Knowledge-Based Systems 30 (2012) 129–135

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data

Jinchao Ji [a], Wei Pang [a,b], Chunguang Zhou [a], Xiao Han [c], Zhe Wang [a,*]

[a] *College of Computer Science and Technology, Jilin University, Changchun 130012, China*
[b] *School of Natural and Computing Sciences, University of Aberdeen, Aberdeen, AB24 3UE, UK*
[c] *College of Mathematics, Jilin University, Changchun 130012, China*

ARTICLE INFO

Article history:
Received 27 April 2011
Received in revised form 6 January 2012
Accepted 6 January 2012
Available online 21 January 2012

Keywords:
Fuzzy clustering
Data mining
Mixed data
Dissimilarity measure
Attribute significance

ABSTRACT

In many applications, data objects are described by both numeric and categorical features. The k-prototype algorithm is one of the most important algorithms for clustering this type of data. However, this method performs hard partition, which may lead to misclassification for the data objects in the boundaries of regions, and the dissimilarity measure only uses the user-given parameter for adjusting the significance of attribute. In this paper, first, we combine mean and fuzzy centroid to represent the prototype of a cluster, and employ a new measure based on co-occurrence of values to evaluate the dissimilarity between data objects and prototypes of clusters. This measure also takes into account the significance of different attributes towards the clustering process. Then we present our algorithm for clustering mixed data. Finally, the performance of the proposed method is demonstrated by a series of experiments on four real world datasets in comparison with that of traditional clustering algorithms.

1. Introduction

The widespread use of information technology has resulted in the generation of data at a phenomenal rate in all areas including business, manufacturing, and healthcare. This explosive growth in stored data has generated an urgent need for new techniques that can transform the vast amount of data into useful information and knowledge. Data mining is, perhaps, the most suitable technique to satisfy this need [11,12,18].

In data mining, clustering is one of the most important techniques. The clustering can be used in many fields such as privacy preserving [21], information retrieval [7] and text analysis [29]. The objective of clustering is to partition a set of data objects into clusters such that data objects in the same cluster are more similar to each other than those in other clusters [15,16,22,23]. Usually, data sets to be mined contain both numeric and categorical attributes [19]. However, most existing algorithms are limited to one of the two data types such as the k-mean, k-mode [16], fuzzy k-mode [17], COOLCAT [5] and G-ANMI algorithm [10]. Cao et al. [8] proposed a new dissimilarity measure for the k-mode algorithm, and Bai et al. [3] presented a new method to simultaneously find initial center and the number of clusters for the categorical data.

Up till now, there is some work for dealing with mixed data. Huang [20] proposed a k-prototype algorithm which integrates the k-means and k-mode to cluster mixed data. Due to the uncertainty of the data, the fuzzy k-prototype algorithm [6], Ahmad and Dey's algorithm [1] and KL-FCM-GM algorithm [9] were proposed to extend the k-prototype algorithm. The KL-FCM-GM algorithm is an extension of the Gath-Geva algorithm [13] which is based on the assumption of data deriving from clusters of Gaussian form, and it is designed for the Gauss-Multinomial distributed data. Zheng et al. [30] presented an evolutionary k-prototype algorithm (EKP). By introducing an evolutionary algorithm framework, the EKP algorithm has the global search ability. Li and Biswas [24] proposed the Similarity-based Agglomerative Clustering (SBAC) which is a hierarchical agglomerative algorithm. The SBAC algorithm adopts the similarity measure defined by Goodall [14] to evaluate the similarity among data objects. However, the quadratic computational cost makes it unacceptable for clustering large datasets. Hsu and Chen [18] proposed a Clustering Algorithm based on the Variance and Entropy (CAVE) for clustering mixed data. However, the CAVE algorithm needs to build the distance hierarchy for every categorical attribute and the determination of distance hierarchy requires the domain expertise. It is not easy to apply this algorithm to the data set where there are many categorical attributes and little domain expertise. Ahmad and Dey [2] also proposed a k-mean type algorithm for mixed data. In their method, the significance of attribute and the distance between categorical values can be evaluated according to the co-occurrence of categorical values, but their method is a hard partition clustering algorithm.

In this paper, we present a novel fuzzy k-prototype algorithm to cluster mixed data. In our method, we integrate mean and fuzzy centroid to represent the prototype of a cluster. Furthermore, we employ a new dissimilarity measure, which takes account of the significance of each attribute and the distance between categorical

* Corresponding author. Fax: +86 0431 85155352.
E-mail addresses: jinchao0374@gmail.com (J. Ji), pangwei@jlu.edu.cn, pang.wei@abdn.ac.uk (W. Pang), cgzhou@jlu.edu.cn (C. Zhou), hanx@jlu.edu.cn (X. Han), wzj0431@gmail.com (Z. Wang).

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2012.01.006

values, to evaluate the dissimilarity between data object and prototype. Then we apply our algorithm to cluster mixed data.

The rest of this paper is organized as follows: we first introduce some notations used throughout this paper in Section 2. In Section 3, the conventional k-prototype algorithm is described. In Section 4, we describe our method. Section 5 presents the experimental evidence that demonstrates the advantage of the proposed algorithm. Finally, Section 6 concludes the paper.

