering result can be represented as a one-dimensional categorical dataset. Based on the principle, we propose a three-stage framework and a simple implementation to cluster the mixed data. In the first stage, the original dataset is divided into several subsets according to the attribute types. Every subset is clustered and produces a categorical vector as the clustering result in the second stage. Finally, all categorical result vectors are composed as a categorical dataset and clustered using the existed algorithm, the result is the answer for clustering the original dataset. The framework can be easily implemented and modified; the paper proposes a algorithm based on k-Means and k-Modes [1] , called kMM, which use k-Means algorithm to cluster the numeric subsets in the second stage and use k-Modes to cluster the new categorical dataset in the third stage.

The experimental results show that the proposed framework and the kMM algorithm are efficient, flexible and the clustering accuracy is higher than the k-prototypes.

2 Related Works

There are all kinds of methods for clustering the mixed data. They can be divided into the following four categories: Attribute Conversions methods, Prototype-based methods, Cluster Ensemble methods and others such as density-based or hierarchical methods. Fig.1 illustrates the taxonomy of mixed data clustering methods.

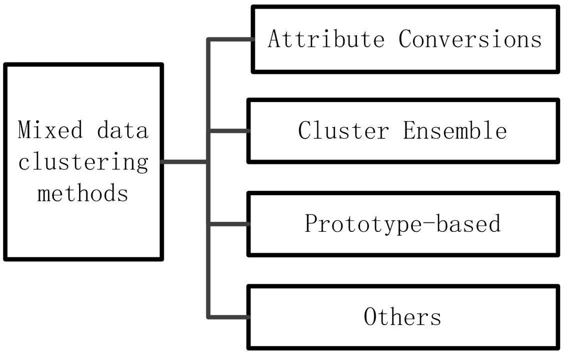

Fig.1 Taxonomy of mixed data clustering methods

2.1 Attribute Conversions methods

The Attribute Conversions methods are easy to understand; they convert the different attribute types into a unified one and then cluster the converted dataset using an existing algorithm for specific types of data. There are two conversion directions: the one is to convert all numerical attributes to categorical attributes and run categorical clustering algorithms such as k-Modes, the other is to convert all categorical attributes to numerical attributes and run numerical clustering algorithms such as k-Means. The most recent example of this kind of methods is SpectralCAT [2] which proposed by Gil David and Averbuch in 2012. It automatically transforms the high-dimensional input data into categorical values according to the Calinski–Harabasz Index, and then clusters the transformed high-dimensional data via spectral clustering. The experiments show that SpectralCAT are generic and suitable to operate on different data types from various domains including high-dimensional data.

The most popular method of Attribute Conversion is convert numerical attributes to categorical attributes, it is an independent research field called discretization. Obtaining the optimal discretization is NP-complete. A vast number of discretization techniques can be found in the literature. It is obvious that when dealing with a concrete problem or data set, the choice of a discretizer will condition the success of the posterior learning task in accuracy, simplicity of the model, etc. Different heuristic approaches have been proposed for discretization, for example, approaches based on information entropy, statistical Chi2 test, likelihood, rough sets, etc. Salvador García etc. provided a survey of discretization methods from a theoretical and empirical perspective in [3] . They presented a taxonomy of more than 80 discretization methods and the criteria used for building it. They categorized all the methods following a hierarchy based on the order: static/dynamic, univariate/ multivariate, supervised/unsupervised, splitting/ merging/hybrid, global/local, direct/incremental, and evaluation measure. Furthermore, the most important discretizers (classic and recent) have been empirically analyzed over a vast number of classification data sets in their paper. They concluded that FUSINTER, ChiMerge, CAIM, and Modified Chi2 offer excellent performances considering all types of classifiers and FUSINTER, Distance, Chi2, MDLP, and UCPD obtain a satisfactory tradeoff between the number of intervals produced and accuracy. Doctor Yu Sang [4] from Dalian University of Technology systematically analyzed existing discretization methods of continuous data and studied them in-depth from different aspects. He divided the discretization methods into the following categories: The discretization method based on statistical independence, the discretization method based on class-properties depend on each other, The discretization method based on information entropy, the discretization method based on the relationship between multiple attributes, and so on. And he proposed a combined single attribute and multi-attribute bottom-up discretization method, a discretization method for disposing high-dimensional data based on nonlinear dimension reduction technique and a data discretization method based on improved chi-square statistic.

Since the clustering analysis is unsupervised, the discretization method used to cluster the mixed dataset must be unsupervised either. The number of unsupervised discretization methods is relatively less than supervised ones. Common examples include EqualWidth(briefly called EQW), EqualFrequecy (briefly called EQF) [3] and the Density-based KDE/TDE method [4].

The second direction of Attribute Conversion is convert categorical attributes to numerical attributes; it is usually used in the model of assessment or evaluation system. The assignment of assessment indicators referred to assign a numerical value to a categorical indicator. It is often assigned by domain experts in specific assessment systems. To the best of our knowledge, the common used assignment method is not reported.

