IOP Conference Series: Earth and Environmental Science

PAPER • OPEN ACCESS

Innovation of Cluster Method for Mixed Data Based on Specific Initialization Process and Attribute Weighting

To cite this article: Xiaoqing Ma 2019 *IOP Conf. Ser.: Earth Environ. Sci.* **252** 052100

View the article online for updates and enhancements.

This content was downloaded from IP address 114.99.54.65 on 09/10/2019 at 21:47

Innovation of Cluster Method for Mixed Data Based on Specific Initialization Process and Attribute Weighting

Xiaoqing Ma*

The Australian National University ANU Canberra, Australia, China

*Corresponding author e-mail: xiaoqm1123@163.com

Abstract. This paper proposes one improved K-Prototype algorithm based on innovations of controlling initialization process and attribute weighting (KP-IW) in order to deal with mixed data containing numeric and categorical attributes. Making initialization of clustering fixed and giving weightings to attributes are two common principles for improving algorithms. However, there are rarely methods regarding numeric or categorical proportion as one new attribute, which will affect the initialization consequence and weight value assigning to attribute because that density distribution of instances is calculated by the combing each attributes and those entire two proportions instead of only the former. There are some more detailed innovations for initialization and weighting, involving auxiliary point, auxiliary clusters and weightings combing linear and exponential effect. And it can be concluded that the KP-IW algorithm is suitable according to the clustering evaluation scores from KP-IW compared with others algorithms.

1. Introductions and Related Works

Clustering is the most studied and widely used algorithm in the classification of unsupervised learning, which is the process of allocating a set of instances of data into several separated groups, named clusters. The purpose of clustering is to make each cluster separated from others as far as possible, and whole instances in one cluster similar as much as possible.

There are so many clustering methods based on different guidelines and appropriate for various datasets with specific characteristics. There are four basic clustering method types based on the primary principles, which are partitioning method, hierarchical method, density-based method and grid-based method [1]. Apart from that, it also can make classifications of clustering method by the feature of dataset that they deal with, which are method based on numeric data, categorical data and mixed data (containing both of them).

As for the dataset with whole attributes being numeric, K-means [2], BIRCH [3], DBSCAN [4] are the most classic methods up to now, which are belong to partitioning method, hierarchical method and density-based method, respectively. And for categorical data, the original typical method was proposed by Huang [5], another common one was present by Cao [6], and both of them are find suitable initialization method based on density of instances distribution.

Paying more attention on the method dealing with mixed data, which is also the target in the paragraph. It should be noted that in real life the mixed data occupied most of data as there is rarely completed numeric or categorical data existing. The earliest method of clustering mixed data can be

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

regarded as the algorithm K-Prototype, proposed by Huang [7], which extends the k-Means algorithm to categorical domains and domains with mixed numeric and categorical values, and it was improved from his original K-Prototype algorithm [8]. After that, researchers have studied the area of clustering mixed data in order to find some general methods which can be appropriate for mixed data, though most of algorithms come up with up to now have limitations more or less.

Across those algorithms, there is one significant type with the application of concept of fuzzy, which means making algorithm have ability of dealing with dataset with ambiguous and uncertainty. This type of algorithm also makes impact on the cluster algorithm with mixed data. The theory of fuzzy was proposed by Zadeh in 1965 at the earliest. Fuzzy k-Means algorithm was present by Ruspini in 1969 and 1973, and after that, Bezdek also proposed the algorithm of fuzzy k-Means. For categorical data, the fuzzy K-Modes proposed by Huang and Ng [9] is describes extensions to the fuzzy k-Means algorithm for clustering categorical data. Another categorical fuzzy K-Modes [10] was come up with to investigate a new variant of algorithm with automated feature weight learning in 2015. For mixed data, Lee and Pedrycz introduced a new generalization called fuzzy p-mode prototype [11], of frequency-based prototypes. Chatzis proposed an extension of the Gath–Geva algorithm to allow for the effective handling of data with mixed attributes, called KL-FCM-GM [12]. Ji et al. proposed WFK-prototype algorithm [13], which is a new fuzzy kprototype clustering algorithm for mixed data.
Another type of improved K-Prototype algorithms is utilizing weight. Based on algorithm focus on automated variable weighting in k-Means [14], which was proposed by Huang et al.. Ji et al. present an improved K-Prototype clustering algorithm [15] for mixed data.

There are some other common algorithms targeting with mixed data. Li and Biswas present a Similarity-Based Agglomerative Clustering (SBAC) [16] algorithm with utilizing a similarity measure, proposed by Goodall [17]. Zheng et al. [18] proposed an evolutionary clustering algorithm, EKP [18], for mixed type data, which focuses on global searching in order to fix problems with initialization and convergence to local optimum. SpectralCAT [19] was proposed by Hsu and Chen, which is also one spectral clustering algorithm for mixed data.

2. Basic Notations and Definitions

There are some mainly notations and notations should be clarified before all steps. Assuming that $\mathbf{A}$ is a set of m attributes and is a set of n instances. Setting that there are m attributes and n instances in each dataset, so,

$$\mathbf{A} = \{A_j \mid j = 1, 2, ..., m\} \tag{1}$$

$$\mathbf{X} = \{X_i \mid i = 1, 2, ..., n\} \tag{2}$$

Furthermore, in the studying background of clustering in mixed data, the considerations of categorical data, numeric data and the classes assigned in the dataset if there is existing should be also added here. Therefore it can be more specific to set $\mathbf{A}^r, \mathbf{A}^c, \mathbf{X}^r, \mathbf{X}^c$ represent the set of numeric attributes, categorical attributes, numeric parts in each instance and categorical parts in each instance. After setting the number of numeric attributes is m, those four sets can be shown as,

$$\begin{cases} \mathbf{A}^r = \{A_j^r \mid j = 1, 2, ..., p\} \\ \mathbf{A}^c = \{A_j^c \mid j = p+1, p+2, ..., m\} \\ \mathbf{X}^r = \{X_i^r \mid i = 1, 2, ..., n\} \\ \mathbf{X}^c = \{X_i^c \mid i = 1, 2, ..., n\} \end{cases} \tag{3}$$

Which is convenient for algorithms and instructions in following steps.

