A New Partition-based Clustering Algorithm For Mixed Data

ZHONG Xian, YU TianBao, and XIA HongXia

***Abstract*—In practical application field, it is common to see the mixed data containing both the numerical attributes and categorical attributes simultaneously. However, most of the given clustering algorithms can only deal with data in single type, either numerical or categorical. Therefore we present a partition-based clustering algorithm for mixed data. First, the multi-Modes representation means of category centres of categorical data in K-prototypes algorithm is given, and then the Euclidean is expanded to mixed data so as to reflect the dissimilarity between the objects and categories more accurately in same framework. On the basis, a partitioned clustering algorithm for processing mixed data is proposed. Finally, the experimental results on UCI dataset show that the proposed algorithm can effectively improve the clustering performance comparing with the K-prototypes algorithm.**

***Index Terms*—K-Prototypes algorithm, Mixed data, Partitioned clustering, Dissimilarity measure.**

I. Introduction

AS an unsupervised learning method, clustering analysis is an important content of research in machine learning and data mining application. According to the thought of "Like attracts like.", clustering analysis is based on certain rules which divide the data objects into a plurality of class. The similarity is high so that the objects seem in the same class, But differences between the larger object are not similar. It has been widely applied in image processing[1], information retrieval[2], bioinformatics[3] and social network analysis and other research fields[4], [5]. In recent years, according to different application of clustering analysis, researchers carried out a series of studies, and put forward some effective clustering algorithm, including partition clustering[6], hierarchical clustering[7], [8], density-based clustering[9], network-based [10]and model based clustering algorithm[11], [12]. However, most of the existing clustering algorithm is used only for data in a single type effectively,data will become invalid in the mixed type. Here the single type of data include either numerical or categorical data. The mixed data exist universally in real life. For example, numerical data attributes to reflect the age, height and weight, while categorical data stand for ethnic, gender, blood type in the description of population datasets. Therefore the design of a good clustering algorithm mixed data is a challenging problem that should be confronted with. To overcome the difficulty, some scholars have performed a depth exploration. For instance, by integrating the K-Means[13], [14], [15] algorithm and K-Modes algorithm [16], [17] directly together, and extending to the K-Prototypes algorithm[18], Huang and others succeeded to solve the clustering problem of mixed data in 1998. In 2002, the literature[19] proposed a hierarchical agglomerative clustering algorithm SBAC (Similarity Based Agglomerative Clustering). The similarity of Good all is used to measure the similarity between object relation and class relation by SBAC algorithm. Among them, attribute weights which do not often occur, are assigned large values by similarity measure. For a numeric attribute, similarity measure not only contain the value of the difference in size, but also contain the unique value to appear. In 2010, the evolution of the k - prototype algorithm[20] is proposedEKP by introducing an evolutionary algorithm framework and it makes the algorithm own global search ability. In 2011, k - the prototype algorithm of extended algorithm KL - FCM - GM[21] is proposed. In 2012, fuzzy k - prototype algorithm is proposed[22] . On the one hand, this algorithm can keep the data set inherent uncertainty a longer time before the determination data which belongs to the prototype. On the other hand, the effect of different dimensions of data on the clustering process is fully considered. Because of its high efficiency and simple implement, partitional clustering algorithm, represented by K-Proto types, has been widely applied. However, dissimilarity measures between numerical attributes and categorical attributes are not uniformed and there exists difficulties of weighting factor selections in K-prototypes algorithm. Therefore, a partitioned clustering algorithm for processing mixed data, which is based on Kprototypes algorithm, is particularly necessary to appear.Therefore, in this paper, we firstly propose the multi Modes representation means of category centres of categorical data in K-prototypes algorithm, and then the Euclidean is expanded to mixed data so as to reflect the dissimilarity between the objects and categories. On the basis, a partitioned clustering algorithm for processing mixed data is proposed. Finally, the experimental results on UCI dataset show that the proposed algorithm can improve the clustering performance effectively and accurately.