2.2 Cluster Ensemble methods

Cluster Ensemble has been proved to be a good alternative when facing cluster analysis problems. It consists of generating a set of clusterings from the same dataset and combining them into a final clustering. The goal of this combination process is to improve the quality of individual data clusterings. The mothod was proposed by A. Strehl and J. Ghosh in 2002 [5].

Every Cluster Ensemble method is made up of two steps: Generation Mechanism and Consensus Function. See Fig.2, let X be the dataset that should be clustered, it contains n data points $X = \{X_1, X_2, X_3, \dots X_n\}$, we can clustering the dataset X for H times using same or different algorithms, they produced H clusterings denoted as $P=\{ P_1,P_2, \dots P_H\}$, here $P_k(k = 1, 2, \dots, H)$ represents as the kth result in the kth clustering process. It is the first step in Cluster Ensemble method named Generation Mechanism.

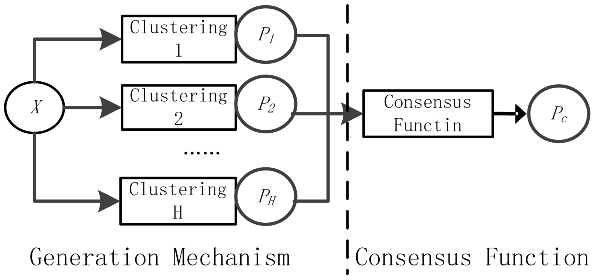

Fig.2 The process of Cluster Ensemble methods

The second step Consensus Function is the main step in any Cluster Ensemble algorithm. Precisely, the great challenge in Cluster Ensemble is the definition of an appropriate Consensus Function, capable of improving the results of single clustering algorithms. In this step, the final data partition or consensus partition P_c, which is the result of any Cluster Ensemble algorithm, is obtained.

There are many mechanisms and functions in both steps. Paper [6] and [7] gave overviews of Cluster Ensemble algorithm or techniques.

In the first step, the weak clustering algorithms are also used. These algorithms make up a set of clusterings using very simple and fast procedures. The Generation Mechanisms include using different object representations such as using different subsets of features or using different modeling of the problem; using different clustering algorithms or using different parameters initializations; using different subsets of objects or projecting the dataset to subspaces. in the generation step, it is advisable to use those clustering algorithms that can yield more information about the data [6].

It is the critical that select a proper Consensus Function in the research of Cluster Ensemble methods. There are two main Consensus Function approaches: objects co-occurrence and median partition [6]. They can be categorized as following types [7]:

- Hypergraph Partitioning: such as CSPA (Similarity Partitioning Algorithm), HGPA (Hypergraph Partitioning Algorithm), MCLA (Meta Clustering Algorithm).
- Voting Approach: such as PV (Plurality Voting), V-M (Voting-Merging), VAC (Voting Active Clusters).
- Mutual Information Algorithm: such as QMI (Quadratic Mutual Information).
- Co-association based functions: such as CTS(Connected-Triple Based Similarity).
- Finite Mixture model: such as CE-EM.

The Cluster Ensemble methods are the mainstream methods in the research of mixed data clustering analysis. ZHAO Yu et al. proposed a Cluster Ensemble method for databases with mixed numeric and categorical values called CEMC(cluster ensemble-based mixed attribute cluster) in 2006 [8], they use the subsets of attributes as the Generation Mechanism. The original dataset contains numerical and categorical attributes; the different type of attributes had been divided into different subsets and then clustered, they defined a average normalized mutual information (ANMI) as the objective function in the Consensus Function step. Experimental results on real datasets show that the clustering accuracy is better than existing mixed numeric and categorical data clustering algorithms. Zengyou He et al. proposed a Cluster Ensemble approach called CEBMDC (Cluster Ensemble Based Mixed Data Clustering) to clustering the mixed numeric and categorical data [10]. The original mixed dataset was divided into two subsets: the pure categorical dataset and the pure numeric dataset in the first step, and the existing clustering algorithms can be easily used in the second step. In the CEBMDC, they used squeezer algorithm to cluster the categorical subset in the first step and used squeezer as the Consensus Function in the second step. LI et al. proposed an incremental clustering algorithm of mixed numerical and categorical data based on Cluster Ensemble, which adopts the results of several clustering to replace that of single clustering and modifies the design of threshold.

2.3 Prototype-based methods

K-Means is a typical prototype-based clustering algorithm. This kind of methods use a “prototype” to represent a cluster, the “prototype” can be a center data point of the cluster. The famous one for clustering the mixed data is k-prototypes algorithm proposed by Huang in 1997 [12]. In the algorithm, *k* prototypes are defined to represent the centers of the clusters; a prototype combined the center of numerical data and the mode of categorical data; a distance function is defined to calculate the dissimilarity between a data point and a prototype, and a method is developed to dynamically update the k prototypes in order to maximize the intra cluster similarity of objects.

