Applied Mathematics and Computation 295 (2017) 1–15

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

A fuzzy SV-*k*-modes algorithm for clustering categorical data with set-valued attributes

Fuyuan Cao[a,*], Joshua Zhexue Huang[b], Jiye Liang[a]

[a] *Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China*
[b] *College of Computer Sciences & Software Engineering, Shenzhen University, Shenzhen 518060, China*

ARTICLE INFO

Keywords:
Categorical data
Set-valued attribute
Set-valued modes
Fuzzy *k*-modes
Fuzzy SV-*k*-modes

ABSTRACT

In this paper, we propose a fuzzy SV-*k*-modes algorithm that uses the fuzzy *k*-modes clustering process to cluster categorical data with set-valued attributes. In the proposed algorithm, we use Jaccard coefficient to measure the dissimilarity between two objects and represent the center of a cluster with set-valued modes. A heuristic update way of cluster prototype is developed for the fuzzy partition matrix. These extensions make the fuzzy SV-*k*-modes algorithm can cluster categorical data with single-valued and set-valued attributes together and the fuzzy *k*-modes algorithm is its special case. Experimental results on the synthetic data sets and the three real data sets from different applications have shown the efficiency and effectiveness of the fuzzy SV-*k*-modes algorithm.

1. Introduction

The *k*-means algorithm is one of the most popular and best-known algorithms for clustering numerical data [1,2]. However, a lot of data in real applications are described by categorical attributes. For example, gender, profession, title, and hobby of customers are usually defined as categorical attributes. Unlike numeric data, categorical values are discrete and unordered. The standard *k*-means clustering process cannot be directly applied to categorical data due to lacking of geometric properties. Huang [3] proposed a *k*-modes algorithm to cluster categorical data by modifying the standard *k*-means clustering process [4]. In the *k*-modes algorithm, Huang used the simple matching dissimilarity measure to compute the distance between two categorical objects and represented the center of a cluster with modes instead of means and gave a frequency-based method to update modes. In [5], Huang further presented a fuzzy *k*-modes algorithm that is the fuzzy version of the *k*-modes algorithm in the framework of the fuzzy *k*-means algorithm [6]. Because of their efficiency in clustering very large categorical data, the *k*-modes and fuzzy *k*-modes algorithms have been widely used in various applications [7–12].

For most of data mining algorithms, a table or matrix is usually used as an input. In this matrix, each row represents an object and each column is an attribute only having a value for each object [13]. However, in real applications, an object may take multiple values in some attributes. For example, many people have more than one hobby in questionnaire. Such a data representation is widespread in many domains, such as retails, insurances and telecommunications. A more general data representation is shown in Table 1.

* Corresponding author.
E-mail addresses: cfy@sxu.edu.cn (F. Cao), zx.huang@szu.edu.cn (J.Z. Huang), ljy@sxu.edu.cn (J. Liang).

http://dx.doi.org/10.1016/j.amc.2016.09.023
0096-3003/

Table 1
An example data set on questionnaire.

ID	Name	Sex	...	Title	Hobby
1	John	M		{CEO, Prof.}	{Sport, Music}
2	Tom	M		{CEO, Chair}	{Reading, Sport}
...	...	...	...	...	...
n	Katty	F		{Prof., Chair}	{Traveling, Music}

Without loss of generality, data in Table 1 can be formulated as follows. Suppose that $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ is a set of n objects and each object is described by m attributes $\{A_1, A_2, \ldots, A_m\}$, where $X_i = (X_{i1}; X_{i2}; \ldots; X_{im})$ and $1 \le i \le n$. Let V^j be the domain values of the attribute A_j in $\mathbf{X}$ and V^{A_j} be the power set of V^j, if $X_{ij} \in V^{A_j}$, we call X_i as a set-valued object and A_j as a set-valued attribute.

To cluster $\mathbf{X}$, the most intuitive method is to convert $V^j (1 \le j \le m)$ into $|V^j|$ binary categorical attributes. The value 0 or 1 indicates the categorical value is absent or present [14]. Although transformation simplifies the representation of set-valued objects, this treatment unavoidably results in semantic information loss, especially in the understandability of clustering results. Moreover, as the number of categorical attributes increases, two set-valued objects are very likely to be similar even if the categorical values they contain are very different [15].

Different distance functions between two objects often result in different cluster structures in clustering algorithms. The attribute values of different set-valued objects usually overlap for a given attribute instead of equal or unequal. For example, the objects 1 and 2 in Table 1 have one overlapping value "CEO" for the attribute *Title*. It is only natural that the dissimilarity measure between two set-valued objects should be in the range of [0,1] instead of {0, 1} for a given attribute. Thus, inherent clusters probably overlap in a data set. The fuzzy *k*-modes algorithm has obtained better results in clustering data with overlapping clusters [9]. Moreover, the fuzzy partition matrix can provide more information to help users to determine the final clustering and to identify the boundary objects.

In this paper, we propose a fuzzy method to cluster objects with set-valued attributes. The main contributions of the paper are outlined as follows:

- We define the center of a cluster as set-valued-modes which is a set-valued object that minimizes the sum of the distance between each object in the cluster and the set-valued modes.
- We develop a way to obtain the fuzzy partition matrix and give a heuristic update way of cluster centers to minimize the objective function.
- We propose a fuzzy SV-*k*-modes algorithm which can partition data with single-valued and set-valued attributes together and the fuzzy *k*-modes algorithm is its special case.
- We analyze the influence of the fuzziness factor for the effectiveness of the fuzzy SV-*k*-modes algorithm.
- Experimental results on the synthetic and real data sets have shown the efficiency and effectiveness of the fuzzy SV-*k*-modes algorithm.

The rest of this paper is structured as follows. Section 2 reviews the hard and fuzzy *k*-modes algorithms. In Section 3, a fuzzy SV-*k*-modes algorithm is presented. In Section 4, we propose an algorithm to generate set-valued data and validate the scalability of the fuzzy SV-*k*-modes algorithm. In Section 5, we show experimental results on the three real data sets from different applications. We draw conclusions in Section 6.

2. The hard and fuzzy *k*-modes algorithms

In this section, we briefly review the *k*-modes [3] and fuzzy *k*-modes [5] algorithms, which have become a very popular technique in clustering categorical data. Both these two algorithms use the simple matching dissimilarity measure for categorical objects, modes instead of means for clusters. They use different methods to update modes in the clustering process for minimizing the objective function. In the *k*-modes algorithm, a mode is composed of the value that occurs most frequently in each attribute for a given cluster. In the fuzzy *k*-modes algorithm, each attribute value of a mode is given by the value that achieves the maximum of the summation of membership degrees in a given cluster. These modifications have removed the numeric-only limitation of the *k*-means and fuzzy *k*-means algorithms [16].

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of n objects described by a set of m categorical attributes $\{A_1, A_2, \ldots, A_m\}$, where $x_i = (x_{i1}; x_{i2}; \ldots; x_{im})$ and $1 \le i \le n$. The simple matching dissimilarity measure between x_i and x_j is defined as

$$d(x_i, x_j) = \sum_{s=1}^{m} \delta(x_{is}, x_{js}), \tag{1}$$

where

$$\delta(x_{is}, x_{js}) = \begin{cases} 0, & if\ x_{is} = x_{js}. \\ 1, & otherwise. \end{cases} \tag{2}$$

A center Q of X, i.e. modes is defined if Q minimizes

$$D(X, Q) = \sum_{i=1}^{n} d(x_i, Q). \tag{3}$$

Here, Q is not necessarily an object of X.

The clustering aim of the k-modes and fuzzy k-modes algorithms is to partition X into k clusters and find $\mathbf{W}$ and $\mathbf{Q}$ that minimize the objective function,

$$\mathbf{F}(\mathbf{W}, \mathbf{Q}) = \sum_{l=1}^{k} \sum_{i=1}^{n} \omega_{li}^{\alpha} d(x_i, Q_l), \tag{4}$$

subject to

$$0 \le \omega_{li} \le 1, 1 \le l \le k, 1 \le i \le n, \tag{5}$$

$$\sum_{l=1}^{k} \omega_{li} = 1, 1 \le i \le n, \tag{6}$$

and

$$0 < \sum_{i=1}^{n} \omega_{li} < n, 1 \le l \le k, \tag{7}$$

where k is a known number of clusters, $\alpha \in [1, \infty]$ is a fuzziness factor, $\mathbf{W} = [\omega_{li}]$ is a k-by-n real matrix and each element indicates the membership degree of object x_i belonging to the lth cluster, $\mathbf{Q} = [Q_1, Q_2, \ldots, Q_k]$ and Q_l is the center of the lth cluster with m categorical attributes. When $\alpha = 1$ and $\omega_{li} = \{0, 1\}$, Eq. (4) corresponds to the objective function of the k-modes algorithm. Huang gave the modes updating methods of these two algorithms in [3,5], respectively.

3. Fuzzy SV-*k*-modes clustering

k-type clustering algorithms, such as the k-means, fuzzy k-means, k-modes and fuzzy k-modes algorithms, consist of three components: (1) distance function, (2) representation of cluster centers, and (3) update process of cluster centers. In this section, we calculate the distance between two set-valued objects using Jaccard coefficient [17] and define the representation of a set of objects as set-valued modes, and give a heuristic update way of cluster centers.