Manuscript received 22, January, 2017; revised 20, February, 2017. This work is supported by National Key Technologies R D Program of China(2012BAH33F03), National Natural Science Foundation of China(61303029), National Natural Science Foundation of China for Young Scholar(61003130),Science and Technology innovation team project of Wuhan (201307020402005).

ZHONG Xian (Email:23668229@qq.com), YU TianBao (corresponding author, Email:ytb1992@whut.edu.cn) and XIA HongXia (Email:xiahx@whut.edu.cn) are with the Department of Software Engineering, School of Computer Science and Technology, WuHan University of Technology, HuBei, 430070 CHINA .

II. K-PROTOTYPES ALGORITHM

Using simple matching to measure the dissimilarity based on category attributes, K-prototypes algorithm[23] is on the basis of K-Means algorithm. So it appears in order to resolve the mixed data clustering problem.

Given a data set $X = \{X_1, X_2, \ldots, X_n\}$, in which we call $X_{\mathrm{i}}(i = 1, 2, \ldots n)$ data points. Here $X_i(1 \leq i \leq n)$ are described by eigenvalues $A_1, A_2, \cdots, A_p, A_{p+1}, \cdots, A_m$. At the same time, let $A_i(1 \leq i \leq p)$ be numeri-

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

cal attributes, and $A_i(p+1 \leq i \leq m)$ be categorical attributes. Here Dom(A_j) means the range of Attribute A_j. Let Dom(A_j)= $\left\{a_j{}^{(1)}, a_j{}^{(2)}, \ldots, a_j{}^{(n_j)}\right\}$, where n_j denotes the number of attribute values of A_j domain for Categorical Data. Object X_i which belongs to X, can be marked by a m-dimension vector, that is $X_i = \{xi1, xi2, \ldots, xip, xi\,(p+1), xi\,(p+2), \ldots, xim\}$. Here $x_{ij}\ Dom(A_j)$, the clustering center is denoted as Z, abbreviated as Z_j (z_{l1}, z_{l2}, …, z_{lm}).

When talking about dissimilarity measures between objects, both numerical attributes and categorical attributes are considerated by the K-Prototypes algorithm. So parameters appears to control how much are contributed by numerical attributes and categorical attributes.

Definition 1 Given a dataset C_l, which belongs to X, is a class in the K-Prototypes algorithm. Z_l is the class center of C_{lh}. The distance metric between Object X and Class center Z_l is defined as follows: $d\,(X_i, Z_l) = d_r\,(X_i, Z_l) + \gamma d_c\,(X_i, Z_l)$

Here it is made of two parts: data attribute and categorical attribute, D_r and d_c represents the dissimilarity between the object and class center, characterized by numeric and categorical attributes respectively. Dr represents the Euclidean distance $d_r(X_i, Z_l) = \sum_{j=1}^{p} (x_{ij} - z_{lj})^2$ d_c denotes simple matching to measure the dissimilarity $d_c(X_i, Z_l) = \sum_{j=p+1}^{m} \delta(x_{ij} - z_{lj})$ Here $\delta(x_{ij}, z_{lj}) = \begin{cases} 1 x_{ij} \neq z_{lj} \\ 0 x_{ij} = z_{lj} \end{cases}$
The minimizing target function is set as
$F(W, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{li} d(X_i, Z_l)$
$w_{li} \in \{0, 1\}, 1 \leq l \leq k, 1 \leq i \leq n; \sum_{i=1}^{k} w_{li} = 1, 1 \leq i \leq n; 0 < \sum_{i=1}^{n} w_{li} < n, 1 \leq l \leq k$

$w_{li} = 1$ indicates the i-th object belongs to the l-th class, and $w_{li} = 0$ says that the i-th object does not belong to the l-th class.

In order to achieve the minimum objective function F under the restrained conditions, The basic steps of K-Prototypes algorithm are as follows:

Step1 Let's select k objects randomly as the initial cluster center in dataset X;

Step2 According to the Definition 1., calculate the distance between the centers of objects and classes, and allocate each object to the nearest class.

