A Fuzzy *k*-Modes Algorithm for Clustering Categorical Data

Zhexue Huang and Michael K. Ng

***Abstract*— This correspondence describes extensions to the fuzzy k-means algorithm for clustering categorical data. By using a simple matching dissimilarity measure for categorical objects and modes instead of means for clusters, a new approach is developed, which allows the use of the k-means paradigm to efficiently cluster large categorical data sets. A fuzzy k-modes algorithm is presented and the effectiveness of the algorithm is demonstrated with experimental results.**

***Index Terms*—Categorical data, clustering, data mining, fuzzy partitioning, k-means algorithm.**

I. Introduction

THE k-means algorithm [1], [2], [8], [11] is well known for its efficiency in clustering large data sets. Fuzzy versions of the k-means algorithm have been reported in Ruspini [15] and Bezdek [3], where each pattern is allowed to have membership functions to all clusters rather than having a distinct membership to exactly one cluster. However, working only on numeric data limits the use of these k-means-type algorithms in such areas as data mining where large categorical data sets are frequently encountered.

Ralambondrainy [13] presented an approach to using the k-means algorithm to cluster categorical data. His approach converts multiple categorical attributes into binary attributes, each using one for presence of a category and zero for absence of it, and then treats these binary attributes as numeric ones in the k-means algorithm. This approach needs to handle a large number of binary attributes when data sets have attributes with many categories. This will inevitably increase both computational cost and memory storage of the k-means algorithm. The other drawback is that the cluster means given by real values between zero and one do not indicate the characteristics of the clusters.

Other algorithms for clustering categorical data include hierarchical clustering methods using Gower's similarity coefficient [6] or other dissimilarity measures [5], the PAM algorithm [9], the fuzzy-statistical algorithms [18], and the conceptual clustering methods [12]. All these methods suffer from a common efficiency problem when applied to massive categorical-only data sets. For instance, the computational complexity of most hierarchical clustering methods is $O(n^2)$ [1] and the PAM algorithm has the complexity of $O(k(n-k)^2)$ per iteration [14], where n is the size of data set and k is the number of clusters.

Manuscript received September 2, 1997; revised October 28, 1998. This work was supported in part by HKU CRCG under Grants 10 201 824 and 10 201 939.

Z. Huang is with Management Information Principles, Ltd., Melbourne, Australia.

M. K. Ng is with the Department of Mathematics, The University of Hong Kong, Hong Kong.

Publisher Item Identifier S 1063-6706(99)06705-3.

To tackle the problem of clustering large categorical data sets in data mining, the k-modes algorithm has recently been proposed in [7]. The k-modes algorithm extends the k-means algorithm by using a simple matching dissimilarity measure for categorical objects, modes instead of means for clusters, and a frequency-based method to update modes in the clustering process to minimize the clustering cost function. These extensions have removed the numeric-only limitation of the k-means algorithm and enable it to be used to efficiently cluster large categorical data sets from real-world databases.

In this paper, we introduce a fuzzy k-modes algorithm which generalizes our previous work in [7]. This is achieved by the development of a new procedure to generate the fuzzy partition matrix from categorical data within the framework of the fuzzy k-means algorithm [3]. The main result of this paper is to provide a method to find the fuzzy cluster modes when the simple matching dissimilarity measure is used for categorical objects. The fuzzy version has improved the k-modes algorithm by assigning confidence to objects in different clusters. These confidence values can be used to decide the core and boundary objects of clusters, thereby providing more useful information for dealing with boundary objects.

II. Notation

We assume the set of objects to be clustered is stored in a database table $\boldsymbol{T}$ defined by a set of attributes $A_1, A_2, \cdots, A_m$. Each attribute A_j describes a domain of values denoted by $DOM(A_j)$ and associated with a defined semantic and a data type. In this letter, we only consider two general data types, *numeric* and *categorical* and assume other types used in database systems can be mapped to one of these two types. The domains of attributes associated with these two types are called numeric and categorical, respectively. A numeric domain consists of real numbers. A domain $DOM(A_j)$ is defined as categorical if it is finite and unordered, e.g., for any $a, b \in DOM(A_j)$, either $a = b$ or $a \neq b$, see for instance [5].

An object X in $\boldsymbol{T}$ can be logically represented as a conjunction of attribute-value pairs $[A_1 = x_1] \wedge [A_2 = x_2] \wedge \cdots \wedge [A_m = x_m]$, where $x_j \in DOM(A_j)$ for $1 \leq j \leq m$. Without ambiguity, we represent X as a vector $[x_1, x_2, \cdots, x_m]$. X is called a categorical object if it has only categorical values. We consider every object has exactly m attribute values. If the value of an attribute A_j is missing, then we denote the attribute value of A_j by ϵ.