2. Notation

Let $X = \{X_1, X_2, \ldots, X_n\}$ denote a set of n data objects to be clustered and X_i $(1 \leqslant i \leqslant n)$ be a data object represented by m attributes $A_1, A_2, \ldots, A_m$. Each attribute A_j describes a domain of values denoted by $\mathrm{Dom}(A_j)$ [17]. The domains of attributes associated with mixed data are numeric and categorical, respectively. The numeric domain is represented by continuous values, and the categorical domain is usually denoted by $\mathrm{Dom}(A_j) = \{a_j^1, a_j^2, \ldots, a_j^t\}$, where t is the number of category values of categorical attribute j. A data object X_i is logically represented as a conjunction of attribute-value pairs

$$[A_1 = x_{i,1}] \wedge [A_2 = x_{i,2}] \wedge \cdots \wedge [A_j = x_{i,j}] \wedge \cdots \wedge [A_m = x_{i,m}],$$

where $x_{i,j} \in \mathrm{Dom}(A_j)$ for $1 \leqslant j \leqslant m$. Without ambiguity, we represent X_i as a vector $[x_{i,1}^r, x_{i,2}^r, \ldots, x_{i,p}^r, x_{i,p+1}^c, x_{i,p+2}^c \ldots, x_{i,m}^c]$ where the first p elements with the superscript r are numeric values and the rest are categorical ones. We assume that every data object has exactly m attributes for the data set considered in this paper.

3. The k-prototype algorithm

The objective of k-prototype is to group the dataset X into k clusters by minimizing the cost function

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} u_{il} d(x_i, Q_l). \tag{3.1}$$

Here Q_l is the representative vector or prototype for a cluster l, u_{il} is an element of the partition matrix $U_{n\times k}$, and $d(x_i, Q_l)$ is the dissimilarity measure defined as follows:

$$d(x_i, Q_l) = \sum_{j=1}^{p} (x_{ij}^r - q_{lj}^r)^2 + \mu_l \sum_{j=p+1}^{m} \delta(x_{ij}^c, q_{lj}^c), \tag{3.2}$$

where $\delta(p, q) = 0$ for $p = q$, and $\delta(p, q) = 1$ for $p \neq q$; $x_{ij}^r (x_{ij}^c)$ is the value of the jth numeric (categorical) attribute for the data object i; $q_{lj}^r (q_{lj}^c)$ is the prototype of the jth numeric (categorical) attribute in the cluster l; μ_l is a weight for categorical attributes in the cluster l. The process of the k-prototype algorithm is described as follows:

Step 1. Select k initial prototypes for k clusters from the date set X.
Step 2. Allocate each data object in X to the cluster whose prototype is the nearest to it according to Eq. (3.2). Update the prototype of the cluster after each allocation.
Step 3. After all data objects have been allocated to a cluster, retest the similarity of data objects against the current prototypes. If a data object is found that its nearest prototype belongs to another cluster rather than its current one, reallocate the data object to that cluster and update the prototypes of both clusters.
Step 4. Repeat Step 3 until no data object has changed clusters after a full cycle test of X.

4. The proposed fuzzy k-prototype algorithm

One motivation of our algorithm is on one hand to preserve the uncertainty inherence in data sets for longer time before decisions are made; and on the other hand to consider the significance of different attributes towards the process of clustering. In this section, we first describe the idea about fuzzy clustering and the fuzzy centroid, which considers the former, and the idea about the distance and the significance, which takes into account the latter. Then, we propose our algorithm, which considers both of the aforementioned issues.

4.1. Fuzzy clustering

Essentially, fuzzy clustering extends traditional hard clustering in the sense that data objects can belong to various clusters with different membership degrees at the same time, whereas data objects can belong to exactly one cluster in traditional hard clustering [17,26]. Most fuzzy clustering algorithms aim to minimize the objective function defined as follows:

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} (u_{il})^{\alpha} d(x_i, Q_l) \quad (\alpha \text{ is the fuzziness coefficient}, 1 < \alpha < \infty), \tag{4.1}$$

where u_{il} indicates the membership degree of the data object i to the cluster l, and it is subject to $\sum_{l=1}^{k} u_{il} = 1$, $0 \leqslant u_{il} \leqslant 1$.

4.2. The Fuzzy centroid

In a fuzzy centroid, each attribute has a fuzzy category value to describe the information distributed in a cluster [25]. For $\mathrm{Dom}(A_j) = \{a_j^1, a_j^2, \ldots, a_j^t\}$, the fuzzy centroid, denoted by $\tilde{V}$, is defined as:

$$\tilde{V} = [\tilde{v}_1, \ldots \tilde{v}_j, \ldots, \tilde{v}_m], \tag{4.2.1}$$

where

$$\tilde{v}_j = a_j^1/\omega_j^1 + a_j^2/\omega_j^2 + \cdots + a_j^{\kappa}/\omega_j^{\kappa} + \cdots + a_j^t/\omega_j^t, \tag{4.2.2}$$

subject to

$$0 \leqslant \omega_j^{\kappa} \leqslant 1, \quad 1 \leqslant \kappa \leqslant t, \tag{4.2.3}$$

$$\sum_{\kappa=1}^{t} \omega_j^{\kappa} = 1, \quad 1 \leqslant j \leqslant m. \tag{4.2.4}$$

Each attribute $\tilde{v}_j \in \tilde{V}$ is a fuzzy category value represented as a fuzzy set $\{a_j^{\kappa}, \omega_j^{\kappa}\}$, which is a convenient notation for a fuzzy set proposed by Zadeh [28], for $1 \leqslant \kappa \leqslant t$. This is determined by the category distribution of attribute A_j for the data objects belonging to the cluster.