Step3 Update the clustering center, among which data attribute is calculated by averaging the value of the same class and part of the frequency that the property value appears is computed to describe categorical attribute in clustering Algorithms;

Step4 Repeat Step2, Step3, never stop until the objective function F no longer changes.

III. A PARTITION-BASED CLUSTERING ALGORITHM FOR MIXED DATA

Achieving an effective clustering of mixed data, algorithm based on the K-Prototypes still has problems urgently needed to solve in the clustering process.

K-Prototypes clustering algorithms is made up of numerical attributes and categorical attributes, among which data attribute is calculated by averaging the value of the same class. and part of the frequency that the property value appears is computed to describe categorical attribute in clustering Algorithms; However, the category centres of categorical data in K-prototypes algorithm, failed to adequately take the effect that other attribute values with non-highest frequency on the clustering center of attribute values into account. It is difficult to accurately reflect the value of objects in the class. And when a property value with highest frequency is more than one, Mode will not be unique. Choosing a different Mode to calculate the dissimilarity measure may get exactly the opposite conclusion, which led to the algorithm unstable.

Secondly, Using simple matching to measure the dissimilarity based on category attributes, (object with the same class center dissimilarity is 0; otherwise, the dissimilarity 1)of objects and class center. This calculation does not accurately confirm the similarity of objects and the other samples in the corresponding class, and the situation that whether the object is to join a class, That not only depends on the difference between the object and the prototype, but also depends on the overall differences between objects and the existing objects in the class. And when one object and the center of the object with multiple classes are in the same degree of difference, K-Prototypes algorithm often join the objects this into a class randomly, they always can not be accurately divided into classes with greater degree of similarity. Although dissimilarity measure integrates Euclidean and simple matching metric, the distance metric mixed data, obtained by a simple accumulation, fails to accurately measure the dissimilarity values and classification of the data portion of the data portion in a unified framework. What's worse, the selection of parameter which controls how much is contributed is a very difficult problem in practical applications when accumulating the dissimilarity of different types of data.

For the above problems, this section first gives a categorical attribute class that represents the central part of the multi Modes way, and expand the traditional Euclidean distance to the mixed data, to reflect the dissimilarity between the objects and categories. Finally, the experimental results show that the proposed algorithm can improve the clustering performance effectively and accurately at the same frame. Furthermore, we present a partition-based clustering algorithm for mixed data.

Definition 2 C_1 represents a class obtained by dataset X in the clustering process, the representation of categorical attribute part of the multi-class center Modes is defined as:

$Z_l^c = (z_{l(p+1)}, z_{l(p+2)}, \ldots, z_{lm})$

Among this,$z_{lj} = \{(a_j^{(1)}, f_{lj1}), (a_j^{(2)}, f_{lj2}), \ldots, (a_j^{(n_j)}, f_{ljn_j})\}$, $p+1 \leq j \leq m$

n_j illustrates the number of different value in numerical range under the j-th classification attribute values; $f_{ljw} (1 \leq w \leq n_j)$ illustrate the occurrence frequency of aw in Class C in numerical range under the j-th classification attribute values.

In order to reveal the categorical attribute class that represents the central part of the multi Modes way more intuitively, we explain these results as follows:

Example 1: Assume that the data shown in Table 1. represent the categorical attribute in the process of clustering. There are $\{X_1, X_2, X_3, X_4, X_5\}$ 5 objects and $\{A_1, A_2, A_3\}$ 3 classification properties.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Table 1. A clustering result data table

	A_1	A_2	A_3
X_1	A	d	f
X_2	A	e	g
X_3	B	e	f
X_4	C	d	f
X_5	C	d	g

According to the calculation method of the categorical attribute class that represents the central part of the multi Modes way in the K-Prototypes algorithm, data shown in Table 1 mean that the Modes of Class Cl is $\{a, d, f\}$ or $\{c, d, f\}$. So there is no unique nature here, it will likely lead to an unstable algorithm. But based on the Definition 2., the more Modes for class center of this class is $Z_l^c = (z_{11}, z_{12}, z_{13})$,

Here $z_{11} = \{(a, 0, 4), (b, 0, 2), (c, 0, 4)\}$, $z_{12} = \{(d, 0, 6), (e, 0, 4)\}$, $z_{13} = \{(f, 0, 6), (g, 0, 2)\}$.