Let $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$ be a set of n objects. Object X_i is represented as $[x_{i,1}, x_{i,2}, \cdots, x_{i,m}]$. We write $X_i = X_k$ if $x_{i,j} = x_{k,j}$ for $1 \leq j \leq m$. The relation $X_i = X_k$ does not mean that X_i and X_k are the same object in the real-world

database, but rather that the two objects have equal values in attributes $A_1, A_2, \cdots, A_m$.

III. Hard and Fuzzy k-Means Algorithms

Let $\boldsymbol{X}$ be a set of n objects described by m numeric attributes. The hard and fuzzy k-means clustering algorithms to cluster $\boldsymbol{X}$ into k clusters can be stated as the algorithms [3], which attempt to minimize the cost function

$$F(W, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{li}^{\alpha}\, d(Z_l, X_i) \tag{1}$$

subject to

$$0 \le w_{li} \le 1, \qquad 1 \le l \le k,\ 1 \le i \le n \tag{2}$$

$$\sum_{l=1}^{k} w_{li} = 1, \qquad 1 \le i \le n \tag{3}$$

and

$$0 < \sum_{i=1}^{n} w_{li} < n, \qquad 1 \le l \le k \tag{4}$$

where $k(\le n)$ is a known number of clusters, $\alpha \in [1, \infty)$ is a weighting exponent, $W = [w_{li}]$ is a k-by-n real matrix, $Z = [Z_1, Z_2, \cdots, Z_k] \in \mathbb{R}^{mk}$,, and $d(Z_l, X_i)(\ge 0)$ is some dissimilarity measure between Z_l and X_i.

Minimization of F in (1) with the constraints in (2)–(4) forms a class of constrained nonlinear optimization problems whose solutions are unknown. The usual method toward optimization of F in (1) is to use partial optimization for Z and W [3]. In this method, we first fix Z and find necessary conditions on W to minimize F. Then we fix W and minimize F with respect to Z. This process is formalized in the k-means algorithm as follows.

Algorithm 1—The k-Means Algorithm:

1) Choose an initial point $Z^{(1)} \in \mathbb{R}^{mk}$. Determine $W^{(1)}$ such that $F(W, Z^{(1)})$ is minimized. Set $t = 1$.
2) Determine $Z^{(t+1)}$ such that $F(W^{(t)}, Z^{(t+1)})$ is minimized. If $F(W^{(t)}, Z^{(t+1)}) = F(W^{(t)}, Z^{(t)})$—then stop; otherwise go to step 3).
3) Determine $W^{(t+1)}$ such that $F(W^{(t+1)}, Z^{(t+1)})$ is minimized. If $F(W^{(t+1)}, Z^{(t+1)}) = F(W^{(t)}, Z^{(t+1)})$—then stop; otherwise set $t = t + 1$ and go to step 2).

The matrices Z and W are calculated according to the following two theorems.

Theorem 1: Let $\hat{Z}$ be fixed and consider Problem (P1)

$$\min_{W} F(W, \hat{Z}) \quad \text{subject to} \quad (2), (3), \text{ and } (4).$$

For $\alpha = 1$, the minimizer $\hat{W}$ of Problem (P1) is given by

$$\hat{w}_{li} = \begin{cases} 1, & \text{if } d(\hat{Z}_l, X_i) \le d(\hat{Z}_h, X_i), \quad 1 \le h \le k \\ 0, & \text{otherwise.} \end{cases}$$

For $\alpha > 1$, the minimizer $\hat{W}$ of Problem (P1) is given by

$$\hat{w}_{li} = \begin{cases} 1, & \text{if } X_i = \hat{Z}_l \\ 0, & \text{if } X_i = \hat{Z}_h,\ h \ne l \\ 1 \Big/ \displaystyle\sum_{h=1}^{k} \left[\frac{d(\hat{Z}_l, X_i)}{d(\hat{Z}_h, X_i)} \right]^{1/(\alpha-1)}, & \text{if } X_i \ne \hat{Z}_l \text{ and } X_i \ne \hat{Z}_h,\ 1 \le h \le k. \end{cases} \tag{5}$$

for $1 \le l \le k$ and $1 \le i \le n$.

The proof of Theorem 1 can be found in [3], [17]. We remark that for the case of $\alpha = 1$ the minimum solution $\hat{W}$ is not unique, so $w_{li} = 1$ may arbitrarily be assigned to the first minimizing index l, and the remaining entries of this column are put to zero.

In the literature the Euclidean norm $d(X, Y) = \sqrt{\sum_{j=1}^{m} |x_j - y_j|^2}$ is often used in the k-means algorithm. In this case, the following result holds [3], [4].