Let $X = \{X_1,X_2,\ldots,X_n\}$ denote a set of n objects and $X_i = [x_{i1},x_{i2},\ldots,x_{im}]$ be an object represented by m attribute values. Let *k* be a positive integer. The objective of clustering *X* is to find a partition which divides objects in *X* into *k* disjoint clusters. The distance function defined as follow:

$$d(X_i, Q_l) = \sum_{j=1}^{m_r} (x_{ij}^r - q_{lj}^r)^2 + \gamma_l \sum_{j=1}^{m_c} \delta(x_{ij}^c, q_{lj}^c) = E^r + E^c \quad (1)$$

where $\delta(p,q)=0$ for $p=q$ and $\delta(p,q)=1$ for $p \neq q$. x_{ij}^r and q_{lj}^r are values of numeric attributes, whereas x_{ij}^c and q_{lj}^c are values of categorical attributes for object *i* and the prototype of cluster *l*. m_r and m_c are the numbers of numeric and categorical attributes. γ_l is a weight for categorical attributes for cluster *l*.

The algorithm is based on the k-means paradigm but removes the numeric data limitation whilst preserving its efficiency. But it has the same obstacle as k-Means, it is sensitive to the initial values. Besides, the k-prototypes introduced a parameter γ_l which must be tuned carefully.

Many improved version had been proposed in recent years. BAI et al. proposed a new Global K-Prototype algorithm (GKP) in 2013 [13], the algorithm randomly selects a sufficiently large number of initial prototypes to account for the global distribution of the data sets. Then, it progressively eliminates the redundant prototypes using an iterative optimization process with an elimination criterion function. They announced that the proposed algorithm significantly improves the clustering accuracy. JI also proposed a weighted fuzzy k-prototypes algorithm (WFK-prototypes) [14] which induced the idea of fuzzy set and fuzzy clustering to deal with the fuzzy nature of data objects, utilized the co-occurrence of attribute values to calculate the impact of attribute in clustering process.

The key point of this kind of methods is the definition of the distance function. Besides the k-prototypes family, there are some other algorithms be proposed. Ahmad and Dey [15] proposed a new mixed data clustering algorithm using new cost function and distance measure based on co-occurrence of values. Yiu-ming Cheung et al. [16] proposed a general clustering framework and an iterative clustering algorithm based on the concept of object-cluster similarity and gives a unified similarity metric which can be simply applied to the data with categorical, numerical, and mixed attributes. Moreover, to circumvent the difficult selection problem of cluster number, they further developed a penalized competitive learning algorithm within the proposed clustering framework. The embedded competition and penalization mechanisms enable this improved algorithm to determine the number of clusters automatically by gradually eliminating the redundant clusters.

2.4 Others

There are many other methods to clustering the mixed data such as density-based or hierarchical methods. Li and Biswas [17] proposed a Similarity-Based Agglomerative Clustering (SBAC) algorithm which used a similarity measure proposed by Goodall and employed an agglomerative algorithm to construct a dendrogram to extract a partition of the data. Hsu and Chen [18] proposed a clustering algorithm CAVE which is based on variance and entropy, it used the variance to measure the similarity of the numeric part of the data and the similarity of the categorical part is measured based on entropy weighted by the distances in the hierarchies. LIAO et al. proposed an algorithm to cluster the hybrid data. The method changes the object's attributes to lattice based on the conception of simple tuples and hyper tuples, uses the numbers of covers to measure the similarity between labels, and chooses the clustering mean-point according to the rule of high covers to high similarity. HUANG and LI proposed an relative density-based clustering algorithm for mixture data sets (RDBC_M) [20]. It can discover the arbitrary shape clusters. HUANG also proposed some data stream cluster algorithm named MCStream [26] and double k-nearest neighbors algorithm [27] using the idea of dimension-oriented distance.

3 Three-stage Clustering Framework

3.1 The principle of clustering

Cluster analysis or simply clustering is the process of partitioning a set of data objects(or observations) into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters [21].

Mathematically, let $X = \{X_1, X_2, X_3 ... X_n\}$ represents the dataset which contains *n* objects, each of which is described by *d* attributes. Where $X_i = (x_{i1}, x_{i2}, . . . , x_{id})^T$ is a vector denotes the *i*th object and x_{ij} is a scalar denoting the *j*th component or attribute of X_i. The number of attributes *d* is also called the dimensionality of the data set. When clustering the dataset *X*, the dataset would be divided into several subsets $X=\{C_1,C_2,...,C_k\}$ (where *k* is the number of clusters) according to the distances or similarities of each data point in the same subset, and each cluster can be represented by a centre point of the subset commonly. After clustered, all objects in a cluster will combine various plausible criteria and requirements such as [22]:

1. share the same or closely related properties;
2. show small mutual distances or dissimilarities;
3. have “contacts” or “relations” with at least one other object in the group;
4. be clearly distinguishable from the complement, i.e., the rest of the objects in the dataset.