According to the Definition 2,a singleton object may also be expressed as the form of multi-Modes, such as categorical data objects $\mathrm{X}_6 = (c, d, f)$, the form of multi-Modes is $\mathrm{X}_6^{'} = (x_{61}^{'}, x_{62}^{'}, x_{63}^{'})$, $x_{61}^{'}$= $\{(\mathrm{a}, 0), (\mathrm{b}, 0), (\mathrm{c}, 1)\}$, $x_{62}^{'}$= $\{(\mathrm{d}, 1), (\mathrm{e}, 0)\}$ and $x_{63}^{'}$= $\{(\mathrm{f}, 1), (\mathrm{g}, 0)\}$.

According to the central part of the multi Modes representation in the categorical attribute class, the extended Euclidean distance metric is as follows.

Definition 3 Given $\mathrm{Z}_i^c = (z_{i(p+1)}, z_{i(p+2)}, \ldots, z_{im})$ and $Z_j^c = (z_{j(p+1)}, z_{j(p+2)}, \ldots, z_{jm})$ as two class centers expressed by multi Modes in categorical attributes for data sets X and here $D_c(Z_i^c, Z_j^c) = \sum_{t=p+1}^{m} \frac{1}{n_t} \sum_{s=1}^{n_t} (f_{its} - f_{jts})^2$ is defined as the Euclidean distance between Z_i^c and Z_j^c.

Definition 4 Let $\mathrm{C}_l \subset X$ be a clustering result in the process of clustering algorithm, and $\mathrm{Z}_l^{'} = \{Z_l^r, Z_l^c\}$ be the class center of C_l, where Z_l^r is the mean value of the center of the class in categorical attributes and Z_l^c is the class center expressed by multi Modes in categorical attributes defined by Definition 2. The definition of Euclidean distance between the object $\mathrm{D}(X_i, Z_l^{'}) = \mathrm{D}_r(X_i, Z_l^r) + \mathrm{D}_c(X_i^{'}, Z_l^c)$

Where $\mathrm{D}_r(X_i, Z_l^r)$ represents numerical attributes part of the Euclidean distance, namely; $\mathrm{D}_r(X_i, Z_l^r) = \sum_{i=1}^{p} (x_{is} - z_{ls})^2$ $\mathrm{D}_c(X_i^{'}, Z_l^c)$ represents the Euclidean distance extended by multi Modes centres in categorical attributes defined by Definition 3. $X_i^{'}$ in the form of objects X_i that represent more modes.

Let X be a hybrid data sets, and apply the extended Euclidean distance given above to K-Prototypes algorithm, the objective function is defined as:

$$\mathrm{F}^{'}(\mathrm{W}, \mathrm{Z}^{'}) = \sum_{l=1}^{k} \sum_{i=1}^{n} \mathrm{w}_{li} \mathrm{D}(X_i, Z_l^{'})$$

$w_{li} \in \{0, 1\}, 1 \leq l \leq k, 1 \leq i \leq n; \sum_{i=1}^{k} w_{li} = 1, 1 \leq i \leq n; 0 < \sum_{i=1}^{n} w_{li} < n, 1 \leq l \leq k$

The optimization problem above is a very complex nonlinear programming problem, the stepwise optimization strategy is that: first fixed cluster centers $Z^{'}$, minimizing the objective function $F^{'}$ to get attached to the matrix W; then fixed membership matrix W, minimizing the objective function $F^{'}$ to get a new cluster center $Z^{'}$; so iterations, never stop until the objective function can not continue to optimize, which is similar to the K-Means algorithm.