Theorem 2: Let $\hat{W}$ be fixed and consider Problem (P2)

$$\min_{Z} F(\hat{W}, Z)$$

where $d(Z_l, X_i)$ is the Euclidean norm. Then the minimizer $\hat{Z}$ of Problem (P2) is given by

$$\hat{Z}_i = \frac{\displaystyle\sum_{i=1}^{n} w_{li}^{\alpha} X_i}{\displaystyle\sum_{i=1}^{n} w_{li}^{\alpha}}, \qquad 1 \le l \le k.$$

Most k-means-type algorithms have been proved convergent and often terminate at a local minimum (see for instance [3], [4], [11], [16], [17]). The computational complexity of the algorithm is $O(tkmn)$ operations, where t is the number of iterations, k is the number of clusters, m is the number of attributes, and n is the number of objects. When $n \gg t, m, k$, it is faster than the hierarchical clustering algorithms whose computational complexity is generally $O(n^2)$ [1]. As for the storage, we need $O(n(m + k) + km)$ space to hold the set of n objects, the cluster centers Z, and the partition matrix W, which, for a large n, is much less than that required by the hierarchical clustering algorithms. Therefore, the k-means algorithm is best suited for dealing with large data sets. However, working only on numeric values limits its use in applications such as data mining in which categorical values are frequently encountered. This limitation is removed in the hard and fuzzy k-modes algorithms to be discussed in the next section.

IV. Hard and Fuzzy K-Modes Algorithms

The hard k-modes algorithm, first introduced in [7], has made the following modifications to the k-means algorithm: 1) using a simple matching dissimilarity measure for categorical objects; 2) replacing the means of clusters with the modes; and 3) using a frequency-based method to find the modes to solve Problem (P2). These modifications have removed the

numeric-only limitation of the k-means algorithm but maintain its efficiency in clustering large categorical data sets [7].

Let X and Y be two categorical objects represented by $[x_1, x_2, \cdots, x_m]$ and $[y_1, y_2, \cdots, y_m]$, respectively. The simple matching dissimilarity measure between X and Y is defined as follows:

$$d_c(X, Y) \equiv \sum_{j=1}^{m} \delta(x_j, y_j) \tag{6}$$

where

$$\delta(x_j, y_j) = \begin{cases} 0, & x_j = y_j \\ 1, & n_j \neq y_j. \end{cases}$$

It is easy to verify that the function d_c defines a metric space on the set of categorical objects. Traditionally, the simple matching approach is often used in binary variables which are converted from categorical variables [9, pp. 28–29]. We note that d_c is also a kind of generalized Hamming distance [10].

The k-modes algorithm uses the k-means paradigm to cluster categorical data. The objective of clustering a set of n categorical objects into k clusters is to find W and Z that minimize

$$F_c(W, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{li}^{\alpha} d_c(Z_l, X_i) \tag{7}$$

with other conditions same as in (1). Here, Z represents a set of k modes for k clusters.[1] We can still use Algorithm 1 to minimize $F_c(W, Z)$. However, the way to update Z at each iteration is different from the method given in Theorem 2. For the hard k-partition (i.e., $\alpha = 1$), Huang [7] has presented a frequency-based method to update Z. This method can be described as follows.

Theorem 3—The Hard k-Modes Update Method: Let $\boldsymbol{X}$ be a set of categorical objects described by categorical attributes $A_1, A_2, \cdots, A_m$ and $DOM(A_j) = \{a_j^{(1)}, a_j^{(2)}, \cdots, a_j^{(n_j)}\}$, where n_j is the number of categories of attribute A_j for $1 \leq j \leq m$. Let the cluster centers Z_l be represented by $[z_{l,1}, z_{l,2}, \cdots, z_{l,m}]$ for $1 \leq l \leq k$. Then the quantity $\sum_{l=1}^{k} \sum_{i=1}^{n} w_{li} d_c(Z_l, X_i)$ is minimized iff $z_{l,j} = a_j^{(r)} \in DOM(A_j)$ where

$$|\{w_{li} | x_{i,j} = a_j^{(r)}, w_{lj} = 1\}| \geq |\{w_{li} | x_{i,j} = a_j^{(t)}, w_{li} = 1\}|, \qquad 1 \leq t \leq n_j \tag{8}$$

for $1 \leq j \leq m$. Here, $|X|$ denotes the number of elements in the set X.

Proof: For a given W, all the inner sums of the quantity $\sum_{l=1}^{k} \sum_{i=1}^{n} w_{li} d_c(Z_l, X_i)$ are nonnegative and independent. Minimizing the quantity is equivalent to minimizing each inner sum. We write the lth inner sum ($1 \leq l \leq k$) as

$$\sum_{i=1}^{n} w_{li} d_c(Z_l, X_i) = \sum_{i=1}^{n} w_{li} \sum_{j=1}^{m} \delta(z_{l,j}, x_{i,j}) = \sum_{j=1}^{m} \sum_{i=1}^{n} w_{li} \delta(z_{l,j}, x_{i,j}) = \sum_{j=1}^{m} n \left(1 - \frac{\{w_{li} | x_{i,j} = z_{l,j}, w_{li} = 1\}|}{n}\right).$$

The inner sum is minimized iff every term $(1 - [|\{w_{li} | x_{i,j} = z_{l,j}, w_{li} = 1\}|]/n)$ is minimal for $1 \leq j \leq m$. Thus the term $|\{w_{li} | x_{i,j} = z_{l,j}, w_{li} = 1\}|$ must be maximal. The result follows. ■

According to (8), the category of attribute A_j of the cluster mode Z_l is determined by the mode of categories of attribute A_j in the set of objects belonging to cluster l.