3.1. Distance between two set-valued objects

Let X_i and X_j be two set-valued objects described by a set of m attributes $\{A_1, A_2, \ldots, A_m\}$, the dissimilarity measure between X_i and X_j is defined as

$$\mathbf{d}(X_i, X_j) = \sum_{s=1}^{m} \delta'(X_{is}, X_{js}), \tag{8}$$

where

$$\delta'(X_{is}, X_{js}) = 1 - \frac{|X_{is} \bigcap X_{js}|}{|X_{is} \bigcup X_{js}|}. \tag{9}$$

Obviously, $d(x_i, x_j)$ is a special case of $\mathbf{d}(X_i, X_j)$.

3.2. Set-valued modes

A center Q' of $\mathbf{X}$ is defined as set-valued modes if Q' minimizes

$$\mathbf{D}(\mathbf{X}, Q') = \sum_{i=1}^{n} \mathbf{d}(X_i, Q'). \tag{10}$$

Here, Q' is not necessarily an object of $\mathbf{X}$. Q is a special case of Q'.

3.3. The fuzzy-SV-k-modes algorithm with heuristic update strategy

In the section, we propose a fuzzy SV-*k*-modes algorithm by extending the fuzzy *k*-modes algorithm. The fuzzy SV-*k*-modes algorithm uses the fuzzy *k*-modes paradigm to cluster categorical data with set-valued attributes. For the fuzzy SV-*k*-modes algorithm, the objective of partitioning $\mathbf{X}$ into k clusters is also to find $\mathbf{W}'$ and $\mathbf{Q}'$ that minimize the objective function,

$$\mathbf{F}'(\mathbf{W}', \mathbf{Q}') = \sum_{l=1}^{k}\sum_{i=1}^{n} \omega_{li}^{\alpha}\mathbf{d}(X_i, \mathbf{Q}_l), \tag{11}$$

subject to

$$0 \le \omega_{li} \le 1, 1 \le l \le k, 1 \le i \le n, \tag{12}$$

$$\sum_{l=1}^{k} \omega_{li} = 1, 1 \le i \le n, \tag{13}$$

and

$$0 < \sum_{i=1}^{n} \omega_{li} < n, 1 \le l \le k, \tag{14}$$

where $\mathbf{W}' = [\omega_{li}]$ is a k-by-n real matrix and each element indicates the membership degree of object X_i belonging to the lth cluster, $\mathbf{Q}' = [\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_k]$, and $\mathbf{Q}_l$ is the set-valued modes of the lth cluster with m set-valued attributes.

Minimization of $\mathbf{F}'$ in Eq. (11) with the constraints in Eqs. (12)–(14) forms a class of constrained nonlinear optimization problems whose solutions are unknown. To optimize $\mathbf{F}'$ in Eq. (11), the usual method is to use partial optimization for $\mathbf{Q}'$ and $\mathbf{W}'$. In this method, we first fix $\mathbf{Q}'$ and find $\mathbf{W}'$ that minimizes $\mathbf{F}'$. Then, we fix $\mathbf{W}'$ and compute $\mathbf{Q}'$ that minimizes $\mathbf{F}'$. The matrix $\mathbf{W}'$ can be obtained by the following theorem.

Theorem 1. *Let* $\hat{\mathbf{Q}}$ *be fixed and minimize* $\mathbf{F}'$ *subject to Eqs. (12)–(14). For* $\alpha > 1$, $\hat{\mathbf{W}}$ *is given by*

$$\hat{\omega}_{li} = \begin{cases} 1, & \text{if} \quad X_i = \hat{\mathbf{Q}}_l, \\ 0, & \text{if} \quad X_i = \hat{\mathbf{Q}}_h, h \ne l \\ \dfrac{1}{\sum_{h=1}^{k}\left[\frac{\mathbf{d}(\hat{\mathbf{Q}}_l, X_i)}{\mathbf{d}(\hat{\mathbf{Q}}_h, X_i)}\right]^{1/(\alpha-1)}}, & \text{if} \quad X_i \ne \hat{\mathbf{Q}}_l \quad \text{and} \quad X_i \ne \hat{\mathbf{Q}}_h, 1 \le h \le k \end{cases} \tag{15}$$

To minimize $\mathbf{F}'(\mathbf{W}', \mathbf{Q}')$ if $\mathbf{W}'$ is fixed, we only need to minimize $\sum_{i=1}^{n} \omega_{li}^{\alpha}\delta'(X_{is}, \mathbf{Q}_{ls})$, the sum of the distance between the objects in $\mathbf{X}$ and $\mathbf{Q}_l$ on the attribute A_s where $s \in \{1, 2, \ldots, m\}$. As the attribute values in $\mathbf{Q}_{ls}$ must be from the values in V^s, the number of categorical values in $\mathbf{Q}_{ls}$ is in the range of $[1, |V^s|]$. If we choose u_s values $\{v_1^s, v_2^s, \ldots, v_{u_s}^s\}$ from V^s as the values of $\mathbf{Q}_{ls}$, there are $C_{|V^s|}^{u_s}$ combinations. Therefore, we need to traverse every combination to find a $\mathbf{Q}_{ls}$, which minimizes $\sum_{i=1}^{n} \omega_{li}^{\alpha}\delta'(X_{is}, \mathbf{Q}_{ls})$. To reduce the complexity of update process, we give a heuristic update strategy to obtain the center of a cluster below.

The frequency of S^j is defined as if S^j is a subset of V^j,

$$f(S^j) = \frac{1}{n}\sum_{i=1}^{n} \nu(S^j, X_{ij}), \tag{16}$$

where

$$\nu(S^j, X_{ij}) = \begin{cases} \frac{|S^j|}{|X_{ij}|}, & if \quad S^j \subseteq X_{ij}. \\ 0, & otherwise. \end{cases} \tag{17}$$

Using the following strategy, we can get $\mathbf{Q}_{lj}$ in the attribute A_j. Suppose that $V^j = \{q_1^j, q_2^j, \ldots, q_{r_j'}^j\}$ is the domain values of the attribute A_j in the lth cluster, we first compute $\sum_{i=1}^{n} f(q_h^j) \times \omega_{li}^{\alpha} (1 \le h \le r_j')$ of all categorical values in V_j, and then rank the categorical values in the descending order of $\sum_{i=1}^{n} f(q_h^j)\omega_{li}^{\alpha}$ in set $V^j = \{q_1^j, q_2^j, \ldots, q_{r_j'}^j\}$. Assume that $\mathbf{Q}_{lj}$ has r_j values. We consider three situations to construct $\mathbf{Q}_{lj}$.

- When $r_j = 1$, if $\sum_{i=1}^{n} f(q_1^j) \times \omega_{li}^{\alpha} > \sum_{i=1}^{n} f(q_t^j) \times \omega_{li}^{\alpha}, t = 2, \ldots, r_j'$, we choose the categorical value $\{q_1^j\}$ as $\mathbf{Q}_{lj}$. If there is more than one the maximum of $\sum_{i=1}^{n} f(q_t^j) \times \omega_{li}^{\alpha} (t \in \{1, 2, \ldots, r_j'\})$ to cluster $\mathbf{Q}_l$, we randomly choose one value as $\mathbf{Q}_{lj}$. This case is similar to the fuzzy *k*-mode algorithm.
- When $r_j = r_j'$, we choose all categorical values in A_j for $\mathbf{Q}_{lj}$ as the center of the cluster.

- When $1 < r_j < r'_j$, we have the following three cases:

  Case 1: If $\sum_{i=1}^{n} f(q_1^j) \times \omega_{li}^{\alpha} \geq \sum_{i=1}^{n} f(q_2^j) \times \omega_{li}^{\alpha} \geq \cdots \geq \sum_{i=1}^{n} f(q_{r_j}^j) \times \omega_{li}^{\alpha} > \sum_{i=1}^{n} f(q_{r_j+1}^j) \times \omega_{li}^{\alpha}$, we choose the first r_j categorical values for $\mathbf{Q}_{lj}$.

  Case 2: If $\sum_{i=1}^{n} f(q_1^j) \times \omega_{li}^{\alpha} \geq \sum_{i=1}^{n} f(q_2^j) \times \omega_{li}^{\alpha} \geq \cdots > \sum_{i=1}^{n} f(q_{r_j}^j) \times \omega_{li}^{\alpha} = \sum_{i=1}^{n} f(q_{r_j+1}^j) \times \omega_{li}^{\alpha} > \cdots \geq \sum_{i=1}^{n} f(q_{r'_j}^j) \times \omega_{li}^{\alpha}$, we firstly choose the first $r_j - 1$ values $Q'' = \{q_1^j, q_2^j, \ldots, q_{r_{j-1}}^j\}$ as part of values for $\mathbf{Q}_{lj}$. If $\sum_{m=1}^{r_j-1} \sum_{i=1}^{n} f(\{q_m^j, q_{r_j}^j\}) \times \omega_{li}^{\alpha} > \sum_{m=1}^{r_j-1} \sum_{i=1}^{n} f(\{q_m^j, q_{r_j+1}^j\}) \times \omega_{li}^{\alpha}$, we choose $\{q_{r_j}^j\}$ as the r_jth value for $\mathbf{Q}_{lj}$, i.e., $\mathbf{Q}_{lj} = \{q_{r_j}^j\} \cup Q''$. If $\sum_{m=1}^{r_j-1} \sum_{i=1}^{n} f(\{q_m^j, q_{r_j}^j\}) \times \omega_{li}^{\alpha} < \sum_{m=1}^{r_j-1} \sum_{i=1}^{n} f(\{q_m^j, q_{r_j+1}^j\}) \times \omega_{li}^{\alpha}$, we choose $\mathbf{Q}_{lj} = \{q_{r_j+1}^j\} \cup Q''$. If $\sum_{m=1}^{r_j-1} f(\{q_m^j, q_{r_j}^j\}) = \sum_{m=1}^{r_j-1} f(\{q_m^j, q_{r_j+1}^j\})$, we choose either $\mathbf{Q}_{lj} = \{q_{r_j}^j\} \cup Q''$ or $\mathbf{Q}_{lj} = \{q_{r_j+1}^j\} \cup Q''$.