Based on the center of the class expressed by multi Modes and extended Euclidean distance metric based on hybrid data partition. a partition-based clustering algorithm for mixed data are described below.

Step1 Let's select k objects randomly as the initial cluster center in dataset X;

Step2 According to the Definition 4, calculate the dissimilarity between the objects and centers of classes, with reference to the extended Euclidean distance metric, and according to the minimization principle divide the data set into the cluster represented by its nearest cluster center;

Step3 Update the clustering center,among which data attribute is calculated by averaging the value of the same class, and the center of classes expressed by multi Modes is for Categorical Attributes according to the Definition 4;

Step4 Repeat Step2, Step3, never stop until the objective $F^{'}$ function no longer changes.

IV. EXPERIMENTAL ANALYSIS

To validate the effectiveness and feasibility of the proposed algorithm,we made a centralized selection of four data sets from UCI real data:Teaching assistant evaluation (TAE), Heart disease (Heart), Australian Credit approval (Credit) and Contraceptive method choice (CMC), and compare the algorithm proposed in this paper(IK-P) and K-Prototypes algorithm(K-P). 4 data sets are described in Table 2 below.

Table 2. Description of the data sets

datasets	objects	numerical attribute	Categorical attribute	class
TAE	151	1	4	3
Heart	303	5	8	2
Credit	690	6	8	2
CMC	1473	2	7	3

In order to assess the effectiveness of the algorithm, accuracy, precision, recall and adjusted rand index [23] four indicators use to evaluate clustering results. X is a set of data set consisting of n objects.

The clustering result is $\mathrm{C} = \{C_1, C_2, \ldots \mathrm{C}_k\}$, and the real division is $\mathrm{P} = \{P_1, P_2, \ldots \mathrm{P}_{k'}\}$. a_i represents the correct number of objects assigned to the i-th class, b_i represents the number of objects mistakenly assigned to the i-th class, c_i represents the object which should be assigned to class i was not assigned to a number of objects, and n_{ij} the number of objects is included between C_i and P_i in common. AC (accuracy), PR (precision), RE (recall) and ARI (adjusted rand index) 4 indicators are defined as follows:

$$AC = \frac{1}{n} \sum_{i=1}^{k} a_i$$

$$PE = \frac{1}{k} \sum_{i=1}^{k} \frac{a_i}{a_i + b_i}$$

$$RE = \frac{1}{k} \sum_{i=1}^{k} \frac{a_i}{a_i + c_i}$$

$$ARI = \frac{\sum_{\mathrm{ij}} \binom{n_{ij}}{2} - E_i D_j \Big/ \binom{n}{2}}{\frac{1}{2}(E_i + D_j) - E_i D_j \Big/ \binom{n}{2}}$$

(Where $E_i = \sum_{\mathrm{i}} \binom{e_i}{2}$, $D_i = \sum_{\mathrm{j}} \binom{d_j}{2}$)

If the clustering results are more close to the real classification of the data set then AC, PE, RE and ARI values is bigger.

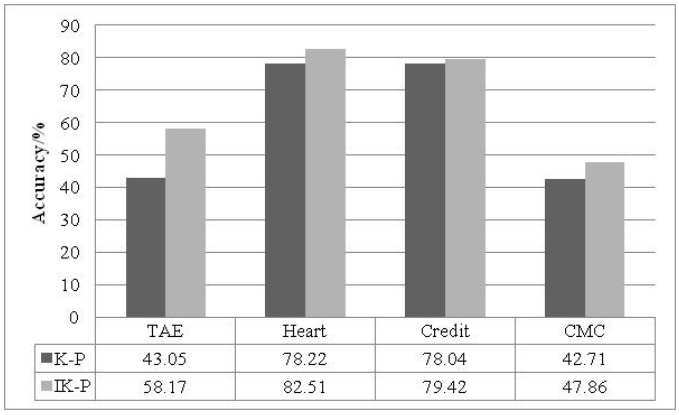
In the experiment, in order to avoid the influence that features in different numerical value turn out different distance calculation, the standardization of numeric data is needed before before clustering. The standard formula is:

$$\mathrm{X}_{\mathrm{ij}}^{'} = \frac{X_{ij} - \min(X_{\cdot j})}{\max(X_{\cdot j}) - \min(X_{\cdot j})}$$