The main problem addressed in the present paper is to find the fuzzy cluster modes ($\alpha > 1$) when the dissimilarity measure defined in (6) is used.

Theorem 4—The Fuzzy k-Modes Update Method: Let $\boldsymbol{X}$ be a set of categorical objects described by categorical attributes $A_1, A_2, \cdots, A_m$ and $DOM(A_j) = \{a_j^{(1)}, a_j^{(2)}, \cdots, a_j^{(n_j)}\}$, where n_j is the number of categories of attribute A_j for $1 \leq j \leq m$. Let the cluster centers Z_l be represented by $[z_{l,1}, z_{l,2}, \cdots, z_{l,m}]$ for $1 \leq l \leq k$. Then the quantity $\sum_{l=1}^{k} \sum_{i=1}^{n} w_{li}^{\alpha} d_c(Z_l, X_i)$ is minimized iff $z_{l,j} = a_j^{(r)} \in DOM(A_j)$ where

$$\sum_{i, x_{i,j} = a_j^{(r)}} w_{li}^{\alpha} \geq \sum_{i, x_{i,j} = a_j^{(t)}} w_{li}^{\alpha}, \qquad 1 \leq t \leq n_j \tag{9}$$

for $1 \leq j \leq m$.

Proof: For a given W, all the inner sums of the quantity $\sum_{l=1}^{k} \sum_{i=1}^{n} w_{li}^{\alpha} d_c(Z_l, X_i)$ are nonnegative and independent. Minimizing the quantity is equivalent to minimizing each inner sum. We write the lth inner sum ($1 \leq l \leq k$) as

$$\sum_{i=1}^{n} w_{li}^{\alpha} d_c(Z_l, X_i) = \sum_{i=1}^{n} w_{li}^{\alpha} \sum_{j=1}^{m} \delta(z_{l,j}, x_{i,j}) = \sum_{j=1}^{m} \sum_{i=1}^{n} w_{li}^{\alpha} \delta(z_{l,j}, x_{i,j}) = \sum_{j=1}^{m} \left[\sum_{t=1}^{n_j} \sum_{i, x_{i,j} = a_j^{(t)}} w_{li}^{\alpha} - \sum_{i, x_{i,j} = z_{l,j}} w_{li}^{\alpha} \right].$$

Since w_{li}^{α} is fixed and nonnegative for $1 \leq l \leq k$ and $1 \leq i \leq n$, the quantity $\sum_{t=1}^{n_j} \sum_{i, x_{i,j} = a_j^{(t)}} w_{li}^{\alpha}$ is fixed and nonnegative. It follows that $\sum_{i=1}^{n} w_{li}^{\alpha} d_c(Z_l, X_i)$ is minimized iff each term $\sum_{i, x_{i,j} = z_{l,j}} w_{li}^{\alpha}$ is maximal. Hence, the result follows. ■

According to Theorem 4, the category of attribute A_j of the cluster mode Z_l is given by the category that achieves the maximum of the summation of w_{li} to cluster l over all categories. If the minimum is not unique, then the attribute of the cluster mode may arbitrarily assigned to the first minimizing index t in (9). Combining Theorems 1 and 4 with Algorithm 1 forms the fuzzy k-modes algorithm in which the modes of clusters in each iteration are updated according

[1] The mode for a set of categorical objects $\{X_1, X_2, \cdots, X_n\}$ is defined as an object Z that minimizes $\sum_{i=1}^{n} d_c(X_i, Z)$ [7].

TABLE I
(a) NUMBER OF OPERATIONS REQUIRED IN THE FUZZY k-MODES ALGORITHM AND (b) THE CONCEPTUAL VERSION OF THE k-MEANS ALGORITHM. HERE $M = \sum_{j=1}^{m} n_j$

Steps	Operations
1 (Initialization)	$O(kmn)$
2	$O(kMn)$
3	$O(kmn)$

(a)

Steps	Operations
1 (Initialization)	$O(kMn)$
2	$O(kMn)$
3	$O(kMn)$

(b)

to Theorem 4 and the fuzzy partition matrix is computed according to Theorem 1. The hard k-mode algorithm [7] is a special case where $\alpha = 1$.

Theorem 5: Let $\alpha \geq 1$. The fuzzy k-modes algorithm converges in a finite number of iterations.