4.3. Distance and significance

For the given data set, let A_i denote a categorical attribute, which has two values x and y. Suppose A_j denotes another categorical attribute; w denotes a subset of values of A_j; and $\sim w$ denotes the complementary set of w. Let $p_i(w/x)$ denote the conditional probability that a data object having value x for A_i has a value belonging to w for A_j. Similarly, $p_i(\sim w/y)$ denotes the conditional probability that a data object having value y for A_i has a value belonging to $\sim w$ for A_j .

Definition 4.3.1. The distance between the pair of values x and y of A_i with respect to the attribute A_j and a particular subset w is defined as follows:

$$\delta_w^i(x, y) = p_i(w/x) + p_i(\sim w/y). \tag{4.3.1}$$

Definition 4.3.2. The distance between attribute values x and y for A_i with respect to attribute A_j, which is denoted by $\delta^{ij}(x, y)$, is given by

$$\delta^{ij}(x,y) = p_i(\vartheta/x) + p_i(\sim\vartheta/y), \tag{4.3.2a}$$

where ϑ is the subset of A_j's values that maximizes the quantity $p_i(w/x) + p_i(\sim w/y)$. Since both $p_i(\vartheta/x)$ and $p_i(\sim\vartheta/y)$ are between 0 and 1, to restrict the value of $\delta^{ij}(x, y)$ between 0 and 1, we redefine $\delta^{ij}(x, y)$ as:

$$\delta^{ij}(x,y) = p_i(\vartheta/x) + p_i(\sim\vartheta/y) - 1.0. \tag{4.3.2b}$$

Eqs. (4.3.2a) and (4.3.2b) state that distance between values x and y of A_i as a function of their co-occurrence probabilities with a set of values of another categorical attribute A_j. In the presence of other categorical attributes, similar distance measures for the pair x and y can be computed with respect to each of these attributes. The absolute distance between the pair of values x and y is thereby computed as the average of all these values. The distance between x and y with respect to a numeric attribute is calculated by discretizing the numeric attribute.

Definition 4.3.3. For a data set with m attributes, inclusive of categorical and numeric attributes which have been discretized, the distance between two distinct values x and y of any categorical attribute A_i is given by

$$\delta(x,y) = \frac{\sum_{j=1,i\neq j}^{m}\delta^{ij}(x,y)}{m-1}. \tag{4.3.3}$$

In the above $\delta(x, y)$ has the following properties:

(1) $0 \leqslant \delta(x, y) \leqslant 1$,
(2) $\delta(x, y) = \delta(y, x)$,
(3) $\delta(x, x) = 0$.

Significance of an attribute defines the importance of that attribute in the data set [4,27]. However, determining the significance of an attribute is task-dependent. It is found that the attributes, which display a good separation of co-occurrence of values into different groups, play a more significant role in clustering of data objects [2]. In other words, an attribute plays a significant role in clustering, provided any pair of its attribute values are well separated against all other attributes, i.e., have an overall high value of $\delta(x, y)$, for all pairs of x and y. Since itself takes account of the grouping of different categorical values, the distance between two categorical values implies the significance of the corresponding attribute within it.

For numeric attributes which are normalized to be considered on the same scale, the distance between two numeric values x and y for an attribute A_i is computed by $d(x, y) = (w_i(x - y))^2$, where w_i denotes the significance of the numeric attribute. To compute the significance of a numeric attribute, we first discretize the numeric attribute. It may be noted that the discretized form is only used for computing the categorical distances and the significance of attributes. Since the values are normalized, we choose the same number of intervals T for all numeric attributes. Each interval is then assigned a categorical value $u[1], u[2], \ldots, u[T]$. Thereafter, for the discretized numeric attribute, we compute the distance of every pair of categorical values $\delta(u[r], u[s])$ in the same way as it is used for categorical values. Then, the significance of a numeric attribute is computed as the average of all pairs $\delta(u[r], u[s])$.

Definition 4.3.4. The significance w_i of a numeric attribute A_i is computed as

$$w_i = \frac{2\sum_{k=1}^{T}\sum_{j>k}^{T}\delta(u[k],u[j])}{T(T-1)}, \tag{4.3.4}$$

where T is the number of intervals taken.

4.4. The proposed algorithm

In this section we introduce the ideas presented in Sections 4.1–4.3 to the original k-prototype algorithm for clustering mixed data. As mentioned in Section 2, we represent a mixed data object X_i as a vector $[x^r_{i,1}, x^r_{i,2}, \ldots, x^r_{i,p}, x^c_{i,p+1}, x^c_{i,p+2} \ldots, x^c_{i,m}]$ where the first p elements with the superscript r are numeric values and the rest are categorical ones. The prototype therefore has two parts: the first p centroids use the mean for numeric attribute; the rest centroids use the fuzzy centroid for categorical one. Similar to the objective function E in the k-prototype algorithm, we propose a new objective function for clustering mixed data objects, which is defined as:

$$E(U,Q) = \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ij}^{\alpha} d(x_i, Q_j). \tag{4.4.1}$$