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

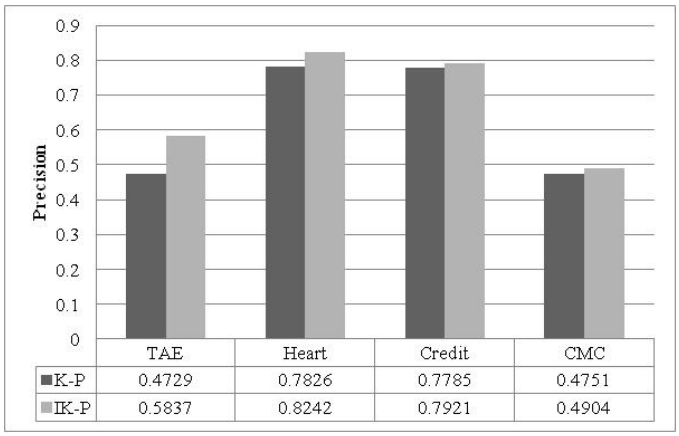
Among this, $1 \leq i \leq n, 1 \leq j \leq p$, X_{ij} means the value of the I data in the j attribute X'_{ij} is the standardized value after X_{ij}. $\max(X_{.j})$ and $\min(X_{.j})$ are the representation of the whole dataset the j properties of the minimum and maximum values respectively.

Due to the effect of choice of the initial cluster centers, cluster partition clustering algorithm may have different clustering results, so select 100 groups of class center in the 4 data sets randomly. Then each algorithm may run 100 times. We can validate the effectiveness of the algorithm by calculating the average clustering quality. On 4 datasets, clustering performance of the proposed algorithms(IK-P) and the K-Prototypes algorithms(K-P) in different evaluation indexes were compared in Fig.1 to Fig.4 as follows:



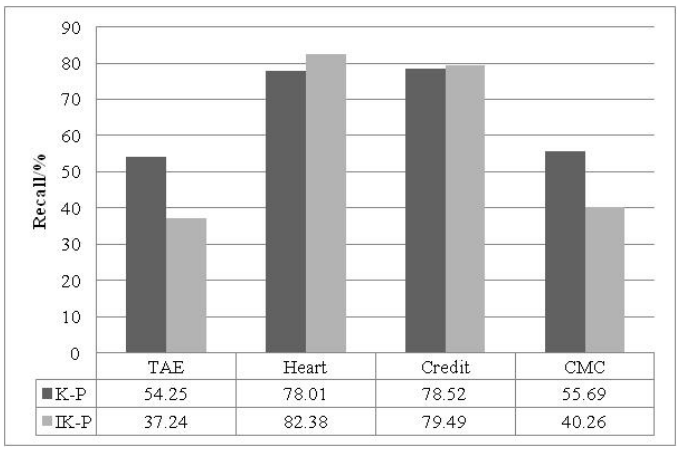
	TAE	Heart	Credit	CMC
K-P	43.05	78.22	78.04	42.71
IK-P	58.17	82.51	79.42	47.86