Proof: We first note that there are only a finite number ($N = \prod_{j=1}^{m} n_j$) of possible cluster centers (modes). We then show that each possible center appears at most once by the fuzzy k-modes algorithm. Assume that $Z^{(t_1)} = Z^{(t_2)}$ where $t_1 \neq t_2$. According to the fuzzy k-modes algorithm we can compute the minimizers $W^{(t_1)}$ and $W^{(t_2)}$ of Problem (P1) for $Z = Z^{(t_1)}$ and $Z = Z^{(t_2)}$, respectively. Therefore, we have

$$F_c(W^{(t_1)}, Z^{(t_1)}) = F_c(W^{(t_1)}, Z^{(t_2)}) = F_c(W^{(t_2)}, Z^{(t_2)}).$$

However, the sequence $F_c(\cdot, \cdot)$ generated by the hard and fuzzy k-modes algorithm is strictly decreasing. Hence the result follows. ∎

We remark that the similar proof concerning the convergence in a finite number of iterations can be found in [16]. We now consider the cost of the fuzzy k-modes algorithm. The computational cost in each step of the fuzzy k-modes algorithm and the conceptual version of the k-means algorithm [13] are given in Table I according to Algorithm 1 and Theorems 1, 2, and 4. The computational complexities of steps 2 and 3 of the fuzzy k-modes algorithm and the conceptual version of the k-means algorithm are $O(kn(m + M))$ and $O(2kMn)$, respectively. Here k is the number of clusters, m is the number of attributes, M ($= \sum_{j=1}^{m} n_j$) is the total number of categories of all attributes, and n is the number of objects. We remark that we need to transform multiple categorical attributes into binary attributes as numeric values in the conceptual version of the k-means algorithm. Thus, when M is large, the cost of the fuzzy k-modes algorithm is significantly less than that of the conceptual version of the k-means algorithm. Similar to the fuzzy k-means-type algorithm, our method requires $O(m(n+k)+km)$ storage space to hold the set of objects $\{X_i\}$, the cluster centers Z and the partition matrix W.

V. EXPERIMENTAL RESULTS

To evaluate the performance and efficiency of the fuzzy k-modes algorithm and compare it with the conceptual k-means algorithm [13] and the hard k-modes algorithm, we carried out several tests of these algorithms on both real and artificial data. The test results are discussed below.

A. Clustering Performance

The first data set used was the soybean disease data set [12]. We chose this data set to test these algorithms because all attributes of the data can be treated as categorical. The soybean data set has 47 records, each being described by 35 attributes. Each record is labeled as one of the four diseases: Diaporthe Stem Canker, Charcoal Rot, Rhizoctonia Root Rot, and Phytophthora Rot. Except for Phytophthora Rot which has 17 records, all other diseases have ten records each. Of the 35 attributes, we only selected 21 because the other 14 have only one category.

We used the three clustering algorithms to cluster this data set into four clusters. The initial means and modes were randomly selected k distinct records from the data set. For the conceptual k-means algorithm, we first converted multiple categorical attributes into binary attributes, using zero for absence of a category and one for presence of it. The binary values of the attributes were then treated as numeric values in the k-means algorithm. For the fuzzy k-modes algorithm we specified $\alpha = 1.1$ (we tried several values of α and found that $\alpha = 1.1$ provides the least value of the cost function F). Unlike the other two algorithms the fuzzy k-modes algorithm produced a fuzzy partition matrix W. We obtained the cluster memberships from W as follows. The record X_i was assigned to the lth cluster if $w_{li} = \max_{1 \leq h \leq k}\{w_{hi}\}$. If the maximum was not unique, then X_i was assigned to the cluster of first achieving the maximum.

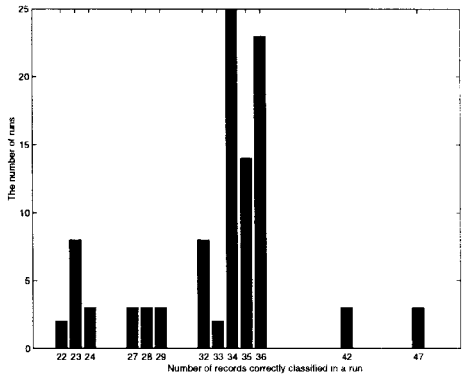
A clustering result was measured by the clustering accuracy r defined as

$$r = \frac{\sum_{l=1}^{k} a_l}{n}$$

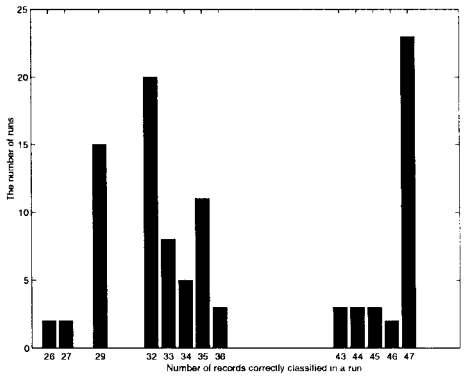
where a_l was the number of instances occurring in both cluster l and its corresponding class and n was the number of instances in the data set. In our numerical tests k is equal to four.