The clusters may be different according to the applications and the clustering algorithms. For numerical data, Lorr [23] suggested that there appear to be two kinds of clusters: compact clusters and chained clusters. A compact cluster is a set of data points in which members have high mutual similarity. Usually, a compact cluster can be represented by a representative point or center. A chained cluster is a set of data points in which every member is more like other members in the cluster than other data points not in the cluster. According to the ownership of each data point, the clustering analysis can be divided into hard clustering and fuzzy clustering. We focus on the hard clustering in this paper. Mathematically, the result of hard clustering algorithms can be represented by a $k \times n$ matrix, see equation (2):

$$U = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \cdots & & & \\ u_{k1} & u_{k2} & \cdots & u_{kn} \end{bmatrix} \qquad (2)$$

Where n denotes the number of records in the data set, k denotes the number of clusters, and u_{ji} satisfies the following criteria:

$$U_{ji} \in \{0,1\}, \quad 1 \le j \le k, 1 \le i \le n \qquad (2.1)$$

$$\sum_{j=1}^{k} U_{ji} = 1, \quad 1 \le i \le n \qquad (2.2)$$

$$\sum_{i=1}^{n} U_{ji} > 0, \quad 1 \le j \le k \qquad (2.3)$$

Constraint (2.1) implies that each object either belongs to a cluster or not. Constraint (2.2) implies that each object belongs to only one cluster. Constraint (2.3) implies that each cluster contains at least one object, i.e., no empty clusters are allowed. We call $U = (u_{ji})$ defined in equation (2) a hard k-partition of the data set X.

Every clustering algorithm is based on the index of similarity or dissimilarity between data points. But how to compute or measure the similarity of the data point is depend on the type of attributes and the measurement calculation on the attributes. A useful (and simple) way to specify the type of an attribute is to identify the properties of numbers that correspond to underlying properties of the attribute. There are four types of attributes [24]: nominal, ordinal, interval and ratio. Nominal and ordinal are collectively referred to as categorical or qualitative attributes, interval and ratio are collectively referred to as quantitative or numeric attributes. Because the operation properties and the similarity measurement are different in numeric and categorical attributes, most off-the-shelf algorithms are designed for specific dataset with single type of attribute. To deal with the mixed dataset, the different methods like above may be used.

In clustering analysis, every attribute despite numeric or categorical will contribute to the final result. The cluster result can be represented by a matrix U like in equation (2). But in hard clustering, the matrix U is sparse, there is only one element is 1 in every row. So we can use a categorical vector to represent the result. It can be denoted in the equation (3).

$$\left\{\begin{matrix} X_{11} & X_{12} & \cdots & X_{1d} \\ X_{21} & X_{22} & \cdots & X_{2d} \\ \cdots & & \cdots & \\ X_{n1} & X_{n2} & \cdots & X_{nd} \end{matrix}\right\} \xrightarrow{f:clustering} \left\{\begin{matrix} C_1 \\ C_2 \\ \cdots \\ C_n \end{matrix}\right\} \qquad (3)$$

Where the matrix in the left is the represent the dataset X, every row is represent a data point. When clustering the dataset X, we use some algorithm f to project the X to a vector $C=\{c_1,c_2,...c_n\}^T$. the C is a categorical vector and it is the clustering result. When data point m and n is clustered into a same cluster, then $c_m=c_n$. Every cluster can be identified by a unique label in the vector C.

From this point of view, all clustering process can be looked as a projection from data matrix X to a categorical vector C.

3.2 Three-stage clustering framework

After analysis the principle of the clustering, we can see that every attribute in the dataset can be used to calculate the similarity between the objects and contribute to the cluster result more or less. So we can use the subspace clustering or multi stage clustering in clustering the mixed data. The clustering result in the former stage is a categorical vector and can be regarded as a categorical attribute of the source dataset. It can be merged into the original dataset to conduct the final result. In general, we present a three-stage clustering framework for clustering the mixed data.

3.2.1 The architecture of the framework

Suppose the mixed dataset *X* contains *n* objects and *d* dimensions. The dataset *X* has *p* numeric attributes and *d-p* categorical attributes, which can be represented as $X = \{ X_1^N, X_2^N, \ldots, X_p^N, X_{p+1}^C, X_{P+2}^C, ..., X_d^C \}$, where X_j^N (*j*=[*1,2,...,p*]) means the *j*th numeric attribute and X_l^C (*l*=[*p+1,p+2,...,d*]) means the *l*th categorical attribute. Then the architecture of the three-stage clustering framework can be illustrated by Fig.3.