Here $U = (u_{ij})_{n\times k}$ is the fuzzy partition matrix, which satisfies $0 \leqslant u_{ij} \leqslant 1$ and $\sum_{j=1}^{k} u_{ij} = 1$. $Q_j = [q_{j1}, q_{j2}, \ldots, q_{jp}, \tilde{v}_{jp+1}, \ldots, \tilde{v}_{jm}]$ is the representative vector or prototype for cluster j; α $(1 < \alpha < \infty)$ is the fuzziness coefficient; and $d(x_i, Q_j)$ is the dissimilarity measure between the data object x_i and the prototype Q_j, which is defined as

$$d(x_i, Q_j) = \sum_{l=1}^{p}(w_l(x^r_{il} - q^r_{jl}))^2 + \sum_{l=p+1}^{m}\varphi(x^c_{il}, \tilde{v}^c_{jl})^2. \tag{4.4.2}$$

So far, we can rewrite the objective function as:

$$E(U,Q) = \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ij}^{\alpha}\left(\sum_{l=1}^{p}(w_l(x^r_{il} - q^r_{jl}))^2 + \sum_{l=p+1}^{m}\varphi(x^c_{il}, \tilde{v}^c_{jl})^2\right). \tag{4.4.3}$$

Here, w_l is calculated by Eq. (4.3.4) and

$$u_{ij} = \left(\sum_{z=1}^{k}\left(\frac{d(x_i,Q_j)}{d(x_i,Q_z)}\right)^{\frac{1}{\alpha-1}}\right)^{-1}, \tag{4.4.4}$$

q^r_{jl} is the centroid for numeric attribute which is defined by

$$q^r_{jl} = \frac{\sum_{i=1}^{n} u_{ij}^{\alpha}\cdot x^r_{il}}{\sum_{i=1}^{n} u_{ij}^{\alpha}}, \tag{4.4.5}$$

$\tilde{v}^c_{jl}$ is the fuzzy centroid for categorical attribute which is defined by

$$\tilde{v}^c_{jl} = a^1_{jl}/\omega^1_{jl} + a^2_{jl}/\omega^2_{jl} + \cdots + a^\kappa_{jl}/\omega^\kappa_{jl} + \cdots + a^t_{jl}/\omega^t_{jl}, \tag{4.4.6}$$

and

$$\varphi(x^c_{il}, \tilde{v}^c_{jl}) = \sum_{\kappa=1}^{t}\tau(x^c_{il}, a^\kappa_{jl}) \times \delta(x^c_{il}, a^\kappa_{jl}), \tag{4.4.7}$$

where $\delta(x^c_{il}, a^\kappa_{jl})$ is calculated by Eq. (4.3.3) and

$$\tau(x^c_{il}, a^\kappa_{jl}) = \begin{cases} 0, & x^c_{il} = a^\kappa_{jl}, \\ \omega^\kappa_{jl}, & x^c_{il} \neq a^\kappa_{jl}, \end{cases} \tag{4.4.8}$$

In Eq. (4.4.6),

$$\omega^\kappa_{jl} = \sum_{i=1}^{n}\gamma(x^c_{il}), \tag{4.4.9}$$

where

$$\gamma(x^c_{il}) = \begin{cases} \frac{u_{ij}^{\alpha}}{\sum_{i=1}^{n}u_{ij}^{\alpha}}, & a^\kappa_{jl} = x^c_{il}, \\ 0, & a^\kappa_{jl} \neq x^c_{il}. \end{cases} \tag{4.4.10}$$

Having presented the detailed calculation methods for all relevant variables, the process of the proposed fuzzy k-prototype algorithm can be described as follows:

Step 1. Given maximum iterative times max *Ite*, the number of clusters k, the value of α, the number of intervals T, and the threshold value ε, randomly choose k distinct data objects from the dataset and convert them to initial prototype $Q^{(t)} = (Q_1, Q_2, \ldots, Q_k)$, and set $t = 0$.

Step 2. Use Eq. (4.4.4) to calculate fuzzy partition matrix $U^{(t)}$.

Step 3. Update the prototype $Q^{(t)}$ to obtain new prototype $Q^{(t+1)}$. For numeric attribute part of prototype we use Eq. (4.4.5) and for categorical attribute part of prototype, we use Eq. (4.4.6), Eq. (4.4.9) and Eq. (4.4.10).

Step 4. If the difference between two adjacent computed values E is no more than the given threshold ε or max *Ite* equals to 0, then stop; otherwise, set $t \leftarrow t + 1$, max *Ite* $\leftarrow$ max *Ite* - 1, then go to Step 2.

The rule of conversion in Step 1 is described as follows: if the jth attribute is numeric, then each $q_{lj}^{r} \in Q_l$ is the value of the numeric value; if the jth attribute is categorical, then each $\tilde{v}_{lj}^{c} \in Q_l$ is assigned value 1 for ω_j^{κ} if $x_{lj}^{c} = a_{lj}^{\kappa}$; 0 for ω_j^{κ} if $x_{lj}^{c} \neq a_{lj}^{\kappa}$, where $1 \leqslant \kappa \leqslant t$.

4.5. Complexity analysis

In this section we consider the time and space complexities of the proposed algorithm. The time complexity consists of the updates of prototypes and partition matrix in each iteration, and the computation of the distance between two attribute values. The computational cost of updating the fuzzy prototypes, partition matrix, and distance are $O(kmn)$, $O(k(p + Nm - Np)n)$, and $O(m^2n + m^2S^3)$, respectively. Here k is the number of clusters; p is the number of numeric attributes; m is the number of all attributes; $N = \max(t)$ is the maximal number of categories for $p < t \leqslant m$; and n is the number of data objects; S is the average number of distinct categorical values. Therefore, the overall time complexity is $O(m^2n + m^2S^3 + k(m + p + Nm - Np)ns)$, where s is the number of iterations required for our algorithm to converge. The time complexity of k-prototype is $O((s + 1)kn)$. So the time complexity of our algorithm is a little higher than that of k-prototype. When the $n \gg k, m, s$, they are faster than the hierarchical clustering algorithm such as the SBAC algorithm whose time complexity is generally $O(n^2)$. For space complexity, it requires $O(k(p + mN - pN) + mn)$ to store prototypes and the data set X, and $O(kn)$ to store the partition matrix U. Thus, the overall space complexity of our algorithm is $O(k(p + n + mN - pN) + mn)$.