Each algorithm was run 100 times. Table II gives the average accuracy (i.e., the average percentages of the correctly classified records over 100 runs) of clustering by each algorithm and the average central processing unit (CPU) time used. Fig. 1 shows the distributions of the number of runs with respect to the number of records correctly classified by each algorithm. The overall clustering performance of both hard and fuzzy k-modes algorithms was better than that of the conceptual k-means algorithm. Moreover, the number of runs with correct classifications of more than 40 records ($r > 0.87$) was much larger from both hard and fuzzy k-modes algorithms than that from the conceptual k-means algorithm. The fuzzy k-modes algorithm slightly outperformed the hard k-modes algorithm in the overall performance. The average CPU time used by the k-modes-type algorithms was much smaller than that by the conceptual k-means algorithm.

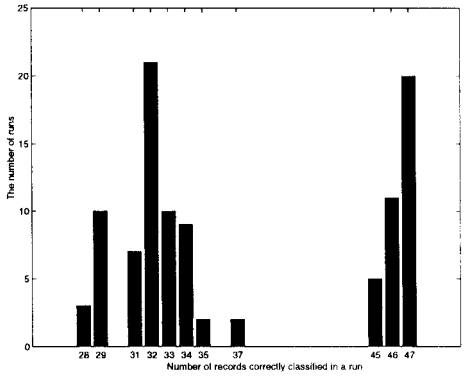
To investigate the differences between the hard and fuzzy k-modes algorithms, we compared two clustering results pro-



(a)



(b)



(c)

Fig. 1. Distributions of the number of runs with respect to the number of the correctly classified records in each run. (a) The conceptual version of the k-means algorithm. (b) The hard k-modes algorithm. (c) The fuzzy k-modes algorithm.

TABLE II
THE AVERAGE CLUSTERING ACCURACY AND AVERAGE CPU TIME IN SECONDS BY DIFFERENT CLUSTERING METHODS

	Conceptual k-means	Hard k-modes	Fuzzy k-modes
Accuracy	0.704	0.782	0.790
CPU time in seconds	0.164	0.024	0.034

duced by them from the same initial modes. Table III gives the modes of four clusters produced by the two algorithms. The modes obtained with the two algorithms are not identical. This indicates that the hard and fuzzy k-modes algorithms

TABLE III
THE MODES OF FOUR CLUSTERS PRODUCED BY (a) HARD k-MODES AND (b) FUZZY k-MODES ALGORITHMS

Z_i	Attributes						
	1	2	3	4	5	6	7
1	0	1	2	0	0	3	1
2	1	1	2	0	0	2	1
3	3	0	2	1	0	1	0
4	5	0	0	2	1	1	2
Z_i	Attributes						
	8	9	10	11	12	13	14
1	1	0	2	0	0	1	1
2	2	1	0	1	0	2	2
3	1	0	2	1	0	3	1
4	1	0	0	1	0	0	3
Z_i	Attributes						
	15	16	17	18	19	20	21
1	0	1	0	0	0	3	0
2	0	0	0	0	0	3	1
3	1	1	0	0	0	0	0
4	0	0	0	2	1	0	0

(a)

Z_i	Attributes						
	1	2	3	4	5	6	7
1	0	1	2	0	0	3	1
2	1	1	2	1	0	3	1
3	6	0	2	1	0	1	0
4	6	0	0	2	1	1	2
Z_i	Attributes						
	8	9	10	11	12	13	14
1	2	0	2	0	0	1	1
2	2	1	0	1	0	2	2
3	1	0	2	1	0	3	1
4	1	0	0	1	0	0	3
Z_i	Attributes						
	15	16	17	18	19	20	21
1	0	1	0	0	0	3	0
2	0	0	0	0	0	3	1
3	1	1	0	0	0	0	0
4	0	0	0	2	1	0	0

(b)

indeed produce different clusters. By looking at the accuracies of the two clustering results, we found that the number of records correctly classified by the hard k-modes algorithm was 43 while the number of records correctly classified by the fuzzy k-modes algorithm was 45. In this case, there was 4.2% increase of accuracy by the fuzzy k-modes algorithm. We found such an increase occurred in most cases. However, in a few cases, the clustering results produced by the hard k-modes algorithm were better than those by the fuzzy k-modes algorithm (see Fig. 1).