  Case 3: If $\sum_{i=1}^{n} f(q_1^j) \times \omega_{li}^{\alpha} \geq \sum_{i=1}^{n} f(q_2^j) \times \omega_{li}^{\alpha} \geq \cdots > \sum_{i=1}^{n} f(q_{r_j-p'}^j) \times \omega_{li}^{\alpha} = \cdots = \sum_{i=1}^{n} f(q_{r_j}^j) \times \omega_{li}^{\alpha} = \sum_{i=1}^{n} f(q_{r_j+1}^j) \times \omega_{li}^{\alpha} = \cdots = \sum_{i=1}^{n} f(q_{r_j+p}^j) \times \omega_{li}^{\alpha} > \sum_{i=1}^{n} f(q_{r_j+p+1}^j) \times \omega_{li}^{\alpha} \geq \cdots \geq \sum_{i=1}^{n} f(q_{r'_j}^j) \times \omega_{li}^{\alpha}$, where p' and p are two integers, we choose the first $(r_j - p' - 1)$ categorical values as $Q'' = \{q_1^j, q_2^j, \ldots, q_{r_j-p'-1}^j\}$. Assume that $\mathbf{Q}^j$ is the set of all combinations of $p' + 1$ categorical values from the next $p' + p + 1$ categorical values. Let $\prod_s$ be a combination in $\mathbf{Q}^j$ that produces the largest sum of frequencies, i.e., $\sum_{i=1}^{n} \sum_{m=1}^{r_j-p'-1} f(\{q_m^j\} \cup \prod_s) \times \omega_{li}^{\alpha} \geq \sum_{i=1}^{n} \sum_{m=1}^{r_j-p'-1} f(\{q_m^j\} \cup \prod_t) \times \omega_{li}^{\alpha}$ where $\prod_t$ is any combination in $\mathbf{Q}^j$ and $\prod_s \neq \prod_t$. We choose $\prod_s$ as the rest values for $\mathbf{Q}_{lj}$, i.e., $\mathbf{Q}_{lj} = \prod_s \cup Q''$.

In the k-modes algorithm, we choose the most frequent categorical value as the mode in a given attribute. For $\mathbf{X}$, only one value cannot adequately represent a cluster in a given attribute. In general cases, we choose $r_j = round(\sum_{i=1}^{n} \frac{|X_{ij}|}{n})$ values as the set-valued mode in the attribute $A_j (1 \leq j \leq m)$.

Based on the above analysis, the fuzzy SV-k-modes algorithm with heuristic strategy is described as follows.

The complexity of the fuzzy SV-k-modes algorithm is analyzed as follows. We only consider two major computational steps:

- Assigning objects to clusters, that is to say, computing membership degrees of objects belonging to clusters. The computational complexity is $\mathcal{O}(|V^j|)$ with respect to A_j. Therefore, the computational complexity for this step is $\mathcal{O}(m \times |V'|)$, where $|V'| = \max\{|V^j|, 1 \leq j \leq m\}$.
- Computing set-valued-modes from the fuzzy matrix. The main goal of updating cluster centers is to find the set-valued modes in each cluster according to the partition matrix $\mathbf{W}'$. The time complexity for this step is $\mathcal{O}(km \times |V'|)$. where $|V'| = \max\{|V^j|, 1 \leq j \leq m\}$.

If the clustering process needs t iterations to converge, the total computational complexity of the proposed algorithm is $\mathcal{O}(nmtk \times |V'|)$, where $|V'| = \max\{|V^j|, 1 \leq j \leq m\}$. It is clear that the time complexity of the proposed algorithm increases linearly as the number of dimensions, objects, or clusters increases.

4. Experiments on synthetic data

In this section, we propose an algorithm to generate data with set-valued attributes and validate the efficiency of the fuzzy SV-k-modes algorithm on synthetic data sets.

4.1. Synthetic data generation method

To the best of our knowledge, we cannot find a method that can generate data with set-valued attributes. To best validate the properties of the fuzzy SV-k-modes algorithm, we need to develop a new method to generate synthetic data with set-valued attributes. Let a synthetic data set $\mathbf{X}$ be a set of n objects $\{X_1, X_2, \ldots, X_n\}$, each of which is described by a set of m set-valued attributes $\{A_1, A_2, \ldots, A_m\}$.

We assume that the set of values of each attribute is given before $\mathbf{X}$ generated. Let V^j denote a set of values of $A_j (j = 1, 2, \ldots, m)$ appearing in the objects of $\mathbf{X}$. If $\mathbf{X}$ is classified into k clusters, then the distributions of attribute values of objects in the same cluster are close to each other while the distributions of attribute values in different clusters have difference. Therefore, we can control the structure of clusters in $\mathbf{X}$ using the distributions of attribute values.

To generate a synthetic data set $\mathbf{X}$ consisting of k clusters $C = \{C_1, C_2, \ldots, C_k\}$, each of which has a particular distribution. We need to use the following parameters.

- k: the number of clusters desired;
- c_i: the number of objects in cluster C_i;
- ρ: the percentage of overlap attribute values between any two clusters.

For simplicity, we suppose that the size of domain values is the same in all attributes and the number of objects in each cluster is equal to n. We use the following steps to generate an object X in cluster C_i.

- With the parameters ρ, k and V^j, we can obtain the domain values of the jth attribute in cluster C_i;
- Select randomly a non empty subset from the domain values of the jth attribute as the jth component of X.
- Use the same way as described in step 2 to generate the rest components of X and assign a label to X.

The detailed generating algorithm is described in *Algorithm 2*, which is abbreviated to *GSDA* (**G**enerating **S**et-valued **D**ata **A**lgorithm).

Algorithm 1 The fuzzy SV-k-modes algorithm with heuristic strategy.

Input:
- **X** : a set of n set-valued objects;
- k : the number of clusters;

Output: $\{C_1, C_2, \ldots, C_k\}$, a set of k clusters;
Method:
Step 1. Randomly choose k objects as $\mathbf{Q}'^{(1)}$. Determine $\mathbf{W}'^{(1)}$ such that $\mathbf{F}'(\mathbf{W}', \mathbf{Q}'^{(1)})$ is minimized with Theorem 1. Set $t = 1$.
Step 2. Determine $\mathbf{Q}'^{(t+1)}$ such that $\mathbf{F}'(\mathbf{W}'^{(t)}, \mathbf{Q}'^{(t+1)})$ is minimized with heuristic strategy. If $\mathbf{F}'(\mathbf{W}'^{(t)}, \mathbf{Q}'^{(t+1)}) = \mathbf{F}'(\mathbf{W}'^{(t)}, \mathbf{Q}'^{(t)})$, then stop; otherwise goto step 3.
Step 3. Determine $\mathbf{W}'^{(t+1)}$ such that $\mathbf{F}'(\mathbf{W}'^{(t+1)}, \mathbf{Q}'^{(t+1)})$ is minimized. If $\mathbf{F}'(\mathbf{W}'^{(t+1)}, \mathbf{Q}'^{(t+1)}) = \mathbf{F}'(\mathbf{W}'^{(t)}, \mathbf{Q}'^{(t+1)})$, then stop; otherwise set $t = t + 1$ and goto step 2.

Algorithm 2 The *GSDA*.