5. Experimental results

To evaluate the effectiveness of our algorithm, we use the proposed algorithm to cluster four mixed data sets: Zoo, Acute Inflammations, Heart Disease and Credit Approval, which are taken from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets.html). For all data sets, the number of intervals T is set 10. The initial prototypes of k-prototype algorithm are k distinct data objects (without missing value) randomly selected from the data set. However, for our method, we convert these k distinct data objects to initial prototypes. In clustering analysis, the clustering accuracy (r) [17] is one of the most commonly used criteria to evaluate the quality of clustering results and thus validate the performance of the clustering algorithms under study. In this paper, we also exploit this kind of accuracy measurement to access the obtained clustering results. The clustering accuracy (r) is defined as

$$r = \frac{\sum_{i=1}^{k} a_i}{n}, \tag{5.1}$$

where a_i is the number of data objects occurring both in ith cluster and its corresponding true class, and n is the number of data objects in the data set. According to this measure, a higher value of r indicates a better clustering result, with perfect clustering yielding a value of $r = 1.0$. Due to the random initial prototypes, our algorithm and the k-prototype algorithm are run 100 trials, and the average accuracy is calculated. In addition, we set max *Ite* = 100 and $\varepsilon = 1.1102 \times 10^{-23}$ in all experiments.

Zoo data set consists of 101 data objects, each of which is described by one numeric attributes and 16 categorical attributes. The last categorical attribute is the class attribute which has 7 values. Fig. 1 gives the impact of varying fuzzy coefficient α on the clustering accuracy r of our algorithm for clustering this data set. From this figure, we can see that the r achieves its best value when α equals to 2.1. In Table 1, we list the clustering accuracy r of our proposed algorithm, KL-FCM-GM, EKP, SBAC, and k-prototype algorithm. The KL-FCM-GM, EKP, SBAC, and k-prototype algorithm give clustering accuracies 0.864, 0.629, 0.426, 0.806, respectively. In contrast, the proposed algorithm gives higher clustering accuracy $r = 0.908$ at $\alpha = 2.1$. Thus, in this case, the proposed algorithm is

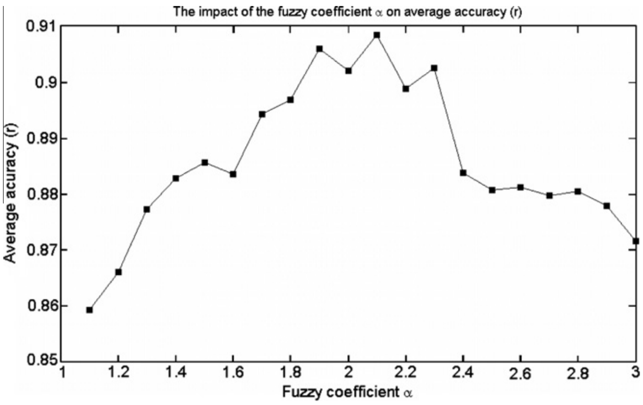

Fig. 1. The impact of the fuzzy coefficient α on the average accuracy (r) of our proposed algorithm for clustering zoo data.

Table 1
Clustering accuracy (r) for clustering zoo data.

Algorithm	Clustering accuracy (r)
K-prototype	0.806
SABC	0.426
EKP	0.629
KL-FCM-GM	0.864 ($\alpha = 1.3$)
Our proposed algorithm	0.908 ($\alpha = 2.1$)

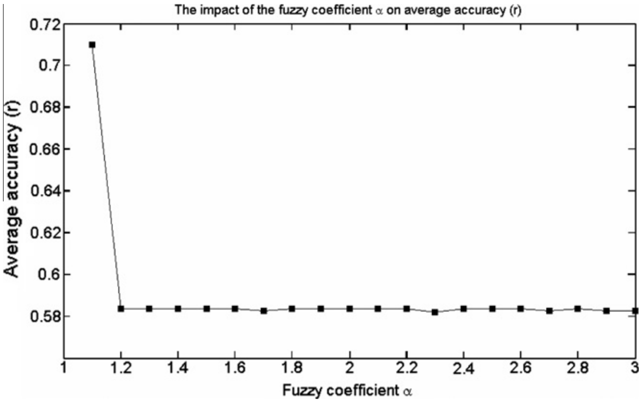

Fig. 2. The impact of the fuzzy coefficient α on the average accuracy (r) of our proposed algorithm for clustering acute inflammations data.

Table 2
Clustering accuracy (r) for clustering acute inflammations.