The partition matrix produced by the fuzzy k-modes algorithm provides useful information for identification of the boundary objects which scatter in the cluster boundaries. This can be shown by the following example. Five records are listed in Table IV together with their dissimilarity values to their corresponding modes, their part of partition matrices, their cluster memberships assigned and their true classes. In Table IV, $*$ denotes the misclassified records. In the clustering result of the hard k-modes algorithm [Table IV(a)], four records X_5, X_{33}, X_{39}, and X_{42} were misclassified. The misclassification of records X_5 and X_{39} was due to the same dissimilarities to the

TABLE IV
(a) THE DISSIMILARITY MEASURE BETWEEN MISCLASSIFIED RECORDS AND THE CLUSTER CENTERS AND THE CORRESPONDING PARTITION MATRICES PRODUCED BY THE HARD k-MODES AND (b) FUZZY k-MODES ALGORITHMS. HERE THE MISCLASSIFIED OBJECTS ARE DENOTED BY *

Record X_j	$d_c(Z_i, X_j)$ 1	2	3	4
1	7	6	13	15
*5	8	8	10	14
*33	7	8	10	12
*39	6	6	11	13
*42	5	7	10	14

w_{ij} 1	2	3	4	cluster assigned	true class
0	1	0	0	2	2
1	0	0	0	1	2
1	0	0	0	1	2
1	0	0	0	1	2
1	0	0	0	1	2

(a)

Record X_j	$d_c(Z_i, X_j)$ 1	2	3	4
*1	6	6	13	15
*5	7	7	10	14
33	8	7	10	12
39	7	4	11	13
42	6	5	10	14

w_{ij} 1	2	3	4	cluster assigned	true class
0.4999	0.4999	0.0001	0.0001	1	2
0.4928	0.4928	0.0139	0.0005	1	2
0.2030	0.7717	0.0218	0.0035	2	2
0.0037	0.9963	0.0000	0.0000	2	2
0.1389	0.8602	0.0008	0.0000	2	2

(b)

modes of clusters Z_1 and Z_2. In such a situation the algorithm arbitrarily assigned them to the first cluster. Such records are called boundary objects, which often cause problems in classification. Some of these misclassifications can be corrected by the fuzzy k-modes algorithm. For instance, in Table IV(b), the classification of object X_{39} was corrected because it has different dissimilarities to the modes of clusters Z_1 and Z_2. However, object X_5 still has a problem. Furthermore, other two objects X_{33} and X_{42}, which were misclassified by the hard k-modes algorithm, were correctly classified by the fuzzy k-modes algorithm. However, object X_1, which was correctly classified by the hard k-modes algorithm was misclassified by the fuzzy one. Because the dissimilarities of the objects X_1 and X_2 to the centers of clusters 1 and 2 are equal, the algorithm arbitrarily clustered them into cluster 1.

From this example we can see that the objects misclassified by the fuzzy k-modes algorithm were boundary objects. But it was not often the case for the hard k-modes algorithm. Another advantage of the fuzzy k-modes algorithm is that it not only partitions objects into clusters but also shows how confident an object is assigned to a cluster. The confidence is determined by the dissimilarity measures of an object to all cluster modes. For instance, although both objects X_{33} and X_{39} were assigned to cluster 2, we are more confident for X_{39}'s assignment because the confidence value $w_{39,2} = 0.9963$ is greater than the confidence value $w_{33,2} = 0.7717$ for cluster 2. In many cases, the dissimilarities of objects to the mode of the assigned cluster may be same but the confidence values of objects assigned to that cluster can be quite different because some objects may also be closer to other cluster modes but other objects are only closer to one of them. The former objects will have less confidence and can also be considered as boundary objects. In many applications, it is reasonable to consider cluster boundaries as zonal areas. The hard k-modes algorithm provides no information for identifying these boundary objects.

B. Efficiency

The purpose of the second experiment was to test the efficiency of the fuzzy k-modes algorithm when clustering large categorical data sets. For the hard k-modes algorithm Huang [7] has reported some preliminary results in clustering a large real data set consisting of 500 000 records, each being described by 34 categorical attributes. These results have shown a good scalability of the k-modes algorithm against the number of clusters for a given number of records and against the number of records for a given number of clusters. The CPU time required for clustering increased linearly as both the number of clusters and the number of records increased.

In this experiment we used an artificial data set to test the efficiency of the fuzzy k-modes algorithm. The data set had two clusters with 5000 objects each. The objects were described by five categorical attributes and each attribute had five categories. This means the maximum dissimilarity between any two objects was five. We purposely divided objects in each inherent cluster into three groups by: 1) $d_c \leq 1$; 2) $d_c = 2$; and 3) $d_c = 3$, where d_c was the dissimilarity measure between the modes of the clusters and objects. Then we specified the distribution of objects in each group as: 1) 3000; 2) 1500; and 3) 500, respectively. In creating this data set, we randomly generated two categorical objects Z_0 and Z_1 with $d_c(Z_0, Z_1) = 5$ as the inherent modes of two clusters. Each attribute value was generated by rounding toward the nearest integer of a uniform distribution between one and six. Then we randomly generated an object X with $d_c(X, Z_i)$ less than or equal to one, two, and three and added this object to the data set. Since the dissimilarity between the two clusters was five, the maximum dissimilarity between each object and the mode was at most three. Nine thousand objects had dissimilarity measure at most two to the mode of the cluster. The generated data set had two inherent clusters. Although we used integers to represent the categories of categorical attributes, the integers had no order.