For each categorical attribute, A_j^c describes a domain of values denoted by $DOM(A_j^c)$. For the vector for each attribute or each instance, we can make some other notations,

$$X_i = (x_{i,1}, x_{i,2}, ..., x_{i,p,}, x_{i,p+1}, ..., x_{i,m}) \tag{4}$$

$$X^{(j)} = (x_{1,j}, x_{2,j}, ..., x_{n,j,}) \tag{5}$$

Obviously, there are also specific descriptions splitting numeric and categorical data, showing as,

$$\begin{cases} X^{(j)r} = (x_{1,j}^r, x_{2,j}^r, ..., x_{n,j}^r) \\ X^{(j)c} = (x_{1,j}^c, x_{2,j}^c, ..., x_{n,j}^c) \\ X_i^r = (x_{i,1}^r, x_{i,2}^r, ..., x_{i,p}^r) \\ X_i^c = (x_{p+1,1}^c, x_{p+2,1}^c, ..., x_{m,1}^c) \end{cases} \tag{6}$$

and we can further update the vector of each instance as $X_i = (x_{i,1}^r, x_{i,2}^r, ..., x_{i,p}^r, x_{i,p+1}^c, ..., x_{i,m}^c)$.

For $DOM(A_j^c)$, letting $s_j = |DOM(A_j^c)|$, and this is the number of elements in the set of domain of values in attribute A_j^c.

If there is a vector of class definition, C, external to whole dataset $\mathbf{X}$, and $C = (c_1, c_2, ..., c_n)$, the number of focused unique class is set as k, which is equal to $|DOM(C)|$.

3. Some Definitions For K-Prototype Clustering Algorithm Based on Specific Initialization Process and Attribute Weighting

3.1. Conjunction of All Categorical Attribute

It can be quoted here of the concept of conjunction of attributed-value pairs, referring in the work of Gowda and Diday in 1991 [20].

Adding one new categorical attribute A_{m+1}^f into whole dataset, which is created by conjuctions of all categorical attributes $(A_{p+1}^c, A_{p+2}^c, ..., A_m^c)$. It also can be shown as form of $[A_{p+1}^c : X^{(p+1)c}] \wedge [A_{p+2}^c : X^{(p+2)c}] \wedge ... \wedge [A_m^c : X^{(m)c}]$, or shaped like $x_{i,m+1}^f = x_{i,p+1}^c \_ x_{i,p+2}^c \_ \cdots \_ x_{i,m}^c$. For example, when $X_i^c = (2,1,4,5,3,6)$, $x_{i,m+1}^f = 2\_1\_4\_5\_3\_6$.

The categorical conjuction attribute A_{m+1}^f is set for analyzing the overall distribution features of categorical part in dataset, which will be used widely in the next algorithms.

3.2. Data Preprocessing

For numeric attribute portion, one of the data normalization method used widely is the min-max normalization. This process performs a linear transformation on the original data. Basically speaking, it maps a value of one attribute to updated value in a new closed interval with new upper and lower boundaries.

Based on the above notations and assumptions, for each A_j^r, this numeric data normalization is to updating as,

$$\begin{aligned} X^{(j)r} = \mathrm{hr}(X^{(j)r}) = \frac{X^{(j)r} - \min(X^{(j)r})}{\max(X^{(j)r}) - \min(X^{(j)r})} \cdot \\ (new\_max - new\_min) + new\_min \end{aligned} \tag{7}$$

Where, lower boundary (new_min) and upper boundary (new_max) are equal to 0 and 1, respectively.

For categorical attribute portion, the purpose preprocessing for categorical data is to simplify the elements in each $DOM(A_j^c)$, which can save time cost and space cost of the following algorithms.

Setting function $r(x)$ to get the rank of value x in the array value of attribute A_j^c in ascending order of $DOM(A_j^c)$, and the preprocessing here is to update as,

$$X^{(j)c} = \mathrm{hc}(X^{(j)c}) = (\mathrm{r}(x_{1,j}^c), \mathrm{r}(x_{2,j}^c), ..., \mathrm{r}(x_{n,j}^c)) \tag{8}$$

3.3. Basic Statistical Calculations and Descriptions Used in this Article

Given $s_j = |DOM(A_j^c)|$, then setting,

$$DOM(A_j^c) = \{a_{j,1}, a_{j,2}, ..., a_{j,s_j}\} = \{a_{j,1}, a_{j,2}, ..., a_{j,|DOM(A_j^c)|}\} \tag{9}$$

And $\mathrm{f}(a_{js}, j)$ denotes frequency of category element a_{js} in attribute A_j^c, where $s = 1, 2, ..., s_j = 1, 2, ..., |DOM(A_j^c)|$.