Algorithm	Clustering accuracy (r)
K-prototype	0.610
SABC	0.508
EKP	0.508
KL-FCM-GM	0.682 ($\alpha = 1.1$)
Our proposed algorithm	0.710 ($\alpha = 1.1$)

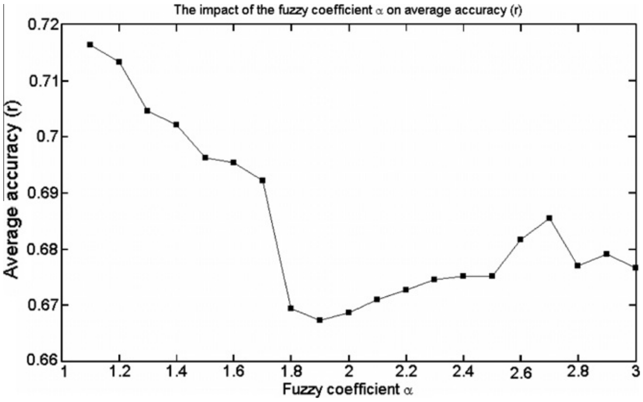

Fig. 3. The impact of the fuzzy coefficient α on the average accuracy (r) of our proposed algorithm for clustering heart disease data (5 classes and 14 attributes).

Table 3
Clustering accuracy (r) for clustering heart disease data (5 classes and 14 attributes).

Algorithm	Clustering accuracy (r)
K-prototype	0.546
SABC	0.545
EKP	0.545
KL-FCM-GM	0.653 ($\alpha = 1.3$)
Our proposed algorithm	0.717 ($\alpha = 1.1$)

4.4%, 27.9%, 48.2%, 10.2% more accurate than the KL-FCM-GM, EKP, SBAC, and k-prototype, respectively.

Acute Inflammations data set consists of 120 data objects, each of which is described by one numeric attributes and 6 categorical attributes. The last categorical attribute is the class attributes which has 2 values. Fig. 2 gives the impact of varying fuzzy coefficient α on the clustering accuracy r of our algorithm for clustering this data set. From this figure, we can see that the r achieves its best value when α equals to 1.1. In Table 2, we list the clustering accuracy r of our proposed algorithm, KL-FCM-GM, EKP, SBAC, and k-prototype algorithm. The KL-FCM-GM, EKP, SBAC, and k-prototype algorithm give clustering accuracies 0.682, 0.508, 0.508, 0.610, respectively. In contrast, the proposed algorithm gives higher clustering accuracy $r = 0.710$ at $\alpha = 1.1$. Thus, in this case, the proposed algorithm is 2.8%, 20.2%, 20.2%, 10% more accurate than the KL-FCM-GM, EKP, SBAC, and k-prototype, respectively.

Heart disease data set consists of 303 patient instances, each of which is described by 6 numeric attributes and 9 categorical attributes. The last two attributes are class attributes. When we use the 15th attribute as its class attribute, the data set has five classes (s1, s2, s3, s4, and H), and the data objects are described by 14 attributes; when we use the 14th attribute as its class attribute, the dataset has two classes (buff, sick), and the data objects are described by 13 attributes. For the first case, in Fig. 3 we give the impact of varying fuzzy coefficient α on the clustering accuracy r of our algorithm. From this figure we can see that the r achieves its best value when α equals to 1.1. In Table 3 we list the clustering accuracy r of our proposed algorithm, KL-FCM-GM, EKP, SBAC, and k-prototype algorithm. The KL-FCM-GM, EKP, SBAC, and k-prototype give clustering accuracies 0.653, 0.545, 0.545, 0.546, respectively. In contrast, the proposed algorithm gives higher clustering accuracy $r = 0.717$ at $\alpha = 1.1$. Thus, in this case, the proposed algorithm is 6.4%, 17.2%, 17.2%, 17.1% more accurate than the KL-FCM-GM, EKP, SBAC, and k-prototype algorithm respectively.

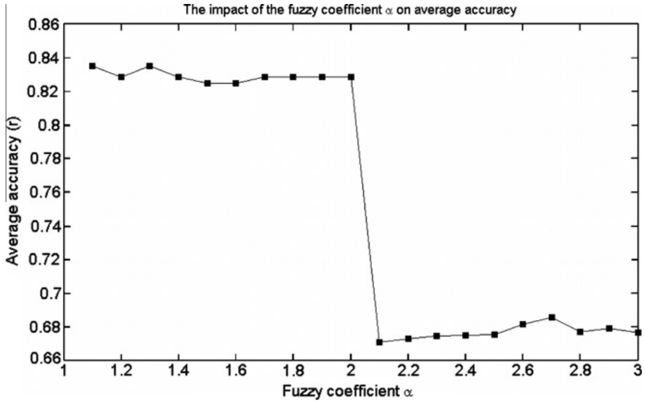

Fig. 4. The impact of the fuzzy coefficient α on average accuracy (r) of our proposed algorithm for clustering heart disease data (2 classes and 13 attributes).

Table 4
Clustering accuracy (r) for clustering heart disease data (2 classes and 13 attributes).

Algorithm	Clustering accuracy (r)
K-prototype	0.577
SABC	0.752
ECOWEB	0.739
AD's algorithm	0.746
EKP	0.545
KL-FCM-GM	0.758 ($\alpha = 1.7$)
Our proposed algorithm	0.835 ($\alpha = 1.3$)

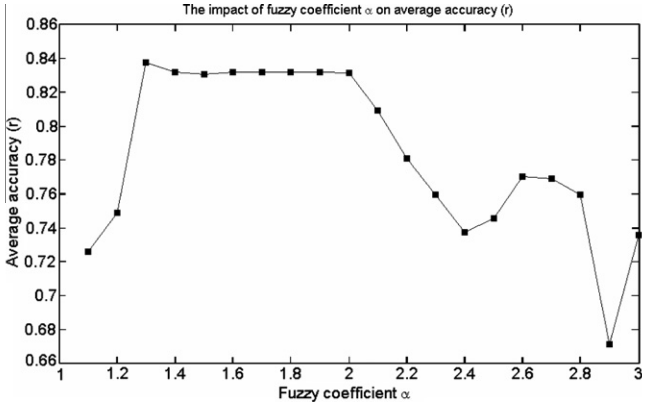

Fig. 5. The impact of fuzzy coefficient α on the average accuracy (r) of our proposed algorithm for clustering credit approval data.