Table V gives the average CPU time used by the fuzzy k-modes algorithm and the conceptual version of the k-means algorithm on a POWER2 RISC processor of IBM SP2. From Table V, we can see that the clustering accuracy of the fuzzy k-modes algorithm was better than that of the conceptual version of the k-means algorithm. Moreover, the CPU time used by the fuzzy k-modes algorithm was five times less than that used by the conceptual version of the k-means algorithm. In this test, as for the comparison, we randomly selected 1000 objects from this large data set and tested this subset with a hierarchical clustering algorithm. We found that the clustering accuracies

TABLE V
AVERAGE CLUSTERING ACCURACY AND AVERAGE CPU TIME REQUIRED IN SECONDS FOR DIFFERENT CLUSTERING METHODS ON 10 000 OBJECTS

	Conceptual k-means	Fuzzy k-modes
Accuracy	0.949	0.992
CPU time in seconds	6.56	1.28

of the hierarchical algorithm was almost the same as that of the fuzzy k-modes algorithm, but the time used by the hierarchical clustering algorithm (4.33 s) was significantly larger than that used by the fuzzy k-modes algorithm (0.384 s). Thus, when the number of objects is large, the hierarchical clustering algorithm will suffer from both storage and efficiency problem. This demonstrates the advantages of the k-modes-type algorithms in clustering large categorical data sets.

VI. CONCLUSIONS

Categorical data are ubiquitous in real-world databases. However, few efficient algorithms are available for clustering massive categorical data. The development of the k-modes-type algorithm was motivated to solve this problem. We have introduced the fuzzy k-modes algorithm for clustering categorical objects based on extensions to the fuzzy k-means algorithm. The most important result of this work is the consequence of Theorem 4 that allows the k-means paradigm to be used in generating the fuzzy partition matrix from categorical data. This procedure removes the numeric-only limitation of the fuzzy k-means algorithm. The other important result is the proof of convergence that demonstrates a nice property of the fuzzy k-modes algorithm.

The experimental results have shown that the k-modes-type algorithms are effective in recovering the inherent clustering structures from categorical data if such structures exist. Moreover, the fuzzy partition matrix provides more information to help the user to determine the final clustering and to identify the boundary objects. Such information is extremely useful in applications such as data mining in which the uncertain boundary objects are sometimes more interesting than objects which can be clustered with certainty.

REFERENCES

[1] M. R. Anderberg, *Cluster Analysis for Applications*. New York: Academic, 1973.
[2] G. H. Ball and D. J. Hall, "A clustering technique for summarizing multivariate data," *Behavioral Sci.*, vol. 12, pp. 153–155, 1967.
[3] J. C. Bezdek, "A convergence theorem for the fuzzy ISODATA clustering algorithms," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-2, pp. 1–8, Jan. 1980.
[4] R. J. Hathaway and J. C. Bezdek, "Local convergence of the fuzzy c-means algorithms," *Pattern Recognition*, vol. 19, no. 6, pp. 477–480, 1986.
[5] K. C. Gowda and E. Diday, "Symbolic clustering using a new dissimilarity measure," *Pattern Recognition*, vol. 24, no. 6, pp. 567–578, 1991.
[6] J. C. Gower, "A general coefficient of similarity and some of its properties," *BioMetrics*, vol. 27, pp. 857–874, 1971.
[7] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining Knowledge Discovery*, vol. 2, no. 3, Sept. 1998.
[8] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[9] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data—An Introduction to Cluster Analysis*. New York: Wiley, 1990.
[10] T. Kohonen, *Content-Addressable Memories*. Berlin, Germany: Springer-Verlag, 1980.
[11] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Symp. Mathematical Statistics and Probability*, Berkeley, CA, 1967, vol. 1, no. AD 669871, pp. 281–297.
[12] R. S. Michalski and R. E. Stepp, "Automated construction of classifications: Conceptual clustering versus numerical taxonomy," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-5, pp. 396–410, July 1983.
[13] H. Ralambondrainy, "A conceptual version of the k-means algorithm," *Pattern Recognition Lett.*, vol. 16, pp. 1147–1157, 1995.
[14] R. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining," in *Proc. 20th Very Large Databases Conf.*, Santiago, Chile, Sept. 1994, pp. 144–155.
[15] E. R. Ruspini, "A new approach to clustering," *Inform. Contr.*, vol. 19, pp. 22–32, 1969.
[16] S. Z. Selim and M. A. Ismail, "K-means-type algorithms: A generalized convergence theorem and characterization of local optimality," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 6, pp. 81–87, Jan. 1984.
[17] M. A. Ismail and S. Z. Selim, "Fuzzy c-means: Optimality of solutions and effective termination of the problem," *Pattern Recognition*, vol. 19, no. 6, pp. 481–485, 1986.
[18] M. A. Woodbury and J. A. Clive, "Clinical pure types as a fuzzy partition," *J. Cybern.*, vol. 4-3, pp. 111–121, 1974.