Input:
- n : the number of objects in each cluster;
- m : the number of attributes;
- V^j: the attribute values in the jth attribute;
- ρ : the overlap percentage of domain values of each attribute in differentclusters;
- k : the number of clusters;

Output: A synthetic data set **X** with label;
Method:
X $= \emptyset$;
for $i = 1$ to k **do**
  for $j = 1$ to m **do**
    Allocate uniformly V^j to k clusters $V_1^j, V_2^j, \ldots, V_k^j$;
  end for
end for
for $i = 1$ to k **do**
  for $p = 1$ to n **do**
    for $q = 1$ to m **do**
      Obtain the domain values of the qth attributes V_i^q in the ith cluster;
      for $h = 1$ to k **do**
        if i! = h **then**
          Compute the number of overlapping attribute values $rationum = round(|V_h^q| \times \rho)$;
          Select randomly $rationum$ values from V_h^q and add them to the V_i^q;
        end if
      end for
      Select randomly r values from V_i^q as the qth component of X;
    end for
    Assign the label i to object X;
    Add X to **X**;
  end for
end for
return **X**;

4.2. Scalability

To test the scalability of the fuzzy SV-k-modes algorithm, we conducted a series of experiments on synthetic data sets. We ran the fuzzy SV-k-modes algorithm by selecting randomly initial cluster centers on synthetic data sets. Considering

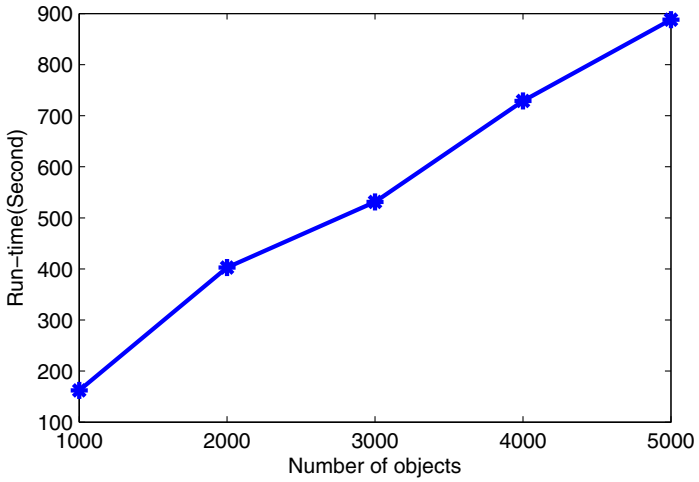

Fig. 1. Scalability of the fuzzy SV-*k*-modes algorithm with data size.

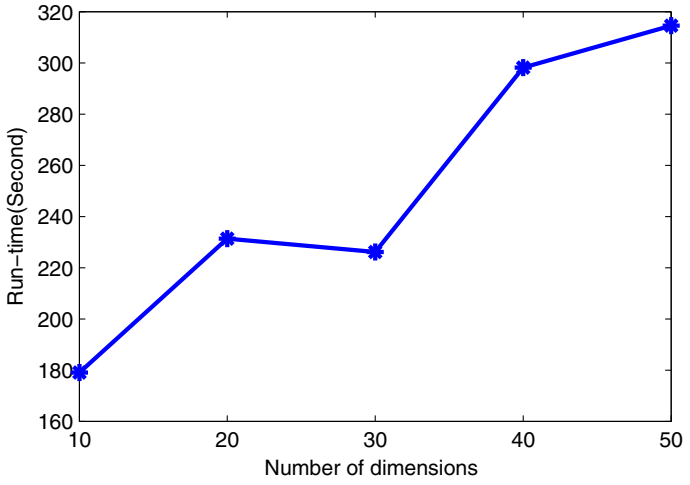

Fig. 2. Scalability of the fuzzy SV-*k*-modes algorithm with data dimensionality.

randomicity of the generating algorithm, we generated 10 synthetic data sets taken as test data sets, where ρ was set to 0.5 in *GSDA*. The average run-time in 10 data sets was taken as experimental results. All of our experiments were conducted on a PC with an Intel Xeon CPU I7 (3.4GHz) and 16GB memory. Experimental results are reported below.

Experiment 1: In this experiment, we fixed the dimensionality to 10, the number of attribute values to 10 in each attribute, the cluster number to 2, and the data size varied from 1000 to 5000 with step 1000.

Fig. 1 shows the scalability of the fuzzy SV-*k*-modes algorithm with data size. It can be seen that this algorithm is linear with respect to the data size. Therefore, the fuzzy SV-*k*-modes algorithm can ensure efficient execution when the data size is large.

Experiment 2: In this experiment, we fixed the data size to 3000, the number of attribute values to 10 in each attribute, the cluster number to 2, and the dimensionality varied from 10 to 50 with step 10.

Fig. 2 shows the scalability of the fuzzy SV-*k*-modes algorithm with dimensionality. It can be seen that the fuzzy SV-*k*-modes algorithm is linear with respect to the dimensionality. Therefore, the fuzzy SV-*k*-modes algorithm can ensure efficient execution for high dimensional data set.

Experiment 3: In this experiment, we fixed the data size to 1000, the number of attribute values to 10 in each attribute, and the dimensionality to 30. For simplicity, 2, 3, 4, 5 and 6 were taken as the number of clusters.

Fig. 3 shows the scalability of the fuzzy SV-*k*-modes algorithm with the number of clusters. It can be seen that the fuzzy SV-*k*-modes algorithm is scalable well to the number of clusters.

Experiment 4: In this experiment, we fixed the data size to 1000, the dimensionality to 10, the cluster to 2, and the number of attribute values from 10 to 50 with step 10.

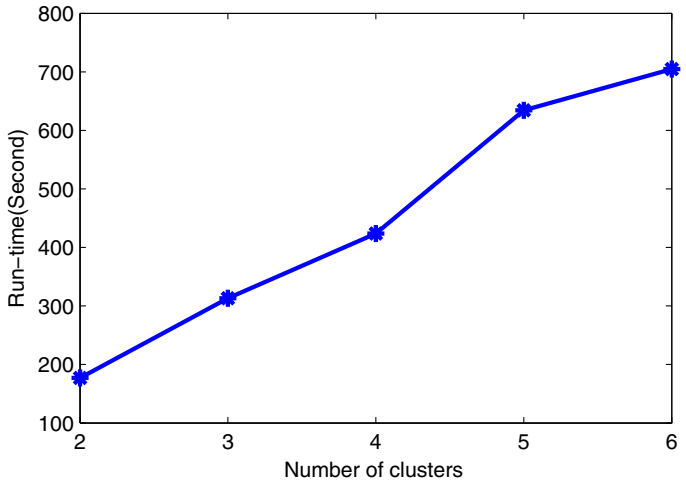

Fig. 3. Scalability of the fuzzy SV-*k*-modes algorithm with the number of clusters.

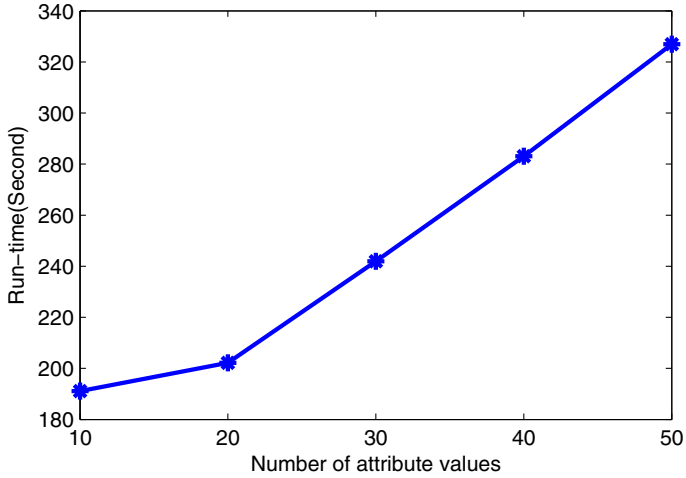

Fig. 4. Scalability of the fuzzy SV-*k*-modes algorithm with the number of attribute values.

Fig. 4 shows the scalability of the fuzzy SV-*k*-modes algorithm with the number of attribute values. We can see that the run-time of the fuzzy SV-*k*-modes algorithm nearly linearly increases with the number of attribute values increasing. This is because that the distributions of the attribute values in each attribute are nonuniform in most cases.

From the above-mentioned analysis, we find that the time complexity of the fuzzy SV-*k*-modes algorithm increases linearly as the number of objects, dimensions, clusters or attribute values increases.

5. Experiments on real data

In this section, we first gave the preprocessing processes of three real data sets and reviewed five external indexes for evaluating clustering quality. We then compared the fuzzy SV-*k*-modes algorithm with the fuzzy *k*-modes algorithm on the three real data sets. Finally, we analyzed the relationship between α and $\mathbf{W}'$ in the fuzzy SV-*k*-modes algorithm.

5.1. Data sets

Although there are many data sets with set-valued attributes in real applications, public set-valued data sets are very rare. To evaluate clustering quality of the fuzzy SV-*k*-modes algorithm, we need to conduct a series of data preprocessing for a given real data set. The main aim of data preprocessing is to decide the size of *k* and the distributions of clusters. The preprocessing processes of the three data sets are described as follows.

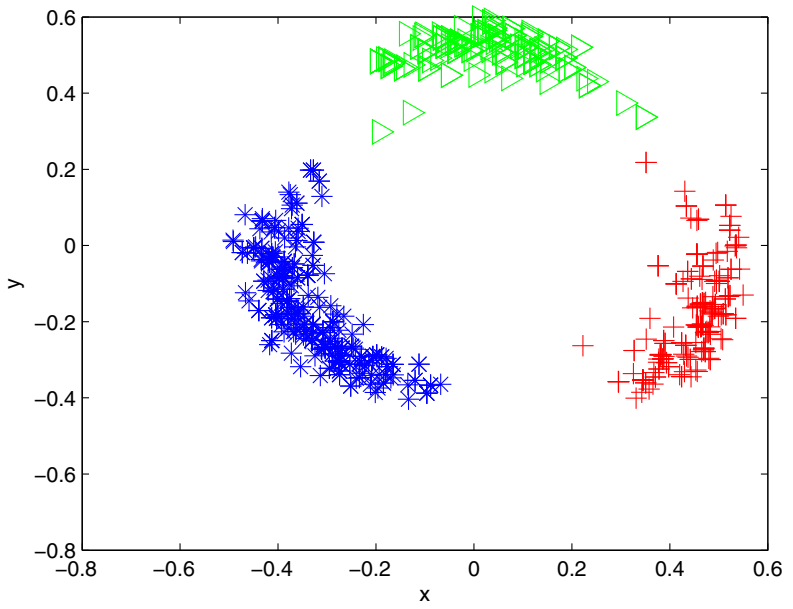

Fig. 5. The distributions of MB data.