In the first stage, the dataset *X* will be divided into several subsets and every subset must have only one kind of attribute type, numeric or categorical. For example, the dataset *X* with *p* numeric attributes and *d-p* categorical attributes can be separated into *m* subsets $S = \{ S_1^N \ldots S_r^N, S_{(r+1)}^C \ldots S_m^C \}$, which contain *r* numeric subsets $\{ S_1^N ... S_r^N \}$ and *m-r* categorical subsets $\{ S_{(r+1)}^C ... S_m^C \}$.

In the second stage, every subset will be clustered using corresponding algorithms, which produce m categorical vectors represent the clustering results. The results can be merged as a new categorical dataset represented by $C=\{C_1,C_2,...,C_m\}$.

In the third stage, the new categorical dataset *C* will be clustered using existing categorical clustering algorithm. The result *U* will be looked as the final result of clustering the original mixed dataset *X*.

The architecture and the process of the three-stage clustering framework are illustrated by Fig.3.

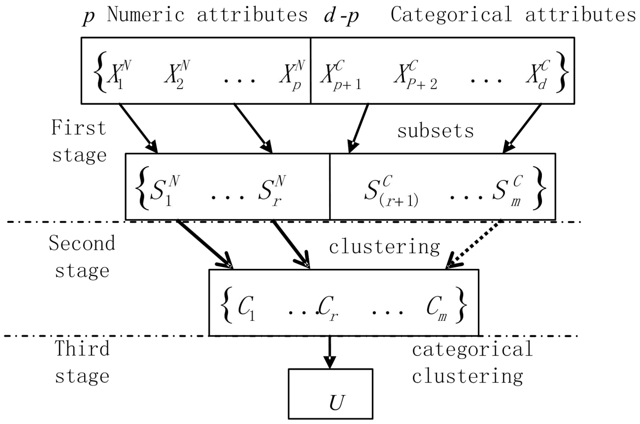
p Numeric attributes d-p Categorical attributes
$\{X_1^N \; X_2^N \; \ldots \; X_p^N$ | $X_{p+1}^C \; X_{P+2}^C \; \ldots \; X_d^C\}$
First stage
subsets
$\{S_1^N \; \ldots \; S_r^N$ | $S_{(r+1)}^C \; \ldots \; S_m^C\}$
Second stage
clustering
$\{C_1 \; ..C_r \; \ldots \; C_m\}$
Third stage
categorical clustering
U

Fig.3 The architecture of the clustering framework

3.2.2 The application of the framework

In the framework, the mixed dataset *X* must be firstly partitioned into several subsets *S* according to the type of attributes. The partition can be orderly, randomly or use some generation mechanisms as in cluster ensemble methods. When the subsets *S* have been generated, according clustering algorithms can be used to cluster every subset. For example, the off-the-shelf algorithms such as k-means, FCM, DBSCAN, or EM can be used to cluster numeric subsets with different distribution or shapes; the k-modes, Squeezer etc can be used to cluster the categorical subsets. The results of clustering every subset can be merged into a categorical dataset, and then the dataset can be clustered by an existing categorical clustering algorithm. The final result is the answer of clustering the mixed dataset.

There are many choices in every stage in the framework; it is a more generalized method. When we partition every numeric attribute as one subset in the first stage and convert every subset into categorical attribute using numeric clustering algorithm in the second stage, it can be looked as a kind of attribute conversions method.

In essence, it is a kind of cluster ensemble method, the generation mechanisms can be used in the first two stages and the consensus function can be used in third stage. But the framework can be extended and modified more easily, we can use different existing methods in every stage in the framework, and we can modify or improve these methods to get better result. For example, we can used a modified k-modes to cluster the categorical dataset in the third stage; or we can use the cluster ensemble methods in the first two stages to generated several categorical result vectors and then get the final result by clustering the merged categorical dataset in the third stage. More detailed improvements may be discussed in the following experimental part.

3.2.3 Key issues of the framework

There are some key issues must be solved when implementing the framework:

a) The partition of the attribute subset. How to divide the dataset according to the type of attributes and how many subsets may be properly partitioned are key issues in the first stage in the framework. The partition criteria should be designed carefully to gain the better result, but it should be analysis how much influence the result of clustering. It is the simplest way that partitions the attribute one by one or randomly.

b) The selection of the clustering algorithms. There are so many existing clustering algorithms to cluster the numeric or categorical datasets in the second and third stages. It is key issues of the framework that deciding which algorithm is better or how to select the best one. It may be guided by the distribution and shapes of the dataset, and the application domain is must be considered.

c) The determination of the prior parameters. The typical prior parameter in clustering analysis is the cluster number. Although there are many approaches to determine the cluster number automatically, the most popular method is to assign it manually in practice. In the three-stage framework, it needs some experimental research to determine the cluster number in second and third stages. The cluster number in the second stage how to influence the final result is needed to analysis. The simplest way is to give a prior number *k* manually in practice.

d) The weight of each subset or attribute. Every subset or attribute contribute more or less in the three-stage mixed data clustering analysis, we can give a weight value for each subset or attribute to reflect their contributions. But how to assign the weight is a key issue in the clustering process. One simplest way is to assign an equal weight value 1.0 for every subset or attribute. Another popular way is using the information entropy approach.