For each attribute A_j^r, the mean value is

$$\mathrm{M}(X^{(j)r}) = \frac{1}{|X^{(j)r}|} \sum_{i=1}^{|X^{(j)r}|} x_{i,j} \tag{10}$$

The variance is

$$\sigma_{X^{(j)r}}^2 = \mathrm{V}(X^{(j)r}) = \frac{1}{|X^{(j)r}|} \sum_{i=1}^{|X^{(j)r}|} (x_{i,j} - \mathrm{M}(X^{(j)r}))^2 \tag{11}$$

And the standard deviation is

$$\sigma_{X^{(j)r}} = (\sigma_{X^{(j)r}}^2)^{1/2} \tag{12}$$

For each attribute A_j^c, the mode element is

$$\mathrm{MO}(X^{(j)c}) = \underset{s \in \{1,2,...,|DOM(A_j^c)|\}}{\arg\max} (\mathrm{f}(a_{js}, j)) \tag{13}$$

the frequency of the mode element is

$$\mathrm{CO}(X^{(j)c}) = \max_{s \in \{1,2,...,|DOM(A_j^c)|\}} (\mathrm{f}(a_{j,s}, j)) \tag{14}$$

Setting the vector of frequency of each value in attribute A_j^c is

$$F_j = (\mathrm{f}(a_{j,1}, j), \mathrm{f}(a_{j,2}, j), ..., \mathrm{f}(a_{j,|DOM(A_j^c)|}, j)) \tag{15}$$

Therefore, apparently, the mean value of F_j is

$$\mathrm{M}(F_j) = \frac{1}{|F_j|}\sum_{s=1}^{|F_j|} \mathrm{f}(a_{j,s}, j) = \frac{1}{|DOM(A_j^c)|}\sum_{s=1}^{|DOM(A_j^c)|} \mathrm{f}(a_{j,s}, j) \tag{16}$$

And variance of F_j is

$$\begin{aligned}\sigma_{F_j}^2 &= \mathrm{V}(F_j) = \frac{1}{|F_j|}\sum_{s=1}^{|F_j|}(\mathrm{f}(a_{j,s}, j) - \mathrm{M}(F_j))^2 \\ &= \frac{1}{|DOM(A_j^c)|}\sum_{s=1}^{|DOM(A_j^c)|}(\mathrm{f}(a_{j,s}, j) - \mathrm{M}(F_j))^2\end{aligned} \tag{17}$$

And the standard deviation of F_j is

$$\sigma_{F_j} = (\sigma_{F_j}^2)^{1/2} \tag{18}$$

3.4. Dissimilarity Measure

For numeric attribute portion, we choose the Minkowski distance of order p here. The distance between two points $X = (x_1, x_2, ..., x_n)$ and $Y = (y_1, y_2, ..., y_n) \in \square^{n}$ is defined as $\mathrm{D}(x, y) = (\sum_{i=1}^{n} | x_i - y_i |^p)^{1/p}$, where p is the value of order. Applying to this paragraph, it should be $\mathrm{Dr}(X_{i1}, X_{i2}) = (\sum_{j=1}^{p} | x_{i1,j}^r - x_{i2,j}^r |^p)^{1/p}$, where p is the number of numeric attributes existing in the targeted dataset as descriptions in above notations and definitions

For categorical attribute portion, the dissimilarity measure is based on simple matching,

$$\mathrm{Dc}(X_{i1}, X_{i2}) = \sum_{j=p+1}^{m} \delta(x_{i1,j}^c, x_{i2,j}^c) \tag{19}$$

Where,

$$\delta(x_{i1,j}^c, x_{i2,j}^c) = \begin{cases} 0, if\ \ x_{i1,j}^c = x_{i2,j}^c \\ 1, if\ \ x_{i1,j}^c \neq x_{i2,j}^c \end{cases}$$

Then we combine those two dissimilarity measures.

For the basic principle in algorithm of K-Prototype proposed by Huang in 1997, the dissimilarity from two points is $\mathrm{D}(X_{i1}, X_{i2}) = \mathrm{D}r(X_{i1}, X_{i2}) + \gamma \cdot \mathrm{Dc}(X_{i1}, X_{i2})$, where γ is the weight for categorical attributes.

And γ is decided by the valued of $\mathrm{V}(X^{(j)r})^{1/2} \,|\, Cl_t$, which means that γ changes spread different clusters.Considering that there is no information of clustering labels before clustering process, the value of weight for categorical attributes is replaced by the overall average standard deviations of all numeric attributes, which is $\frac{1}{p}\sum_{j'=1}^{p}\sigma_{X^{(j')r}}$.

Furthermore, Huang also proposed that the a suitable γ lies between $\frac{1}{3}(\frac{1}{p}\sum_{j'=1}^{p}\sigma_{X^{(j')r}})$ and $\frac{2}{3}(\frac{1}{p}\sum_{j'=1}^{p}\sigma_{X^{(j')r}})$, and the default choice is $\frac{1}{2}(\frac{1}{p}\sum_{j'=1}^{p}\sigma_{X^{(j')r}})$.

The innovations about the dissimilarity measure of combing numeric part and categorical part in this article is to turn the linear relative weight of categorical compared with numerical part into two part, which contains linear importance relative to numeric portion and exponential weight controlling categorical portion itself, which are marked as γ and λ.

Considering of regarding $\mathbf{A}^r$ as one entity, which means that we can change the part of $\frac{1}{p}\sum_{j'=1}^{p}\sigma_{X^{(j')r}}$ in default choice of γ into the overall average standard deviations of $\mathrm{hr}((|x_{1,j}^r-\mathrm{M}(X^{(1)r})|,|x_{2,j}^r-\mathrm{M}(X^{(2)r})|,\ldots,|x_{p,j}^r-\mathrm{M}(X^{(p)r})|))$, named γ_1; and changing the part of $\frac{1}{2}$ in default choice into the heuristic used for $\mathbf{A}_{m+1}^f$, which is equal to $1-\sum_{s=1}^{|DOM(A_{m+1}^f)|}(\mathrm{f}(a_{m+1,s},m+1)/)^2$, named γ_2. The innovative γ is $\frac{\gamma_1}{\gamma_2}$.