5.1.1. Market basket data

Market basket data, which have been used earlier to evaluate association rules algorithms, are used in our study and can be downloaded from Data website[1]. This market basket data contain 1001 customers and each customer has 7 transactional records described by four attributes, which are Customer_Id, Time, Product_Name, Product_Id. As each customer has the same value in the attribute Time, we deleted this attribute. In addition, attributes Product_Name and Product_Id represent the same meaning and we only consider the attribute Product_Id. Thus, each customer has at most 7 values in the attribute Product_Id and can be transformed a set-valued object. The preprocessing of the market basket data is described as follows. We firstly visualized the market basket data using multidimensional scaling techniques [18] where the dissimilarity matrix was obtained by Eq. (8). And then, we selected objects whose coordinate values are in the range of $(x < -0.2, y < 0.2)$, $(x > 0.2, y < 0)$ and $y > 0.4$ in the coordinate system to generate a new market basket data set (abbr. MB), which has 703 objects. The distributions of MB data set are shown in Fig. 5.

From Fig. 5, we can obviously see that MB data can be divided into 3 clusters.

5.1.2. Microsoft web data

Microsoft web data set can be downloaded from UCI [19] and its associated task is Recommender-Systems. The data record the use of www.microsoft.com by 37,711 anonymous, randomly selected users. For each user, the data list all the areas of the website that the user visited in a one-week timeframe. Therefore, each user is a set-valued object described by two attributes. One attribute is User_Id, the other is the areas of the website. The preprocessing of this data is summarized as follows: firstly select users who visited the number of websites is greater than 8 to generate a temporary data; then select objects whose coordinate values of x are greater than 0.1 and less than -0.1 in the coordinate system after visualizing the temporary data to generate a new web data set (abbr. MW) which has 962 objects and includes 9857 records. The distributions of MW data are shown in Fig. 6.

From Fig. 6, obviously the number of clusters of MW data can be set to 2 in the fuzzy SV-k-modes algorithm.

5.1.3. MovieLens data

MovieLens data can be downloaded from the MovieLens website[2]. Depending on the size of the set, this data were classified into MovieLens 100k, MovieLens 1M and MovieLens 10M. MovieLens data contain rating information, user information, movie information and tag information.

We selected MovieLens 1M data to evaluate the fuzzy SV-k-modes algorithm. In MovieLens 1M data, the rating information contains 1,000,209 anonymous ratings of approximately 3900 movies made by 6040 MovieLens users who joined MovieLens in 2000. Each record of the data set represents one rating of one movie, and has the following format: User_Id::Movie_Id::Rating::Timestamp. Each user has at least 20 rating records and each rating was made on a 5-star scale.

[1] http://www.datatang.com/datares/go.aspx?dataid=613168.

[2] http://grouplens.org/datasets/movielens/.

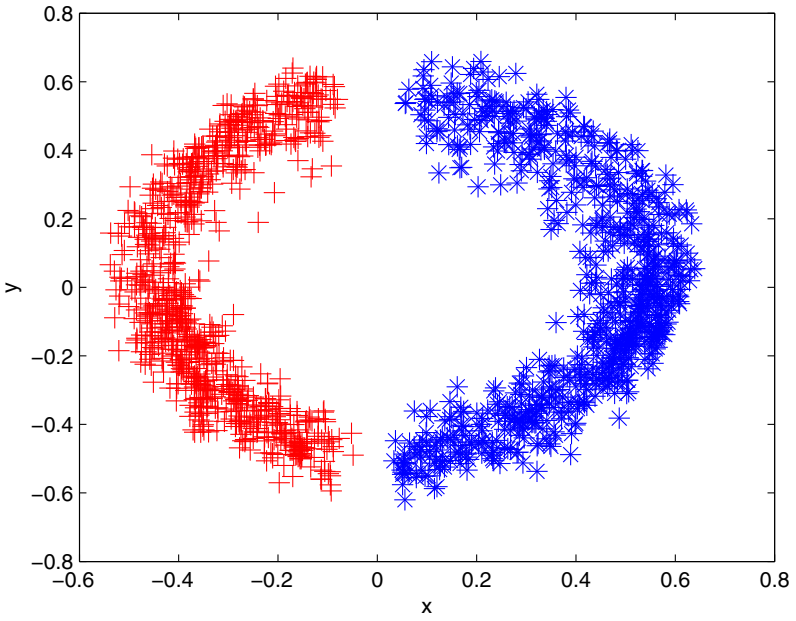

Fig. 6. The distributions of MW data.

Table 2
Summary of the three real data sets after preprocessing.

Data set	Objects	Attributes	Records	k	C_1	C_2	C_3
MB	703	2	4921	3	309	217	177
MW	962	2	9857	2	366	596	
URM	2306	6	2306	3	855	810	641

According to Movies data file structure Movie_Id::Title::Genres, we can find that each Moive_Id corresponds to more than one genre. Thus, the rating format can be transformed User_Id::Genres::Rating::Timestamp, where Genres is a set-valued attribute.

In addition, in MovieLens 1M data, each user provided some demographic information, such as Gender, Age, Occupation and Zip-code. Age was divided into seven categories according to the range of age. There are 21 attribute values for Occupation.

Joining the rating information with the user information by User_Id, we generated a new rating data having 6040 records. Each record is described by User_Id, Gender, Age, Occupation, Zip-code, Genres, Rating and Timestamp, where User_Id, Gender, Age, Occupation, Zip-code and Timestamp are six single-valued attributes while Genres and Rating are two set-valued attributes. Therefore, the new rating data can be used to evaluate the fuzzy SV-k-modes algorithm. As the domain values of Zip-code and Timestamp have too many different values, these two attributes were not been considered. We selected 2306 objects whose coordinate values are in the range of ($x < 0$, $y > 0$) and ($x > 2$, $0 < y < 1$) from the coordinate system after visiualizing the new rating data as a new user rating data set (abbr. URM). The distributions of URM data set are shown in Fig. 7.

From Fig. 7, we can divide URM data into 3 clusters and each cluster has some outliers.

The detailed information of the three real data set after preprocessing is summarized in Table 2.

5.2. Evaluation indexes

Given a categorical set-valued data set $\mathbf{X}$, let $C = \{C_1, C_2, \ldots, C_k\}$ be a clustering result of $\mathbf{X}$, $P = \{P_1, P_2, \ldots, P_k'\}$ be a real partition in $\mathbf{X}$. The overlap between C and P can be summarized in a contingency table as shown in Table 3, where n_{ij} denotes the number of objects in common between C_i and P_j, $n_{ij} = |C_i \bigcap P_j|$. c_i is the number of objects in C_i, and p_j is the number of objects in P_j.

With Table 3, Accuracy (AC), Precision (PE), Recall (RE), Adjusted rand index (ARI) and Normalized mutual information (NMI) are defined as follows:

$$AC = \frac{1}{n} \max_{j_1 j_2 \ldots j_k \in S} \sum_{i=1}^{k} n_{ij_i},$$

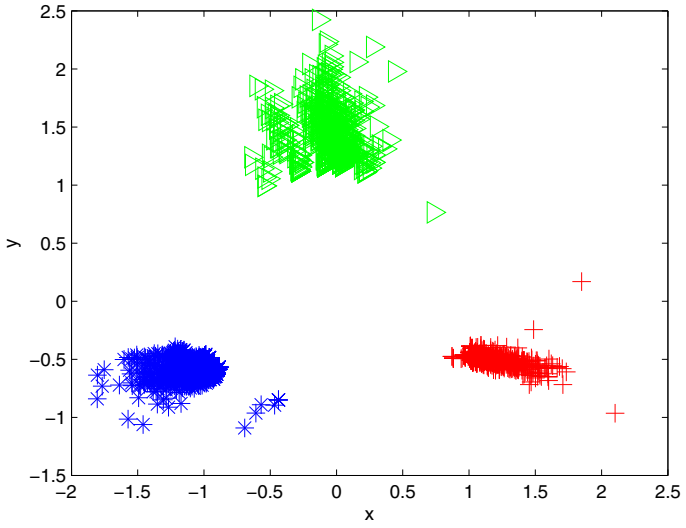

Fig. 7. The distributions of URM data.

Table 3
The contingency table.