When we use the three-stage framework to cluster the mixed dataset, the above key issues must be solved first. In the next subsection, we'll give a simple implementation of the framework and give a simple discuss of how to solve these issues.

3.3 The implementation of the three-stage clustering algorithm

In order to verify the applicability of the mentioned three-stage clustering framework, we present a simple implementation of the three-stage clustering algorithm called kMM based on k-Means and k-Modes in this subsection. The proposed kMM algorithm has solved the above four issues in a simple way:

a) In the first stage, the original mixed dataset X has been partitioned into two subsets according to the type of attributes. All numeric attributes are divided into one numeric subset X^N and all categorical attributes in another categorical subset X^C.

b) In the second stage, the k-Means algorithm was used to cluster the numeric subset and the result was merged into the categorical subset to construct a new categorical dataset and the k-Modes algorithm was used to cluster the categorical dataset in the third stage.

c) The cluster number *k* was assigned manually before clustering in the second and third stages.

d) The weight values of each subset or attribute are simply assigned, which is equally 1.0.

All the issues are solved in a very simply way, but it is a simple usable algorithm, based on the original kMM algorithm, there are many variants can be proposed.

The original kMM algorithm can be described in Table 1.

Table.1. The original kMM algorithm

```
Algorithm1: original kMM algorithm
Input: mixed dataset $X=\{X^N,X^C\}$ (where $X^N$ is the numeric subset and $X^C$ is the categorical subset);
Cluster number $K$
Output: cluster result vector $U$
Process:
1. Clustering the numeric subset $X^N$ using k-Means: $C_1=kMeans(X^N, K)$.
2. merge the result $C_1$ into the categorical subset $X^C$ and get a new categorical dataset $C=merge(C_1, X^C)$.
3. Clustering the new dataset $C$ using k-Modes and get the final result $U=kModes(C, K)$.
```

The original kMM algorithm uses the famous off-the-shelf algorithms like k-Means and k-Modes. It is easy to implement. It has the advantages of the both algorithms. It is efficient and can be used to cluster the large mixed dataset.

The computational complexity of kMM can be easily estimated from the process of algorithm1 in table 1. It can be denoted as $O(t_1Knd_1)+O(t_2Knd_2)$, where t_1,t_2 is the iteration times of k-Means and k-Modes, d_1,d_2 represent the dimension number of numeric subset and the dimension number of categorical dataset separately, K is the cluster number which assigned manually before clustering; n represents the number of objects in the original dataset.

4 Experimental analysis

To investigate the effectiveness of the proposed framework and kMM algorithm, we applied it to various three common mixed data sets obtained from UCI Machine Learning Data Repository [25] and compared its performance with k-prototypes. The algorithms were implemented in MATLAB2012a and all the experiments run on a laptop with Intel(R) Core(TM)2 T6670 CPU, 2.20 GHz main frequency, and 3GB DDR2 667 RAM.

Table.2. Brief introduction of UCI datasets

Abbr.	*name*	*Instances*	*numeric*	*categorical*	*decision*	*description*
Acute	Acute Inflammations	120	1	5	2	the pathological physiological indexes used to judge the acute inflammation
Credit	Credit Approval	653/690	6	9	1	customer relationship data of users who want to apply for a credit
Heart	Heart Disease	270	6	7	1	Used to judge the presence of heart disease in the patient

4.1 The datasets

Acute Inflammations, Heart Disease and Credit Approval are three popular mixed datasets from UCI data repository. They are usually used in the research of classification but also commonly used in the research of clustering later. There is one or more numeric and categorical attributes and one or two decision attribute in these datasets. The detail information has summarized in the Table 2.

The Acute dataset was created by a medical expert as a data set to test the expert system, which will perform the presumptive diagnosis of two diseases of the urinary system. It contains one numeric, five categorical and two decision attributes, so we transfer the two decision attributes to one when calculate the clustering accuracy.

The Credit dataset contains data from credit card organization, where customers are divided into two classes. It is a mixed data set with eight categorical and six numeric features. It contains 690 instances belonging to two lasses – negative (383) and positive (307). It contains 37 instances with missing values, so we use remained 653 instances in our experiment.

The Heart dataset generated at the Cleveland Clinic is a mixed data set with eight categorical and five numeric features which have been extracted from a larger set of 75. It contains 270 instances belonging to two classes – normal (150) and heart patient (120).