Fig. 1. Algorithm accuracy of different data sets



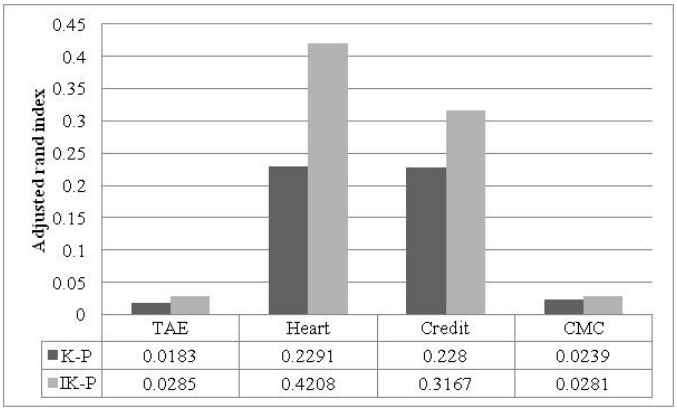
	TAE	Heart	Credit	CMC
K-P	0.4729	0.7826	0.7785	0.4751
IK-P	0.5837	0.8242	0.7921	0.4904

Fig. 2. Algorithm purity of different data sets



	TAE	Heart	Credit	CMC
K-P	54.25	78.01	78.52	55.69
IK-P	37.24	82.38	79.49	40.26

Fig. 3. Algorithm recall rate of different data sets

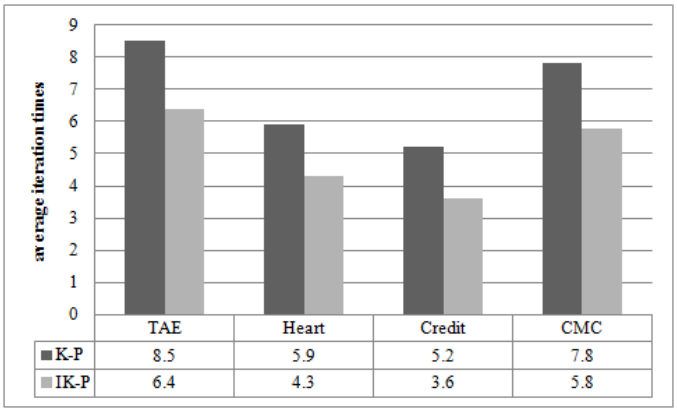


	TAE	Heart	Credit	CMC
K-P	0.0183	0.2291	0.228	0.0239
IK-P	0.0285	0.4208	0.3167	0.0281

Fig. 4. Algorithm Teresa index of different data sets

According to the datasets Teaching assistant evaluation Heart disease, Australian Credit approval and Contraceptive method choice, this paper proposed clustering algorithm based on partition, which is better than traditional K-Prototypes algorithm on the whole, and get better clustering effect by analyzing Fig.1 to Fig.4.

To further analyze the effectiveness of the algorithm proposed in this paper, we compare the IK-P algorithm with the traditional K-Prototypes algorithm on 4 data sets which are mentioned above from the aspect of the number of iterations, as shown in Fig.5. The result in Fig.5 is the average result of 100 times experiments which are separately done on the 4 data sets by using the two algorithms, and the class centers are randomly selected.



	TAE	Heart	Credit	CMC
K-P	8.5	5.9	5.2	7.8
IK-P	6.4	4.3	3.6	5.8

Fig. 5. The average number of iterations of different data sets

As can be seen from Fig.5, the number of iterations partition-based clustering algorithm on the 4 data sets proposed in this chapter is far less than the K-Prototypes algorithm. It can be known that partition-based clustering algorithm proposed in this chapter also has some advantages in aspect of the number of iterations.

These results show that the representation of in this paper class centers expressed by multi Modes in categorical attributes and extended Euclidean distance are valid.

V. Conclusion and Future Work

In order to solve the traditional K-Prototypes algorithm problems, and apply for the mixed data containing both numerical and categorical attributes. Firstly the representation of class centers expressed by multi-Modes in categorical attributes are proposed and this will be extended based on the Euclidean distance to the mixed data, and then

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

design a partition-based clustering algorithm for mixed data. Compared to the traditional K-Prototypes algorithm, with a series of experiments, K-Prototypes is superior to traditional clustering algorithm on the overall proposed and have more effective result for the actual data.