The parameter γ is one linear parameter controlling the importance ratio of all numeric attributes to that of all categorical attributes. Apart from this linear parameter, there is also one exponential parameter λ inserting here, used for controlling the importance variation of categorical attributes themselves, which varies across different purpose of measuring dissimilarity. The whole dissimilarity measure is from two points is shaped like the formula as follow:

$$\mathrm{D}(X_{i1},X_{i2})=\mathrm{Dr}(X_{i1},X_{i2})+\gamma\cdot(\mathrm{Dc}(X_{i1},X_{i2}))^{\lambda/|\mathbf{A^c}|} \tag{20}$$

As for the suitable choice for λ, they will be specifically introduced in the following paragraph.

4. K-Prototype Clustering Algorithm Based on Specific Initialization Process and Attribute Weighting

4.1. *Initialization Process*

The initialization process is to selecting initial central points. In order to get the specific k initial central points, there are three classical initialization method for central points during the process of K-prototype. The first method raised by Huang [7], notated as ‘I’, is to selects the first k distinct records from the data set. Huang [7] also present the second initialization method in this paragraph, notated as ‘H’, with considering the all descending-orders of domain values’ frequency distribution for all categorical attributes. Cao [6] came up with another method, notated as ‘C’, which makes use of the average density of each instance based on the frequency of attribute values.

In the method of KP-BD-P, for the initialization method, we also take advantage of the frequency dentist of all categorical attributes, but more focus on the overall frequency dentist of all categorical attributes conjunction, which is the frequency of A_{m+1}^f as above notation showing.

Firstly, finding the auxiliary point for initialization process, $P^{(0)}$,in order to search other real central points. Seeking all instances whose value in A_{m+1}^f is same as the mode element of the conjunction attribute A_{m+1}^f, which is $\mathrm{MO}(A_{m+1}^f)$. It can be denoted by $\mathbf{P^{(0)}}$ for the set of those satisfied instances.

$$\mathbf{P^{(0)}}=\{X_i\mid x_{i,m+1}^f=\mathrm{MO}(A_{m+1}^f)\}=\{p\} \tag{21}$$

In the set $\mathbf{P^{(0)}}$, assuming that the auxiliary point $P^{(0)}$ consists of the mean value of each numeric attributes and the identical categorical-attribute-pattern across all instances in $\mathbf{P^{(0)}}$.

$$\begin{aligned} P^{(0)} = (&X^{(1)r} \mid \mathbf{P^{(0)}}, X^{(2)r} \mid \mathbf{P^{(0)}}, ..., X^{(p)r} \mid \mathbf{P^{(0)}}, \\ &x^{c}_{i0,p+1}, x^{c}_{i0,p+2}, ..., x^{c}_{i0,m}, x^{f}_{i0,m+1}) \end{aligned} \tag{22}$$

Where,

$$i_0 = \forall i \in \{i \mid x^{f}_{i,m+1} = \mathrm{MO}(A^{f}_{m+1})\}$$

And then,

$$P^{(0)} = \arg\min_{P' \in \mathbf{P^{(0)}}} (\mathrm{D}^1(P^{(0)}, P')) \tag{23}$$

$\mathrm{D}^1(X_{i1}, X_{i2})$ is the function of $\mathrm{D}(X_{i1}, X_{i2})$ where λ can be any values as all conjunction attribute values are identical in $\mathbf{P^{(0)}}$. Assuming λ is $|\mathbf{A^c}|$, which is m-p here.

After getting the auxiliary point $P^{(0)}$, starting the process of initializing k central points, which is mainly finding the t-th initial central point based on the dissimilarity measures from each points of remaining part to the previous acquired central points as well as the auxiliary point $P^{(0)}$.

Usually, the number of clusters number should be equal to the number of targeted class (k), which means t=1,2,…,k, and the method to find t-th initial point is to find the point whose average dissimilarity from itself to each of the previous acquired central points as well as the auxiliary point $P^{(0)}$ is the maximum. This step can make sure that the whole clustering process is based on k part as separately as possible.

However, when standard deviation of frequency in attribute A^{f}_{m+1} is larger than 1, it can be supposed that the distribution of all instance is decentralized. In this situation, it is likely to omit some important points which should have been initial central points during the next initialization process. To deal with this problem, method to find t-th initial point should turn into the way of finding the point whose average dissimilarity from itself to each of the previous acquired central points as well as the auxiliary point $P^{(0)}$ is the minimum. Apart from this modification, the auxiliary clusters also should be taken into account, which means we update k with the value of adding ($\log_{10} \lfloor \mathrm{n} \rceil$) to the original k, where $\lfloor \mathrm{x} \rceil$ means round value of the number of instances.

After each steps of finding one central point, we will find the set of points whose categorical conjunction attribute is same as that of the newly chose central point, and deduct it from the whole remaining instances. The deduction part is set as $\mathbf{P^{(t)}} = \{X_i \mid x^{f}_{i,m+1} = P^{(t)}_{m+1}\}$.

Therefore, the next steps for getting the t-th central points is set as following, where $t = 1, 2, ..., k-1$.

When $\sigma_{F_{m+1}} > 1$, $P^{(t)} = \arg\max_{X_{i'} \in \mathbf{X} - \sum_{t'=0}^{k-1}(\mathbf{P^{(t')}})} (\sum_{t=1}^{k-1} \mathrm{D}^2(P^{(t)}, X_{i'}))$, and k is the number of targeted classes.

When $\sigma_{F_{m+1}} \le 1$, $P^{(t)} = \arg\min_{X_{i'} \in \mathbf{X} - \sum_{t'=0}^{k-1}(\mathbf{P^{(t')}})} (\sum_{t=1}^{k-1} \mathrm{D}^2(P^{(t)}, X_{i'}))$, and k is the number of targeted class adding up the $\log_{10} \lfloor \mathrm{n} \rceil$.

And the all k used in the following steps are based on the choice deciding during this period.