For the second case where the data object in heart disease is described by 13 attributes and we use the 14th attribute of data as its class attribute. Ahmad and Dey have proposed a fuzzy k-mean type clustering algorithm (AD's algorithm) for mixed data. They use Heart Disease (the second case) to show the effectiveness of their algorithm. The comparison with other algorithms such as SBAC [24] and ECOWEB [31] are also shown in their paper [1]. We take those published results for comparison. In Fig. 4 we illustrate the impact of fuzzy coefficient α on the clustering accuracy r of our

Table 5
Clustering accuracy (r) for clustering credit approval data.

Algorithm	Clustering accuracy (r)
K-prototype	0.562
SABC	0.555
EKP	0.682
KL-FCM-GM	0.584 ($\alpha = 2.3$)
Our proposed algorithm	0.838 ($\alpha = 1.3$)

method. From this figure we can see that r achieves its best when $\alpha = 1.3$. In Table 4 we summarize the clustering accuracies of the seven algorithms. Clustering results in Table 4 show that our algorithm is 7.7%, 29%, 8.9%, 9.6%, 8.3%, and 25.8% more accurate than KL-FCM-GM, EKP, AD's algorithm, ECOWEB, SABC, and k-prototype algorithm, respectively.

Credit Approval data set contains data objects from credit card organizations, where customers are divided into two classes. It is also a mixed data set with 10 categorical and six numeric attributes (the last categorical one is class attribute). It contains 690 instances belonging to two classes: negative (383) and positive (307). Fig. 5 displays the impact of fuzzy coefficient α on the clustering accuracy of our method. In Fig. 5, the clustering accuracy r achieves its best when α equals 1.3. Table 5 summarizes the clustering accuracies of five algorithms. From this table we can see that our proposed algorithm is 25.4%, 15.6%, 28.3%, and 27.6% more accurate than KL-FCM-GM, EKP, SABC, and k-prototype, respectively.

From Figs 1–5 we can see that, given the proposed dissimilarity measure, the representation for the center of a cluster, and the number of intervals, the clustering accuracy of our algorithm is only related to the fuzzy coefficient, and is not influenced by the number of numeric (categorical) attributes in different data sets. The fuzzy coefficient reflects the overlap degree of clusters in a data set. The higher the fuzzy coefficient is, the higher overlap degree of clusters in a data set is. In all experiments we utilize the performance-based criterion, which is commonly used in FCM type algorithms, to determine the optimal value for the fuzzy coefficient α. Specifically, we perform our algorithm with the fuzzy coefficient α from 1.1 to 3.0 at an increment 0.1 for all data set. Then we choose the optimal value for α according to the clustering accuracy r of the obtained clustering results. The results in Figs 1–5 demonstrate that each data set has its own optimal value of fuzzy coefficient.

Moreover, the results in Tables 1–5 show that our method achieves better results in comparison with other algorithms. The reason is described as follow: The KL-FCM-GM algorithm is an extension of the Gath-Geva algorithm which is based on the assumption of data deriving from clusters of Gaussian form. This algorithm is suitable to deal with the Gauss-Multinomial distributed data. The EKP algorithm has the global search ability due to the introduction of an evolutionary algorithm framework, but its representation of the center of a cluster is the same as that of k-prototype, which may results in information loss. Furthermore, KL-FCM-GM considers the weight of cluster and EKP takes into account the weight of the different attribute categories, i.e., the whole categorical or numeric attributes. However, in real data set different attributes may have different significances for clustering process. In our method, we give a new representation for the center of a cluster, and a new dissimilarity measure. By introducing these new features, our method can not only fully exploit the power of fuzzy sets in classifying vague data objects in the boundaries of regions, but also take into account the significance of each attribute towards the clustering process and the distance between categorical values. The above-mentioned features make our algorithm obtain better results.

6. Conclusion

Mixed data are ubiquitous in real world databases. In this paper, we proposed a fuzzy c-mean type clustering algorithm to cluster these type of data. In our method, we integrate the fuzzy centroid and mean to represent the prototype of a cluster, and use a new measure to evaluate the dissimilarity between data objects and the prototype of a cluster. In comparison with other algorithm, our algorithm has two main contributions:

Firstly, by using the fuzzy centroid our algorithm can preserve the uncertainty inherence in data sets for longer time before decisions are made, and is therefore less prone to falling into local optima in comparison with other clustering algorithms.

Secondly, our algorithm takes account of the significance of different attributes towards clustering by using the new measure to evaluate the dissimilarity between the data objects and the cluster's prototype.

Because of these advantages our algorithm can achieves higher clustering accuracy, which has been demonstrated by experimental results. However, like many other fuzzy k-means type algorithms, our algorithm also needs to determine the value of fuzzy coefficient α in advance. In the experiments we use the performance-based criterion to determine the optimal value for the fuzzy coefficient α. How to determine the appropriate value for fuzzy coefficient α in advance still remains an open question. In the near future we will focus on investigating this problem.