REFERENCES

[1] M. Gong, Y. Liang, J. Shi, W. Ma, and J. Ma, "Fuzzy c-means clustering with local information and kernel metric for image segmentation," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 22, no. 2, pp. 573–584, 2013.

[2] X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, and J. Han, "Cluscite: effective citation recommendation by information network-based clustering," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 821–830.

[3] Y. Si, P. Liu, P. Li, and T. P. Brutnell, "Model-based clustering for rna-seq data." *Bioinformatics*, vol. 30, no. 2, pp. 197–205, 2014.

[4] D. W. R. Xu, "Survey of clustering algorithms for manet," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, p. 645678, 2005.

[5] J. G. Sun, J. Liu, and L. Y. Zhao, "Clustering algorithms research," *Journal of Software*, vol. 19, no. 19, 2008.

[6] R. T. Aldahdooh and W. Ashour, "Dimk-means distance-based initialization method for k-means clustering algorithm," *International Journal of Intelligent Systems Applications*, vol. 5, no. 2074-904X, pp. 41–51, 2013.

[7] Y. Malitsky, A. Sabharwal, H. Samulowitz, and M. Sellmann, "Algorithm portfolios based on cost-sensitive hierarchical clustering," 2013.

[8] S. Zahra, M. A. Ghazanfar, A. Khalid, M. A. Azam, U. Naeem, and A. Prugel-Bennett, "Novel centroid selection approaches for kmeans-clustering based recommender systems," *Information Sciences*, vol. 320, no. C, pp. 156–189, 2015.

[9] Y. Kim, K. Shim, M. S. Kim, and J. S. Lee, "Dbcure-mr: An efficient density-based clustering algorithm for large data using mapreduce," *Information Systems*, vol. 42, no. 2, pp. 15–35, 2013.

[10] S. Wang and Y. Chen, "Hasta: A hierarchical-grid clustering algorithm with data field," *International Journal of Data Warehousing Mining*, vol. 10, no. 2, pp. 39–54, 2014.

[11] J. Jacques and C. Preda, "Model-based clustering for multivariate functional data," *Computational Statistics Data Analysis*, vol. 71, no. 3, pp. 92–106, 2014.

[12] J. Y. Chen and H. H. He, "A fast density-based data stream clustering algorithm with cluster centers self-determined for mixed data," *Information Sciences An International Journal*, vol. 345, no. C, pp. 271–293, 2016.

[13] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[14] M. S. Yang and Y. C. Tian, "Bias-correction fuzzy clustering algorithms," *Information Sciences*, vol. 309, pp. 138–162, 2015.

[15] A. Saha and S. Das, "Categorical fuzzy k -modes clustering with automated feature weight learning," *Neurocomputing*, vol. 166, no. C, pp. 422–435, 2015.

[16] J. Y. Chen and H. H. He, "Research on density-based clustering algorithm for mixed data with determine cluster centers automatically," *Acta Automatica Sinica*, 2002.

[17] Z. Huang, "A fast clustering algorithm to cluster very large categorical data sets in data mining," *Research Issues on Data Mining Knowledge Discovery*, pp. 1–8, 1998.

[18] ——, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[19] C. Li and G. Biswas, "Unsupervised learning with mixed numeric and nominal data," *IEEE Transactions on Knowledge Data Engineering*, vol. 14, no. 4, pp. 673–690, 2002.

[20] Z. Zheng, M. Gong, J. Ma, and L. Jiao, "Unsupervised evolutionary clustering algorithm for mixed type data," in *IEEE Congress on Evolutionary Computation, Cec 2010, Barcelona, Spain, 18-23 July*, 2010, pp. 1–8.

[21] 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 An International Journal*, vol. 38, no. 7, pp. 8684–8689, 2011.

[22] J. Ji, W. Pang, C. Zhou, X. Han, and Z. Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data," *Neurocomputing*, vol. 120, no. 10, pp. 590–596, 2013.

[23] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)