$\mathrm{D}^2(X_{i1}, X_{i2})$ is the function of $\mathrm{D}(X_{i1}, X_{i2})$, where $\lambda / |\mathbf{A^c}| \ge 1$ as strengthening importance variation of categorical attributes themselves during the part of initialization process.The λ chose in the function $\mathrm{D}(X_{i1}, X_{i2})$ in next step of labeling is same as this step with same reason. For the suitable value of λ here, with consideration of strengthening and the level of mixed structure of categorical attributes over numeric attributes, we assume that λ is equal to $|\mathbf{A}^{\mathrm{c}}| \times \lfloor |\mathbf{A}^{\mathrm{c}}| / |\mathbf{A}^{r}| \rceil$, which is $(m-p) \times \lfloor (m-p)/p \rceil$. And it should have one upper limitation, which means, if the value is larger than m, turning the value of λ into m. The λ chose here is notated as λ_1, so does it in the next step of labeling.

4.2. *Finding the Labels for Remaining Instances*

After initialization and getting the set of k central points, $\mathbf{P} = \{P^{(t)} \mid t = 1, 2, ..., k\}$, we continue to grouped the whole data into k clusters by assigning label of cluster to each instance. For each instance, the labeling process is to find the nearest central point and record the label of this central point. Setting the vector of labels for each instance is $x^{l}_{i,m+2}$, it should be noted that when dealing with any instance of the set of k central points, the instance's label is the label of this central point itself. This whole period are showing as,

$$x^{l}_{i,m+2} = \begin{cases} \underset{t \in \{1,2,...,k\}}{\arg\min}(D^2(X_i, P^{(t)})), if\ X_i \notin \mathbf{P} \\ t, if\ X_i \in \mathbf{P},\ if\ X_i = P^{(t)} \end{cases} \tag{24}$$

Where,

$$i \in \{1, 2, ..., n\}\ and\ t \in \{1, 2, ..., k\}$$

Then making aggregation for instances with same label and getting k clusters with instances, shape as $Cl_t = \{X_i \mid x^{l}_{i,m+2} = t\}$.

4.3. *Locating New Central Points*

For each cluster, in process of locating new central point, the main algorithm is to seek the point nearest to the traditional focus point. The traditional focus point in the t-th cluster, $P^{(t)}\,'$, consists of the mean value of each numeric attribute and the mode element of each categorical attribute, written as,

$$\begin{gathered} P^{(t)}\,' = (\mathrm{M}(X^{(1)r}) \mid Cl_t, \mathrm{M}(X^{(2)r}) \mid Cl_t, ..., \mathrm{M}(X^{(p)r}) \mid Cl_t, \\ \mathrm{MO}(A^{c}_{p+1}) \mid Cl_t, \mathrm{MO}(A^{c}_{p+2}) \mid Cl_t, ..., \mathrm{MO}(A^{c}_{m}) \mid Cl_t) \end{gathered} \tag{25}$$

As for the measure of nearness, apart from the dissimilarity from each point to the traditional focus point in specific cluster, there is another factor added into the level of nearness, which is the sum of frequency of one instance spread out whole categorical attribute, showing as,

$$\begin{aligned} P^{(t)} &= \underset{X_{i'} \in Cl_t}{\arg\min} \left(\frac{D^3(P^{(t)}\,', X_{i'})}{\log_e(\sum_{j=p+1}^{m} \mathrm{f}((X_{i'})_j, j) \mid Cl_t)} \right) \\ &= \underset{X_{i'} \in Cl_t}{\arg\min}(D^3(P^{(t)}\,', X_{i'}) / \log_e(\sum_{j=p+1}^{m} \mathrm{f}((X_{i'})_j, j) \mid Cl_t)) \end{aligned} \tag{26}$$

Where, $(X_{i'})_j$ is the value of Attribute A^{c}_{j} in the instance $X_{i'}$.

It also should be pointed out that $\mathrm{D}^3(X_{i1}, X_{i2})$ is the function of $\mathrm{D}(X_{i1}, X_{i2})$ where $\lambda / |\mathbf{A}^{\mathrm{c}}| \leq 1$ as reducing importance variation of categorical attributes themselves during the part of recentering period. As for the suitable value for λ here, by analyzing the features of the function $\mathrm{Dc}(X_{i1}, X_{i2})$, only value far less than 1 can make effect of reducing. Therefore λ assumed here, notated as λ_2, is set as 1.

4.4. Iteration

Going back to the step of labeling, recentering, and making iterations. The whole algorithm will stop if the max iteration times set up has been reached up or the set of new central points is identical to the previous one. Afterward, output the set of central points and labeling vector of clustering result.

5. Experiments and Result

5.1. Six Datasets Descriptions

As for automobile dataset, we apply this algorithm to cluster Automobile dataset. After deleting missing values, this dataset contains 159 instances and 25 attributes, which can be split into 16 numeric attributes and 9 categorical attributes. The targeted classes' number of this dataset is 2, where class 1 with value 0 represents the auto is more risky than its price indicates, class 2 with value 1 represents the auto is more safe than its price indicates. This dataset is marked as 'A'.

As for cylinder Bands dataset, we apply this algorithm to cluster Cylinder Bands dataset. After deleting missing values, this dataset contains 277 instances and 39 attributes, which can be split into 20 numeric attributes and 19 categorical attributes. The targeted classes' number of this dataset is 2, where class 1 with value 0 represents band for whether cylinder banding existing, class 2 with value 1 represents band for whether cylinder banding existing. This dataset is marked as 'B'.

As for credit Approval dataset, we apply this algorithm to cluster Credit Approval dataset. After deleting missing values, this dataset contains 653 instances and 15 attributes, which can be split into 6 numeric attributes and 9 categorical attributes. The targeted classes' number of this dataset is 2, where class 1 with value 0 represents '+' for credit approval, class 2 with value 1 represents '-' for credit approval. This dataset is marked as 'C'.