4.2 Experimental setup

In the experiment, we compare the clustering accuracy of k-prototypes and the proposed kMM algorithm in clustering the above three datasets. The parameters and process of kMM algorithm is introduced in subsection 3.3; the parameter $\gamma_l=1/2\ \sigma$ (σ represents the average standard deviation of the numeric attributes) for all clusters according to the research in paper [12] which recommend a suitable γ lies between 1/3 σ and 2/3 σ.

We ran the two algorithms 100 times on the Credit and Heart datasets. Since the Acute dataset has two decision attributes, we transfer the two binary decision attributes to one attributes with the cluster number K=4, we ran the two algorithms 20 times on the Acute dataset and calculate the clustering accuracy manually.

It's a very simple and basic experiment used to verify the applicability and efficiency of the proposed kMM algorithm and three-stage clustering framework. The result will be show in the next sub-subsection.

4.3 Results and discussion

We recorded the maximum, average and minimum clustering accuracy of the two algorithms run on the three datasets, the result was given in the Table 3.

Table.3.Clustering results of kMM & k-prototypes

Clustering Accuracy		Acute	Heart	Credit
Maximum	k-prototypes	0.783	0.593	0.649
	kMM	**0.900**	**0.811**	**0.824**
Average	k-prototypes	**0.673**	0.590	0.556
	kMM	0.658	**0.726**	**0.700**
Minimum	k-prototypes	**0.542**	**0.589**	**0.510**
	kMM	0.525	0.511	0.508

In order to compare the results more intuitive, we illustrate the results as bar charts in the Fig.4.

As show in the Table 3 and Fig.4, the minimum clustering accuracies of kMM are all smaller than that of k-prototypes on three datasets and the maximum clustering accuracies of kMM are all bigger than that of k-prototypes on above datasets. This means that the search space of kMM algorithm is larger than that of the k-prototypes but the stability is poorer which may be produced and accumulated by the poor stabilities of k-Means and k-Modes using in the second and third stages.

From the sub figure d) in Fig.4, we can see that the average clustering accuracy of kMM is a little smaller (-0.015) than k-prototypes on the Acute dataset but that is much bigger (0.136 and 0.144 separately) than that of k-prototypes on the Heart and Credit datasets, which means that the proposed kMM algorithm can be applied to clustering the mixed datasets in practice.

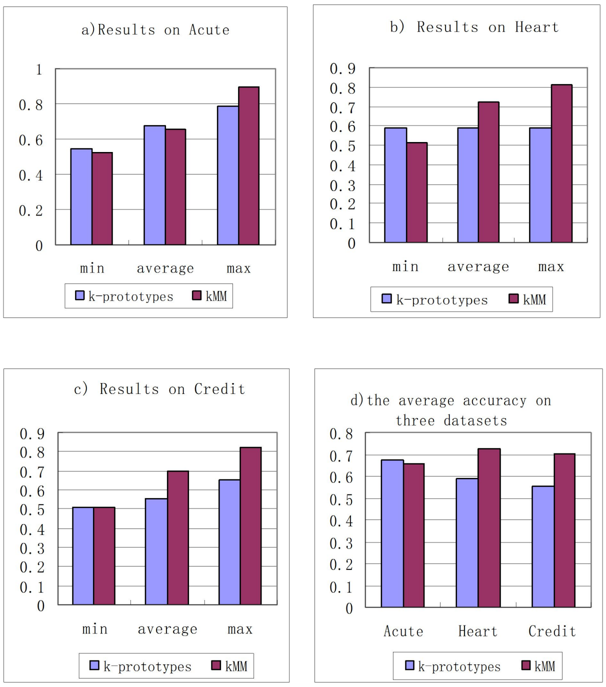

Fig.4. The clustering results on datasets

5 Conclusion

The experimental results show that the kMM algorithm based on the three-stage framework is simple and practical, it achived better accuracy in clustering real world mixed datasets without parameter tuning as in k-prototypes. The three-stage framework for clustering mixed data is an intuitive approach and easy to understand. There are four key issues should be considered when improving or optimizing the correlated algorithm: the partition of the attribute subset; the selection of the clustering algorithms; the determination of the prior parameters such as the cluster number K; the weight of each subset or attribute. All of these should be researched carefully and these are our key areas in the future work.

References:

[1] Zheyue Huang. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. Research Issues on Data Mining & Knowledge Discovery, 1-8..

[2] Gil David,Amir Averbuch. SpectralCAT: Categorical Spectral Clustering of Numerical and Nominal Data. *Pattern Recognition*, Vol.45, No.1 2012, pp.416-433.

[3] Salvador Garcia,Julián Luengo,José Antonio Sáez, et al. A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning. *Knowledge and Data Engineering, IEEE Transactions on,* Vol.25, No.4, 2013, pp.734-750.

[4] Yu Sang. *Research on Discretization Methods for Continuous Data*. DaLian: Dalian University of Technology, 2012.