Acknowledgements

The authors are very grateful to the editors and reviewers for their valuable comments and suggestions. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61175023, National High-Tech Program of China under Grant Nos. 2007AA04Z114, 2009AA02Z307, project of science and technology innovation platform of computing and software science (985 engineering), the Fundamental Research Funds for the Central Universities (Grant No. 201103195), and the Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, China.

References

[1] A. Ahmad, L. Dey, Algorithm for fuzzy clustering of mixed data with numeric and categorical attributes, ICDCIT, LNCS 3816 (2005) 561–572.
[2] A. Ahmad, L. Dey, A k-mean clustering algorithm for mixed numeric and categorical data, Data & Knowledge Engineering 63 (2) (2007) 503–527.
[3] L. Bai, J.Y. Liang, C.Y. Dang, An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data, Knowledge-Based Systems 24 (6) (2011) 785–795.
[4] J. Basak, R.K. De, S.K. Pal, Unsupervised feature selection using a neuro-fuzzy approach, Pattern Recognition Letters 19 (11) (1998) 997–1006.
[5] D. Barbara, J. Couto, Y. Li, COOLCAT: An entropy-based algorithm for categorical clustering, in: Proceedings of the Eleventh International Conference on Information and, Knowledge Management, 2002, pp. 582–589.
[6] J.C. Bezdek, J. Keller, R. Krisnapuram, Fuzzy Models and Algorithms for Pattern Recognition and Image Processing, Kluwer Academy Publishers, Boston, 1999.
[7] G. Bordogna, G. Pasi, A quality driven Hierarchical Data Divisive Soft Clustering for information retrieval, Knowledge-Based Systems 26 (1) (2012) 9–19.
[8] F.Y. Cao, J.Y. Liang, D.Y. Li, L. Bai, C.Y. Dang, A dissimilarity measure for the k-Modes clustering algorithm, Knowledge-based System 26 (1) (2012) 120–127.
[9] 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 38 (7) (2011) 8684–8689.
[10] S.C. Deng, Z.Y. He, X.F. Xu, G-ANMI: A mutual information based genetic clustering algorithm for Categorical data, Knowledge-based Systems 23 (2) (2010) 144–149.
[11] M.H. Dunham, Data Mining-introductory and Advanced Topics, Prentice-Hall, New Jersey, 2003.
[12] U.M. Fayyad, G. Piatesky-Shapiro, P. Smyth, R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI press, California, 1996.
[13] I. Gath, A.B. Geva, Unsupervised optimal fuzzy clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (7) (1989) 773–781.

[14] D.W. Goodall, A new similarity index based on probability, Biometrics 22 (4) (1966) 882–907.

[15] J. Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001.

[16] Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, Data Mining and Knowledge Discovery 2 (3) (1998) 283–304.

[17] Z. Huang, M.K. Ng, A fuzzy k-modes algorithm for clustering categorical data, IEEE Trans. Fuzzy System 7 (4) (1999) 446–452.

[18] C.C. Hsu, Y.C. Chen, Mining of mixed data with application to catalog marketing, Expert Systems with Applications 32 (1) (2007) 12–27.

[19] C.C. Hsu, Y.P. Huang, Incremental clustering of mixed data based on distance hierarchy, Expert Systems with Applications 35 (3) (2008) 1177–1185.

[20] Z. Huang, Clustering large data sets with mixed numeric and categorical values, in: Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, World Scientific, Singapore, 1997, pp. 21–34.

[21] M.Z. Islam, L. Brankovic, Privacy preserving data mining: a noise addition framework using a novel clustering technique, Knowledge-Based Systems 24 (8) (2011) 1214–1223.

[22] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, New Jersey, 1988.

[23] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a survey, ACM Computing Survey 31 (3) (1999) 264–323.

[24] C. Li, G. Biswas, Unsupervised learning with mixed numeric and nominal data, IEEE Transactions on Knowledge and Data Engineering 14 (4) (2002) 673–690.

[25] D.W. Kim, K.H. Lee, D. Lee, Fuzzy clustering of categorical data using fuzzy centroids, Pattern Recognition Letters 25 (11) (2004) 1263–1271.

[26] S. Nascimento, B. Mirkin, F. Moura-Pries, A fuzzy clustering model of data and fuzzy c-means, The Ninth IEEE International Conference on Fuzzy Systems 2 (2) (2000) 302–307.

[27] D.S. Yeung, X.Z. Wang, Improving performance of similarity-based clustering by feature weight learning, IEEE Transactions on Analysis and Machine Intelligence 24 (4) (2002) 556–561.

[28] L.A. Zadeh, A fuzzy set theoretic interpretation of linguistic hedges, Journal of Cybernetics 2 (3) (1972) 4–34.

[29] W. Zhang, T. Yoshida, X.J. Tang, Q. Wang, Text clustering using frequent itemsets, Knowledge-Based Systems 23 (5) (2011) 379–388.

[30] Z. Zheng, M.G. Gong, J.J. Ma, L.C. Jiao, Unsupervised evolutionary clustering algorithm for mixed type data, in: IEEE Congress on Evolutionary Computation (CEC), 2010, pp. 1–8.

[31] Y. Reich, S.J. Fenves, The Formation and Use of Abstract Concepts in Design, Concept Formation Knowledge and Experience in Unsupervised Learning, Morgan Kaufmann, 1991.