	P_1	P_2	...	$P_{k'}$	*Sums*
C_1	n_{11}	n_{12}	...	n_{1k}	c_1
C_2	n_{21}	n_{22}	...	n_{2k}	c_2
⋮	⋮	⋮	⋱	⋮	⋮
C_k	$n_{k'1}$	$n_{k'2}$	...	$n_{k'k}$	c_k
Sums	p_1	p_2	...	$p_{k'}$	

$$PE = \frac{1}{k}\sum_{i=1}^{k}\frac{n_{ij_i^*}}{p_i}$$

$$RE = \frac{1}{k'}\sum_{i=1}^{k'}\frac{n_{ij_i^*}}{c_i},$$

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i\binom{n_{ij}}{2}\sum_j\binom{n_{ij}}{2}\right]/\sum_i\binom{c_i}{2}}{\frac{1}{2}\left[\sum_i\binom{c_i}{2}+\sum_j\binom{p_j}{2}\right] - \left[\sum_i\binom{n_{ij}}{2}\sum_j\binom{n_{ij}}{2}\right]},$$

$$NMI = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k'} n_{ij}\log(\frac{n_{ij}n}{c_i p_j})}{\sqrt{\sum_{i=1}^{k} c_i\log(\frac{c_i}{n})\sum_{j=1}^{k'} p_j\log(\frac{p_j}{n})}}.$$

where $n_{1j_1^*}+n_{2j_2^*}+\cdots+n_{kj_k^*} = \max_{j_1 j_2\cdots j_k\in S}\sum_{i=1}^{k} n_{ij_i}$ $(j_1^* j_2^*\ldots j_k^*\in S)$ and $S = \{j_1 j_2\ldots j_k : j_1, j_2, \ldots, j_k \in \{1, 2, \ldots, k\}, j_i \neq j_t \text{ for } i \neq t\}$ is a set of all permutations of $1, 2, \ldots, k$. In these experiments, we let $k = k'$, i.e., the number of clusters to be found was equal to the number of classes in the data set. In general, the higher the values of *AC*, *PE*, *RE*, *ARI* and *NMI* are, the better the clustering results are.

5.3. Clustering results

For fuzzy k-type clustering algorithms, the fuzziness factor α is an important parameter that influences the results of clustering algorithms. In the fuzzy k-means algorithm, Pal and Bezdek [20] suggested taking $\alpha \in [1.5, 2.5]$ and Yu [21] gave a theoretical upper bound for α. From the perspective of cluster validation, Zhou [22] considered that the optimal interval of α is [2.5, 3]. Wu [23] recommended that α was set to 4 when a data set contains noise and outliers. Xiong [24] found

Table 4
Comparison results of the fuzzy k-modes and fuzzy SV-k-modes algorithms with different α on MB data.

		AC	*PE*	*RE*	*ARI*	*NMI*
$\alpha = 1.1$	Fuzzy k-modes	0.7982 ± 0.1106	0.8017 ± 0.1042	0.7662 ± 0.1272	0.5645 ± 0.2106	0.5453 ± 0.1713
	Fuzzy SV-k-modes	**0.8755** ± 0.1412	**0.8785** ± 0.1354	**0.8616** ± 0.1512	**0.7333** ± 0.2526	**0.7092** ± 0.2188
$\alpha = 1.3$	Fuzzy k-modes	0.7796 ± 0.0863	0.7801 ± 0.0813	0.7475 ± 0.1214	0.5151 ± 0.1710	0.5111 ± 0.1383
	Fuzzy SV-k-modes	**0.8742** ± 0.1314	**0.8793** ± 0.1268	**0.8611** ± 0.1411	**0.7231** ± 0.2448	**0.7019** ± 0.2076
$\alpha = 1.5$	Fuzzy k-modes	0.7488 ± 0.0989	0.7531 ± 0.0950	0.7123 ± 0.1152	0.4625 ± 0.1835	0.4699 ± 0.1498
	Fuzzy SV-k-modes	**0.8656** ± 0.1083	**0.8785** ± 0.0964	**0.8488** ± 0.1262	**0.6945** ± 0.2115	**0.6807** ± 0.1690
$\alpha = 1.7$	Fuzzy k-modes	0.7258 ± 0.0670	0.7258 ± 0.0674	0.7070 ± 0.1219	0.4203 ± 0.1167	0.4275 ± 0.1001
	Fuzzy SV-k-modes	**0.8367** ± 0.1286	**0.8460** ± 0.1222	**0.8082** ± 0.1523	**0.6412** ± 0.2478	**0.6335** ± 0.2065
$\alpha = 1.9$	Fuzzy k-modes	0.7016 ± 0.0932	0.7103 ± 0.0998	0.6967 ± 0.1401	0.3792 ± 0.1540	0.3952 ± 0.1450
	Fuzzy SV-k-modes	**0.8855** ± 0.1036	**0.8953** ± 0.0947	**0.8638** ± 0.1282	**0.7379** ± 0.2010	**0.7220** ± 0.1626
$\alpha = 2.1$	Fuzzy k-modes	0.6682 ± 0.0682	0.6782 ± 0.0714	0.6882 ± 0.1390	0.3149 ± 0.1176	0.3303 ± 0.1021
	Fuzzy SV-k-modes	**0.8751** ± 0.1176	**0.8807** ± 0.1183	**0.8548** ± 0.1387	**0.7190** ± 0.2143	**0.6963** ± 0.1842
$\alpha = 2.3$	Fuzzy k-modes	0.6606 ± 0.0684	0.6680 ± 0.0710	0.6610 ± 0.1104	0.2941 ± 0.1114	0.3153 ± 0.1042
	Fuzzy SV-k-modes	**0.8829** ± 0.1186	**0.8904** ± 0.1146	**0.8661** ± 0.1346	**0.7318** ± 0.2192	**0.7100** ± 0.1908
$\alpha = 2.5$	Fuzzy k-modes	0.6469 ± 0.0627	0.6647 ± 0.0546	0.6844 ± 0.1343	0.2807 ± 0.0951	0.3107 ± 0.0814
	Fuzzy SV-k-modes	**0.9125** ± 0.0842	**0.9183** ± 0.0790	**0.9027** ± 0.0983	**0.7844** ± 0.1636	**0.7572** ± 0.1327
$\alpha = 2.7$	Fuzzy k-modes	0.6391 ± 0.0446	0.6541 ± 0.0470	0.7378 ± 0.1136	0.2693 ± 0.0702	0.2896 ± 0.0552
	Fuzzy SV-k-modes	**0.8462** ± 0.1280	**0.8561** ± 0.1225	**0.8264** ± 0.1425	**0.6609** ± 0.2357	**0.6470** ± 0.1987
$\alpha = 2.9$	Fuzzy k-modes	0.6170 ± 0.0341	0.6296 ± 0.0425	0.7389 ± 0.1345	0.2409 ± 0.0613	0.2684 ± 0.0571
	Fuzzy SV-k-modes	**0.8778** ± 0.1125	**0.8826** ± 0.1133	**0.8579** ± 0.1338	**0.7248** ± 0.2027	**0.6980** ± 0.1782

Table 5
Comparison results of the fuzzy k-modes and fuzzy SV-k-modes algorithms with different α on MW data.

		AC	*PE*	*RE*	*ARI*	*NMI*
$\alpha = 1.1$	Fuzzy k-modes	0.7498 ± 0.0954	0.7619 ± 0.0892	0.7373 ± 0.1200	0.2741 ± 0.2043	0.2399 ± 0.1588
	Fuzzy SV-k-modes	**0.8527** ± 0.0968	**0.8824** ± 0.0961	**0.8112** ± 0.1233	**0.5230** ± 0.2268	**0.4890** ± 0.2094
$\alpha = 1.3$	Fuzzy k-modes	0.7335 ± 0.0875	0.7439 ± 0.0793	0.7177 ± 0.1231	0.2347 ± 0.1860	0.2053 ± 0.1467
	Fuzzy SV-k-modes	**0.8690** ± 0.0897	**0.8987** ± 0.0915	**0.8303** ± 0.1131	**0.5663** ± 0.2063	**0.5329** ± 0.1850
$\alpha = 1.5$	Fuzzy k-modes	0.7250 ± 0.0639	0.7309 ± 0.0529	0.7076 ± 0.0998	0.2060 ± 0.1255	0.1692 ± 0.0897
	Fuzzy SV-k-modes	**0.8884** ± 0.0495	**0.9163** ± 0.0571	**0.8568** ± 0.0520	**0.6084** ± 0.1202	**0.5660** ± 0.1225
$\alpha = 1.7$	Fuzzy k-modes	0.7234 ± 0.0679	0.7345 ± 0.0557	0.7037 ± 0.1015	0.2055 ± 0.1303	0.1747 ± 0.0889
	Fuzzy SV-k-modes	**0.8708** ± 0.0870	**0.9017** ± 0.0839	**0.8314** ± 0.1120	**0.5695** ± 0.2046	**0.5357** ± 0.1838
$\alpha = 1.9$	Fuzzy k-modes	0.7375 ± 0.0707	0.7535 ± 0.0534	0.7297 ± 0.1034	0.2361 ± 0.1388	0.2126 ± 0.0909
	Fuzzy SV-k-modes	**0.8610** ± 0.0933	**0.8936** ± 0.0856	**0.8212** ± 0.1173	**0.5441** ± 0.2211	**0.5117** ± 0.1975
$\alpha = 2.1$	Fuzzy k-modes	0.7343 ± 0.0829	0.7560 ± 0.0677	0.7169 ± 0.1216	0.2332 ± 0.1690	0.2171 ± 0.1153
	Fuzzy SV-k-modes	**0.8638** ± 0.0873	**0.8932** ± 0.0887	**0.8274** ± 0.1036	**0.5509** ± 0.2046	**0.5145** ± 0.1908
$\alpha = 2.3$	Fuzzy k-modes	0.7231 ± 0.0820	0.7495 ± 0.0678	0.7039 ± 0.1239	0.2109 ± 0.1650	0.2070 ± 0.1140
	Fuzzy SV-k-modes	**0.8754** ± 0.0729	**0.9074** ± 0.0695	**0.8371** ± 0.0948	**0.5757** ± 0.1745	**0.5404** ± 0.1585
$\alpha = 2.5$	Fuzzy k-modes	0.7280 ± 0.0822	0.7509 ± 0.0678	0.7073 ± 0.1246	0.2206 ± 0.1650	0.2091 ± 0.1133
	Fuzzy SV-k-modes	**0.8629** ± 0.0882	**0.8943** ± 0.0869	**0.8253** ± 0.1056	**0.5480** ± 0.2078	**0.5144** ± 0.1892
$\alpha = 2.7$	Fuzzy k-modes	0.7256 ± 0.0834	0.7494 ± 0.0752	0.6994 ± 0.1192	0.2161 ± 0.1686	0.2040 ± 0.1219
	Fuzzy SV-k-modes	**0.8723** ± 0.0812	**0.9006** ± 0.0812	**0.8352** ± 0.1025	**0.5710** ± 0.1959	**0.5325** ± 0.1844
$\alpha = 2.9$	Fuzzy k-modes	0.7409 ± 0.0813	0.7600 ± 0.0749	0.7303 ± 0.1087	0.2473 ± 0.1674	0.2284 ± 0.1216
	Fuzzy SV-k-modes	**0.8508** ± 0.0996	**0.8890** ± 0.0818	**0.8069** ± 0.1285	**0.5166** ± 0.2374	**0.4916** ± 0.2024