As for flags dataset, we apply this algorithm to cluster Flags dataset. After deleting missing values, this dataset contains 194 instances and 28 attributes, which can be split into 10 numeric attributes and 18 categorical attributes. The targeted classes' number of this dataset is 8, where classes from 1 to 8 with value from 0 to 7 represent 8 religion classes (0=Catholic, 1=Other Christian, 2=Muslim, 3=Buddhist, 4=Hindu, 5=Ethnic, 6=Marxist, 7=Others), respectively. This dataset is marked as 'F'.

As for heart Disease dataset, we apply this algorithm to cluster Heart Disease dataset. After deleting missing values, this dataset contains 297 instances and 13 attributes, which can be split into 6 numeric attributes and 7 categorical attributes. The targeted classes' number of this dataset is 2, where class 1 with value 0 represents no presence of heart disease, class 2 with value 1 represents presence of heart disease. This dataset is marked as 'H'.

As for zoo dataset, we apply this algorithm to cluster Zoo dataset. After deleting missing values, this dataset contains 101 instances and 16 attributes, which can be split into 1 numeric attributes and 15 categorical attributes. The targeted classes' number of this dataset is 7, where classes from 1 to 7 with value from 0 to 6 represent seven animal types, respectively. This dataset is marked as 'Z'.

5.2. Evaluation Indicators for Different Clustering Algorithms

The most common measure is accuracy. The accuracy r is $r = (n\sum_{t=1}^{k} n_t)/n$, where n_t is the number of instances belonging to cluster t.

The second indicator is the average purity of clusters, which is $Pur = (\sum_{t=1}^{k}(|C_t^d| / |C_{t}|)) / k$,where $|C_t^d|$ is the number of instances whose dominant label is cluster t, and $|C_t|$ is that number of all instances in cluster t.

There are some other external evaluation indicators of clustering, Adjusted Rand Index (ARI) [21], Adjusted Mutual Information (AMI) [22], Homogeneity (HM) [23], Completeness (CM) [23] and V-Measure (VM) [23].

5.3. Results Comparing with Existing Basic Algorithms

By choosing the assuming values of λ_1 and λ_2 , we can get the whole evaluation indicators of the clustering result based on this set of parameters. Making comparisons to the whole evaluation indicators clustered by K-prototype Algorithm with initialization methods of 'I','C' and 'H'. Those comparison tables are showing as follow.

Table 1. Clustering Performance Evaluation Scores for Dataset of Automobile (A)

Algorithms	**Different Evaluation Scores**						
	r	**Pur**	**ARI**	**AMI**	**HM**	**CM**	**VM**
K-prototypes -C	0.653	0.650	0.070	0.068	0.064	0.093	0.069
K-prototypes -H	0.610	0.603	0.030	0.029	0.024	0.042	0.030
K-prototypes-I	0.652	0.659	0.081	0.075	0.071	0.096	0.077
KP-IW	**0.780**	**0.792**	**0.272**	**0.271**	**0.267**	**0.309**	**0.271**

Table 2. Clustering Performance Evaluation Scores for Dataset of Cylinder Bands (B)

Algorithms	**Different Evaluation Scores**						
	r	**Pur**	**ARI**	**AMI**	**HM**	**CM**	**VM**
K-prototypes -C	0.653	0.655	0.022	0.046	0.018	0.021	0.030
K-prototypes -H	0.663	0.653	0.023	0.049	0.019	0.030	0.031
K-prototypes-I	0.646	0.645	0.016	0.032	0.012	0.015	0.021
KP-IW	**0.661**	**0.629**	**0.027**	**0.054**	**0.023**	**0.054**	**0.036**

Table 3. Clustering Performance Evaluation Scores for Dataset of Credit Approval (C)

Algorithms	**Different Evaluation Scores**						
	r	**Pur**	**ARI**	**AMI**	**HM**	**CM**	**VM**
K-prototypes -C	0.718	0.722	0.190	0.190	0.189	0.244	0.190
K-prototypes -H	0.802	0.801	0.281	0.281	0.280	0.365	0.281
K-prototypes-I	0.754	0.762	0.221	0.213	0.212	0.274	0.217
KP-IW	**0.819**	**0.818**	**0.316**	**0.315**	**0.314**	**0.407**	**0.315**

Table 4. Clustering Performance Evaluation Scores for Dataset of Flags (F)

Algorithms	**Different Evaluation Scores**						
	r	**Pur**	**ARI**	**AMI**	**HM**	**CM**	**VM**
K-prototypes -C	0.488	0.543	0.215	0.261	0.146	0.081	0.236
K-prototypes -H	0.487	0.533	0.215	0.268	0.147	0.087	0.238
K-prototypes-I	0.477	0.532	0.204	0.248	0.135	0.082	0.224
KP-IW	**0.521**	**0.570**	**0.243**	**0.309**	**0.178**	**0.101**	**0.272**

Table 5. Clustering Performance Evaluation Scores for Dataset of Heart Disease (H)

Algorithms	**Different Evaluation Scores**						
	r	**Pur**	**ARI**	**AMI**	**HM**	**CM**	**VM**
K-prototypes -C	0.814	0.816	0.320	0.316	0.314	0.404	0.318
K-prototypes -H	0.825	0.826	0.332	0.328	0.327	0.420	0.330
K-prototypes-I	0.817	0.823	0.327	0.318	0.316	0.401	0.323
KP-IW	**0.825**	**0.825**	**0.330**	**0.328**	**0.326**	**0.420**	**0.329**

Table 6. Clustering Performance Evaluation Scores for Dataset of Zoo (Z)