[5] Strehl, A., Strehl, E., & Ghosh, J. (2002). Cluster Ensembles - A Knowledge Reuse Framework for Combining Partitionings. Journal of Machine Learning Research, Vol.3, pp.583-617.

[6] Sandro Vega-pons, José Ruiz-shulcloper. A Survey of Clustering Ensemble Algorithms . *International Journal of Pattern Recognition and Artificial Intelligence*, Vol.25, No.3, 2011, pp.337-372.

[7] Reza Ghaemi,Md Sulaiman,Hamidah Ibrahim, et al.A Survey: Clustering Ensembles Techniques..Proceedings of World Academy of Science: Engineering & Technology, Vol.50, No.38, 2009, pp.644-653.

[8] ZHAO Yu, LI Bing, LI Xiu, LIU Wenhuang, REN Shouju. Cluster ensemble method for databaseswith mixed numeric and categorical values, *Journal of Tsinghua Univercity (Sci & Tech)* , Vol.46, No.10 ,2006, pp.1673-1676.

[9] Zengyou He, Xiaofei Xu, Shengchun Deng. Squeezer: an Efficient Algorithm for Clustering Categorical Data. *Journal of Computer Science and Technology*, Vol.17, No.5, 2002, pp.611-624.

[10] Zengyou He, Xiaofei Xu, Shengchun Deng. Clustering Mixed Numeric and Categorical Data: a Cluster Ensemble Approach. *Arxiv Preprint Cs/0509011*, 2005(1):1.

[11] LI Tao-ying, CHEN Yan, ZHANG Jin-song, QIN Sheng-jun. Incremental clustering algorithm of mixed numerical and categorical data based on clustering ensemble . *Control and Decision*, Vol.27, No.4 , pp.603-608.

[12] ZHEXUE Huang. Clustering Large Data Sets with Mixed and Numeric and Categorical values, *Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining,(PAKDD)*,1997, pp.21-34.

[13] BAI Tian, JI Jin-chao, HE Jia-liang, ZHOU Chun-guang, New clustering method of mixed-attribute data. *Journal of Jilin University (Engineering and Technology Edition)*, Vol.43, No.1, 2013, pp.130-134.

[14] JI Jin-chao, *Research on algorithms for the data with multidimensional mixed attributes* . JiLin: JiLin University, 2013.

[15] Amir Ahmad, Lipika Dey. A K-mean Clustering Algorithm for Mixed Numeric and

Categorical Data. *Data and Knowledge Engineering*, Vol.63, No.2, 2007, pp.503-527.

[16] Yiu-ming Cheung, Hong Jia. Categorical-and-numerical-attribute Data Clustering Based on a Unified Similarity Metric Without Knowing Cluster Number. *Pattern Recognition*, Vol.46, No.8, 2013, pp.2228-2238.

[17] Cen Li, Gautam Biswas. Unsupervised Learning with Mixed Numeric and Nominal Data. *Knowledge and Data Engineering*, IEEE Transactions on, Vol.14, No.4, 2002, pp.673-690.

[18] Chung-Chian Hsu,Yu-Cheng Chen. Mining of Mixed Data with Application to Catalog Marketing. *Expert Systems with Applications*, Vol.32, No.1, 2007, pp.12-23.

[19] LIAO Zhi-f ang, LUO Hao, FAN Xiao-ping, LIU Ke-zhun, New hybrid data orientation cluster algorithm . *Control and Decision,* Vol.24, No.5, 2009, pp.697-700+705.

[20] HUANG De-cai, LI Xiao-chang, Incremental relative density-based clustering algorithm for mixture datasets . *Control and Decision,* Vol.28, No.6, 2013, pp.815-822.

[21] Jiawei Han, Micheline Kamber, Jian Pei. Data Mining: *Concepts and Techniques*, Morgan Kaufmann, 2006.

[22] Guojun Gan, Chaoqun Ma, Jianhong Wu. *Data Clustering: Theory, Algorithms, and Applications*, Siam,2007.

[23] Maurice Lorr. *Cluster Analysis for Social Scientists*, Jossey-bass San Francisco, 1983.

[24] P.N. Tan,M. Steinbach,V. Kumar. *Introduction to Data Mining*, Pearson Addison Wesley, 2006.

[25] Kevin Bache, Moshe Lichman. UCI Machine Learning Repository. Irvine, Ca: University of California, School of Information and Computer Science. 2013. [http://archive.ics.uci.edu/ml].

[26] HUANG De-cai, WU Tian –hong, Density -based clustering algorithm for mixture data sets. *Control and Decision,* Vol.25, No.3, 2010, pp.416-421.

[27] HUANG De-cai SHEN Xian-qiao LU Yi-hong, Double k-nearest Neighbors of Heterogeneous Data Stream Clustering Algorithm, *Computer Science,* Vol.40, No.10, 2013, pp.226-230.