that the k-means algorithm often produce "uniform effect" in clustering imbalance data sets. That is to say, the k-means clustering algorithm makes clusters have relatively uniform sizes for a given data set. In [25], we studied the uniform effect of the fuzzy k-means algorithm and found the uniform effect phenomenon becomes more obvious as the fuzziness factor α increases for imbalance data sets. In the fuzzy k-modes algorithm, Huang [5] set $\alpha = 1.1$ because it provided the least value of object function. Although there have been many studies on the selection of α for the fuzzy k-type algorithms, there is still not one generally accepted criterion [22].

In this experiment, we compared the clustering results of the fuzzy SV-k-modes and fuzzy k-modes algorithms and the size of α was set from 1.1 to 2.9 with step length of 0.2. We randomly ran 50 times the two algorithms that use the same initial cluster centers in each run. Experimental results on the three real data sets are shown in Tables 4–6. The value following "±" is the standard deviation of average values.

From Tables 4–6, we can find that the fuzzy SV-k-modes algorithm is obviously superior to the fuzzy k-modes algorithm. In addition, we can see that $\alpha = 1.1$ is the optimal value for the fuzzy k-modes algorithm and the fuzzy SV-k-modes algorithm is not sensitive to the fuzziness factor α. When $\alpha > 1.5$, the fuzzy k-modes algorithm cannot obtain effective clustering results in URM data because it usually generates one cluster in iteration process. The "———" symbol means the fuzzy k-modes algorithm cannot generate an effective partition in URM data set. Therefore, we consider that the fuzzy SV-k-modes algorithm is an effective method in clustering set-valued objects.

Table 6
Comparison results of the fuzzy k-modes and fuzzy SV-k-modes algorithms with different α on URM data.

		AC	*PE*	*RE*	*ARI*	*NMI*
$\alpha = 1.1$	Fuzzy k-modes	0.6356 ± 0.1430	0.6348 ± 0.1548	0.6094 ± 0.1510	0.3457 ± 0.2097	0.3828 ± 0.2305
	Fuzzy SV-k-modes	**0.7411** ± 0.0842	**0.7370** ± 0.0901	**0.7104** ± 0.0816	**0.5077**± 0.1273	**0.5170** ± 0.1220
$\alpha = 1.3$	Fuzzy k-modes	0.6484 ± 0.1684	0.6426 ± 0.1652	0.6202 ± 0.1727	0.3767 ± 0.2491	0.4053 ± 0.2674
	Fuzzy SV-k-modes	**0.7481** ± 0.1322	**0.7544** ± 0.1313	**0.7220** ± 0.1264	**0.5102** ± 0.2072	**0.5134** ± 0.1825
$\alpha = 1.5$	Fuzzy k-modes					
	Fuzzy SV-k-modes	0.7890 ± 0.1126	0.8183 ± 0.1143	0.7602 ± 0.1117	0.5720 ± 0.1779	0.5808 ± 0.1773
$\alpha = 1.7$	Fuzzy k-modes					
	Fuzzy SV-k-modes	0.7475 ± 0.1016	0.7696 ± 0.0943	0.7200 ± 0.1041	0.5168 ± 0.1549	0.5193 ± 0.1332
$\alpha = 1.9$	Fuzzy k-modes					
	Fuzzy SV-k-modes	0.7405 ± 0.0999	0.7480 ± 0.0978	0.7029 ± 0.1032	0.5012 ± 0.1558	0.5091 ± 0.1456
$\alpha = 2.1$	Fuzzy k-modes					
	Fuzzy SV-k-modes	0.7511 ± 0.0712	0.7606 ± 0.0942	0.7107 ± 0.0847	0.5078 ± 0.1259	0.5186 ± 0.1141
$\alpha = 2.3$	Fuzzy k-modes					
	Fuzzy SV-k-modes	0.7815 ± 0.0616	0.8162 ± 0.0775	0.7467 ± 0.0703	0.5645 ± 0.0905	0.5724 ± 0.0930
$\alpha = 2.5$	Fuzzy k-modes					
	Fuzzy SV-k-modes	0.7730 ± 0.1296	0.7885 ± 0.1499	0.7520 ± 0.1216	0.5523 ± 0.2002	0.5580 ± 0.1872
$\alpha = 2.7$	Fuzzy k-modes					
	Fuzzy SV-k-modes	0.8001 ± 0.0745	0.8138 ± 0.1049	0.7663 ± 0.0867	0.5879 ± 0.1040	0.5804 ± 0.1041
$\alpha = 2.9$	Fuzzy k-modes					
	Fuzzy SV-k-modes	0.8012 ± 0.1281	0.8154 ± 0.1345	0.7753 ± 0.1341	0.5861 ± 0.2090	0.5705 ± 0.1925

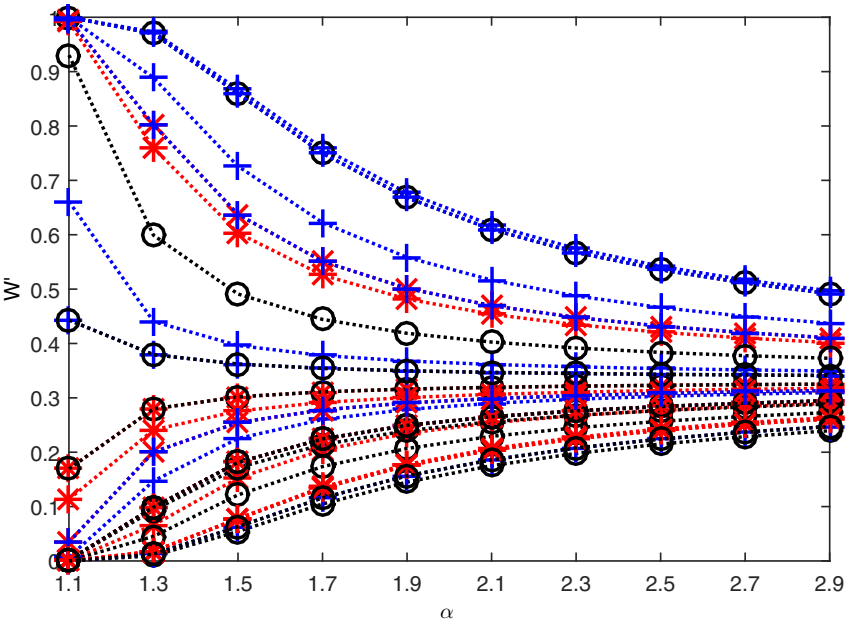

Fig. 8. Relationship between α and membership degrees on MB data.

*5.4. Relationship between α and **W'***

The size of α affects the membership degrees that an object is assigned to of different clusters. In this experiment, we analyzed the relationship between α and $\mathbf{W}'$ on the three data sets and α was set from 1.1 to 2.9 with the step length 0.2. For convenience, in each data set we only visualized the variety of the membership degrees of the first 10 objects with α increasing. The relationship between α and $\mathbf{W}'$ on the three real data sets is shown in Figs. 8–10, where the symbols "$*$", "+" and "o" represent different cluster labels, respectively.

From Figs. 8–10, we can see that the membership degrees that an object is assigned to different clusters decrease as α increases.