Algorithms	**Different Evaluation Scores**						
	r	**Pur**	**ARI**	**AMI**	**HM**	**CM**	**VM**
K-prototypes -C	0.905	0.874	0.800	0.868	0.775	0.735	0.832
K-prototypes -H	0.881	0.859	0.728	0.810	0.695	0.619	0.767
K-prototypes-I	0.828	0.835	0.712	0.747	0.671	0.600	0.728
KP-IW	**0.921**	**0.894**	**0.884**	**0.891**	**0.869**	**0.902**	**0.888**

By analyzing above six tables, it can be concluded that out proposed algorithm KP-IW is rather appropriate than others based on most of those datasets and clustering evaluation measures

5.4. Results Comparing with Other Innovation Algorithms Appearing Recently

There are also other recently algorithms of clustering with mixed data, which have been introduced in the part of introductions. Selecting Algorithm whose experiments contain dataset existing in those six datasets and making comparisons, showing as follow,

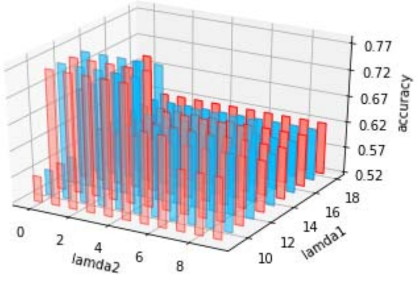
Table 7. Accuracy of clustering with KP-IW and other common algorithms

Algorithms	**Different Datasets**		
	C	**H**	**Z**
KP-IW	0.819	0.825	0.921
KL-FCM-GM	0.574	0.758	0.426
SBAC	0.555	0.752	--
EKP	0.682	0.545	0.901
WFK-prototype	0.838	0.835	0.908
SpectralCAT	0.770	0.820	0.930

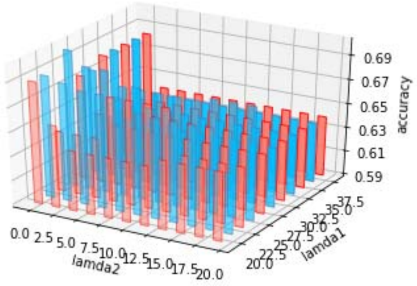
From the above TABLE VIII, it can be found that our proposed KP-IW algorithm is more effective than most of existing common algorithms of clustering for mixed data.

5.5. Comparing the Effect from Different Combinations Choices of lambda1 and lambda2

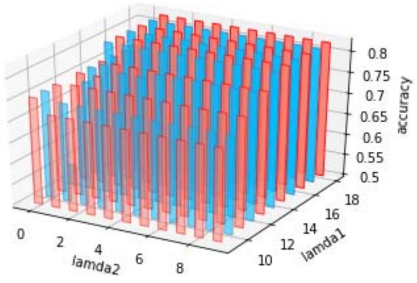
In order to verify whether the assumptions of λ_1 and λ_2 are suitable, we continue to assign more possible values to λ_1 and λ_2. For λ_1, we set it change from $1\cdot|\mathbf{A}^c|$ to $2\cdot|\mathbf{A}^c|$ to with a step of $0.1\cdot|\mathbf{A}^c|$, and for λ_2, we set it change from $0\cdot|\mathbf{A}^c|$ to $1\cdot|\mathbf{A}^c|$ with a step of $0.1\cdot|\mathbf{A}^c|$. The plots of (a), (b), (c), (d), (e) and (f) in Fig. 1 are to show how the exponential parameter sets $\{\lambda_1, \lambda_2\}$ influence the accuracy of clustering experiment results across those six datasets.


0.77
0.72
0.67
0.62
0.57
0.52
accuracy
0
2
4
6
8
lamda2
10
12
14
16
18
lamda1

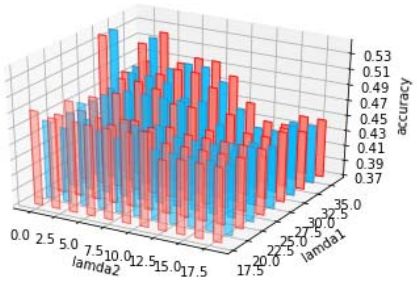
(a)


0.69
0.67
0.65
0.63
0.61
0.59
accuracy
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
lamda2
20.0
22.5
25.0
27.5
30.0
32.5
35.0
37.5
lamda1

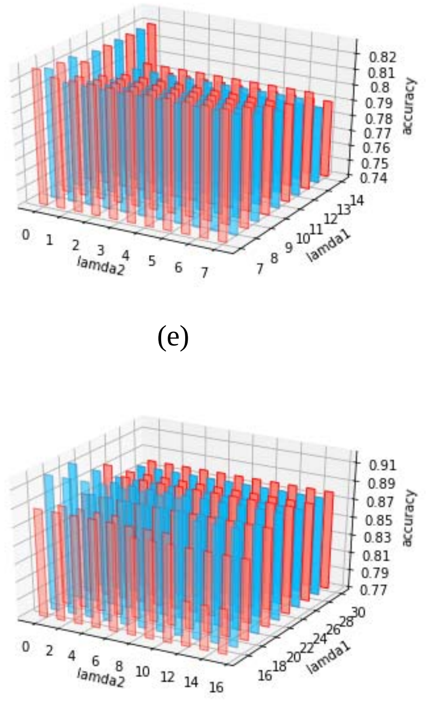
(b)


0.8
0.75
0.7
0.65
0.6
0.55
0.5
accuracy
0
2
4
6
8
lamda2
10
12
14
16
18
lamda1

(c)


0.53
0.51
0.49
0.47
0.45
0.43
0.41
0.39
0.37
accuracy
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
lamda2
17.5
20.0
22.5
25.0
27.5
30.0
32.5
35.0
lamda1

(d)



(e)

(f)

Figure 1. The impact of lambda1 and lambda2 on the accuracy(r) of clustering: (a) Automobile, (b)Cylinder Bands, (c)Credit Approval, (d)Flags, (e)Heart Disease, (f)Zoo

Table 8. Table 8 Accuracy(r) of clustering with different set of lambda in KP-IW

Lambda Setting	**Different Datasets**					
	A	**B**	**C**	**F**	**H**	**Z**
Score based on parameters from best result	0.78	0.704	0.819	0.541	0.828	0.921
Score based on assumed parameters	0.78	0.661	0.819	0.521	0.825	0.921

Comparing measure scores based on λ_1 and λ_2 from the best result with that from the above assumptions, we can find that there is no significant improvements for most of those datasets by taking more choice of λ_1 and λ_2 and selecting the best result, which verifies that the above assumptions about how to choose λ_1 and λ_2 is appropriate.

6. Conclusions

In this paragraph, we present the KP-IW algorithm which innovates in initializations process and attribute weighting.

During the initialization process, the innovations consist of choosing appropriate method of finding next central points, the utilization of auxiliary point and auxiliary clusters.

As for weighting of attribute significance, we add exponential weighting to the traditional linear weighting. The whole algorithm is controlling by several parameters based on the features of target data itself, which can cluster more effectively and avoid relying on parameter turning too much.

There are two main processes of comparing. The former one contrasts evaluation of clustering result from KP-IW based on hypothetical weighting parameters with that from utilizing various tempt of parameters to the established KP-IW algorithm, which can verified that the guideline of choosing weighting is appropriate. The latter one is focus on the comparison of evaluations of clustering result from KP-IW with other existing algorithms suitable for mixed data, which is to show that the feasibility and superior performance can be demonstrated by the compared results from KP-IW algorithm together with others.

References

[1] J. Han and M. Kamber, Data Mining Concepts and Techniques,3rd ed, USA:Morgan Kaufmann, 2001, pp. 490-491

[2] S. P. Lloyd, “Least squares quantization in PCM,” IEEE Transactions on Information Theory, vol. 28, pp.129–137, 1982.

[3] T. Zhang, R. Ramakrishnan, andM. Livny, “BIRCH: an efficient data clustering method for very large databases,” in Proceedings of the ACMSIGMOD International Conference onManagement of Data, ACM.Montreal.Canada, June 1996, pp. 103–114.

[4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD ’96), Portland.Ore.USA, August 1996.

[5] Z. Huang, “A fast clustering algorithm to cluster very large categorical data sets in data mining,” in Research Issues on Data Mining and Knowledge Discovery,Tuscon.Ariz.USA: ACM Press, 1997, pp. 1–8.

[6] F. Cao and J. Liang, “A new initialization method for categorical data clustering,” in Expert Systems with Applications, vol. 36, pp. 10223–10228–433, 2009.

[7] Z. Huang, “Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values,” in Data Mining and Knowledge Discovery, vol. 2, pp. 283–304, 1998.

[8] Z. Huang, “Clustering large data sets with mixed numeric and categorical values,” in Proceedings of the the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining, World Scientific Publishing.Singapore, 1997, pp. 21–34.

[9] Z. Huang and M. K.Ng, “Afuzzy k-modes algorithmfor clustering categorical data,” IEEE Transactions on Fuzzy Systems, vol.7, pp. 446–452, 1999.

[10] A. Saha and S. Das, “Categorical fuzzy k-modes clustering with automated feature weight learning,” Neurocomputing, vol. 166,pp. 422–435, 2015.

[11] M. Lee and W. Pedrycz, “The fuzzy C-means algorithm with fuzzy P-modeproto types for clustering objects having mixed features,” in Fuzzy Sets and Systems, vol. 160, pp. 3590–3600, 2009.

[12] S. P. Chatzis, “A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional,” Expert Systems with Applications, vol. 38, pp. 8684–8689, 2011.

[13] J. Ji, W. Pang, C. Zhou, X. Han, and Z. Wang, “A fuzzy kprototype clustering algorithm for mixed numeric and categorical data,” Knowledge-Based Systems, vol. 30, pp. 129–135, 2012.

[14] J. Huang, M. K. Ng, H. Rong, and Z. Li, “Automated Variable Weighting in k-Means Type Clustering,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, pp. 656-668, May 2005.

[15] J. Ji, T. Bai, C. Zhou, C. Ma, and Z.Wang, “An improved k-prototypes clustering algorithm for mixed numeric and categorical data,” in Neurocomputing, vol. 120, pp. 590–596, 2013.

[16] C. Li and G. Biswas, “Unsupervised learning with mixed numeric and nominal data,” IEEE Transactions on Knowledge and Data Engineering, vol. 14, pp. 673–690, 2002.

[17] D. W. Goodall, “A new similarity index based on probability,”Biometrics, vol. 22, pp. 882–907,

1966.

[18] Z. Zheng, M. Gong, J. Ma, L. Jiao, and Q. Wu, “Unsupervised evolutionary clustering algorithm for mixed type data,” in Proceedings of the IEEE Congress on Evolutionary Computation, Barcelona,.Spain, 2010, pp. 1–8.

[19] G. David and A. Averbuch, “SpectralCAT: categorical spectral clustering of numerical and nominal data,” in Pattern Recognition,vol. 45, pp. 416–433, 2012.

[20] K. C. Gowda and E. Diday, “Symbolic clustering using a new dissimilarity measure,” in Pattern Recognition, vol. 24, pp. 567–578, 1991.

[21] L. hubert and P. Arabie, “Comparing partitions,” in Journal of Classification, vol. 2, pp. 193–218, December 1985.

[22] N. X. Vinh, J.Epps, and J. Bailey, “Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance,” in Journal of Machine Learning Research, vol. 11,pp. 2837-2854, 2010

[23] A. Rosenberg and J. Hirschberg, “V-Measure: A conditional entropy-based external cluster evaluation measure,” in Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, June 2007, pp. 410–420