6. Conclusions

In real applications, data sets with set-valued characteristic have become ubiquitous. In this paper, we proposed a fuzzy SV-k-modes algorithm that is an extension version of the fuzzy k-modes algorithm for clustering data with set-valued

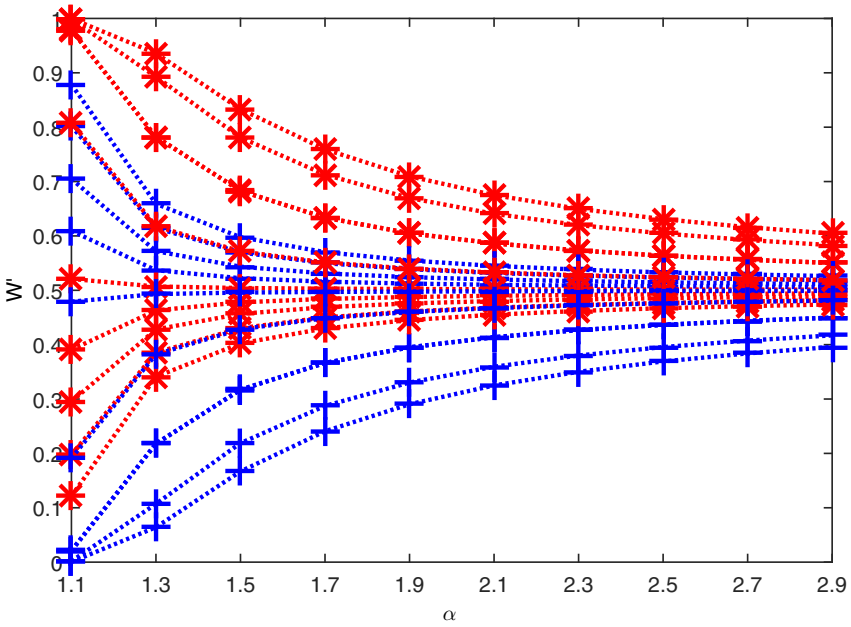

Fig. 9. Relationship between α and membership degrees on MW data.

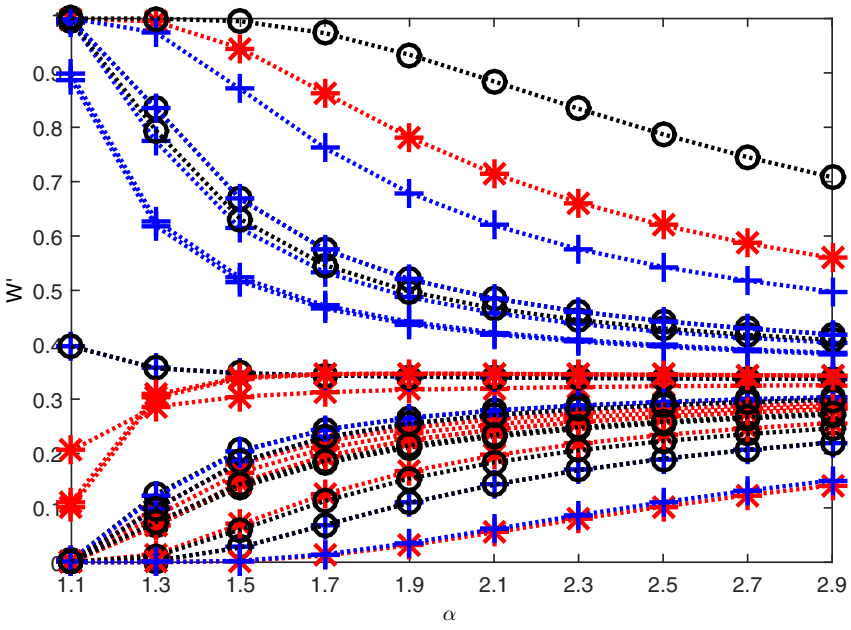

Fig. 10. Relationship between α and membership degrees on URM data.

attributes. In the proposed algorithm, we defined the distance between two set-valued objects and gave the representation and heuristic update ways of cluster prototype. Experimental results on the synthetic and real data sets have shown the efficiency and effectiveness of the fuzzy SV-*k*-modes algorithm in clustering data with set-valued attributes. These modifications made the fuzzy SV-*k*-modes algorithm can cluster data with single-valued and set-valued attributes together and the fuzzy *k*-modes algorithm is its special case.

Acknowledgments

The authors would like to thank Prof. Jian Pei at Simon Fraser University for his valuable suggestions. We are also very grateful to the editor and reviewers for their valuable comments on our paper. This work was supported by the

National Natural Science Foundation of China (under Grants 61573229, 61473194, 61305073, 61432011 and U1435212), the Natural Science Foundation of Shanxi Province (under Grant 2015011048), the Shanxi Scholarship Council of China (under Grant 2016–003) and the National Key Basic Research and Development Program of China (973) (under Grant 2013CB329404).

References

[1] A.K. Jain, Data clustering: 50 years beyond *k*-means, Pattern Recognit. Lett. 31 (8) (2010) 651–666.
[2] M.-Y. Cheng, K.-Y. Huang, H.-M. Chen, *k*-means particle swarm optimization with embedded chaotic search for solving multidimensional problems, Appl. Math. Comput. 219 (6) (2012) 3091–3099.
[3] Z. Huang, Extensions to the *k*-means algorithm for clustering large data sets with categorical values, Data Min. Knowl. Discov. 2 (3) (1998) 283–304.
[4] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, California, USA, 1967, pp. 281–297.
[5] Z. Huang, M.K. Ng, A fuzzy *k*-modes algorithm for clustering categorical data, IEEE Trans. Fuzzy Syst. 7 (4) (1999) 446–452.
[6] J.C. Bezdek, A convergence theorem for the fuzzy isodata clustering algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 2 (1) (1980) 1–8.
[7] S.W. Purnami, J.M. Zain, T. Heriawan, An alternative algorithm for classification large categorical dataset: *k*-mode clustering reduced support vector machine, Int. J. Database Theory Appl. 4 (1) (2011) 19–30.
[8] M. Al-Razgan, C. Domeniconi, D. Barbará, Random subspace ensembles for clustering categorical data, in: Supervised and Unsupervised Ensemble Methods and Their Applications, Springer, 2008, pp. 31–48.
[9] T.A. Thornton-Wells, J.H. Moore, J.L. Haines, Dissecting trait heterogeneity: a comparison of three clustering methods applied to genotypic data, BMC Bioinf. 7 (1) (2006) 204.
[10] B. Andreopoulos, A. An, X. Wang, Clustering the internet topology at multiple layers, WSEAS Trans. Inf. Sci. Appl. 2 (10) (2005) 1625–1634.
[11] V. Manganaro, S. Paratore, E. Alessi, S. Coffa, S. Cavallaro, Adding semantics to gene expression profiles: new tools for drug discovery, Curr. Med. Chem. 12 (10) (2005) 1149–1160.
[12] F. Cao, J.Z. Huang, J. Liang, Trend analysis of categorical data streams with a concept change method, Inf. Sci. 276 (2014) 160–173.
[13] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, third ed., Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011.
[14] H. Ralambondrainy, A conceptual version of the *k*-means algorithm, Pattern Recognit. Lett. 16 (11) (1995) 1147–1157.
[15] F. Giannotti, C. Gozzi, G. Manco, Clustering transactional data, in: Principles of Data Mining and Knowledge Discovery, Springer, 2002, pp. 175–187.
[16] M.K. Ng, M.J. Li, J.Z. Huang, Z. He, On the impact of dissimilarity measure in *k*-modes clustering algorithm, IEEE Trans. Pattern Anal. Mach. Intell. 29 (3) (2007) 503–507.
[17] P. Jaccard, Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines, Bulletin de la Société Vaudoise des Sciences Naturelles 37 (1901) 241–272.
[18] S. Schiffman, L. Reynolds, F. Young, Introduction to Multidimensional Scaling: Theory, Methods, and Applications, Academic Press, 1981.
[19] K. Bache, M. Lichman, UCI machine learning repository, 2014.
[20] N.R. Pal, J.C. Bezdek, On cluster validity for the fuzzy c-means model, IEEE Trans. Fuzzy Syst. 3 (3) (1995) 370–379.
[21] J. Yu, Q. Cheng, H. Huang, Analysis of the weighting exponent in the FCM, IEEE Trans. Syst. Man Cybern. Part B Cybern. 34 (1) (2004) 634–639.
[22] K. Zhou, C. Fu, S. Yang, Fuzziness parameter selection in fuzzy c-means: the perspective of cluster validation, Sci. China Inf. Sci. 57 (11) (2014) 1–8.
[23] K.-L. Wu, Analysis of parameter selections for fuzzy c-means, Pattern Recognit. 45 (1) (2012) 407–415.
[24] H. Xiong, J. Wu, J. Chen, *k*-means clustering versus validation measures: a data-distribution perspective, IEEE Trans. Syst. Man Cybern. Part B Cybern. 39 (2) (2009) 318–331.
[25] J. Liang, L. Bai, C. Dang, F. Cao, The *k*-means type algorithms versus imbalanced data distributions, IEEE Trans. Fuzzy Syst. 20 (4) (2012) 728–745.