Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Pattern Recognition Letters 25 (2004) 1263–1271

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Fuzzy clustering of categorical data using fuzzy centroids

Dae-Won Kim [a,*], Kwang H. Lee [a,b], Doheon Lee [b]

[a] *Department of Electrical Engineering & Computer Science, KAIST, 373-1, Kusung-dong, Yusung-gu, Daejon 305701, South Korea*

[b] *Department of BioSystems and Advanced Information Technology Research Center (AITRC), KAIST, Kusung-dong, Yusung-gu, Daejon, South Korea*

Received 15 July 2003; received in revised form 4 March 2004
Available online 14 May 2004

Abstract

In this paper the conventional fuzzy *k*-modes algorithm for clustering categorical data is extended by representing the clusters of categorical data with fuzzy centroids instead of the hard-type centroids used in the original algorithm. Use of fuzzy centroids makes it possible to fully exploit the power of fuzzy sets in representing the uncertainty in the classification of categorical data. To test the proposed approach, the proposed algorithm and two conventional algorithms (the *k*-modes and fuzzy *k*-modes algorithms) were used to cluster three categorical data sets. The proposed method was found to give markedly better clustering results.

Keywords: Fuzzy clustering; *k*-modes algorithm; Fuzzy *k*-modes algorithm; Categorical data; Fuzzy centroid

1. Introduction

Clustering has emerged as a popular technique for pattern recognition, image processing, and, most recently, data mining. Clustering algorithms are increasingly required to deal with large scale data sets containing categorical data as well as numeric data, particularly in the context of data mining. A variety of clustering algorithms have been proposed for clustering categorical data, for example the hierarchical method using Gower's similarity coefficient (Gower, 1971; Gowda and Diday, 1991; Kaufman and Rousseeuw, 1990; Michalski and Stepp, 1983; Woodbury and Clive, 1974). However, as Huang and Ng pointed out (Huang and Ng, 1999), these algorithms become prohibitively inefficient when applied to large data sets containing only categorical data.

One method that has proved particularly efficient is the *k*-means-type algorithm (Jain and Dubes, 1998; Jain et al., 1999). Huang recently developed the *k*-modes algorithm by extending the standard *k*-means algorithm with a simple matching dissimilarity measure for categorical data, and a frequency-based method to update centroids in the clustering (Huang, 1998). This extended method is not constrained by the numeral-only

[*] Corresponding author. Tel.: +82-42-869-5561; fax: +82-42-869-9000/8680.
E-mail address: dwkim@if.kaist.ac.kr (D.-W. Kim).

0167-8655/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2004.04.004

limitation of the k-means algorithm, and has been shown to give efficient clustering performance in real-world databases. Furthermore, Huang and Ng introduced the fuzzy k-modes algorithm, a generalized version of the k-modes algorithm (Huang and Ng, 1999). Fuzzy set theory and fuzzy logic are suited to deal with uncertainties, and fuzzy clustering models have proved a particularly promising solution in a variety of areas (Zadeh, 1972; Dubois and Prade, 1980; Bezdek et al., 1999). The fuzzy k-modes algorithm generates the fuzzy partition matrix from categorical data with the framework of the fuzzy k-means-type algorithm (Bezdek, 1981; Bezdek et al., 1999), and improves on the k-modes algorithm by assigning confidence degrees to data in different clusters.

In most fuzzy versions of clustering algorithms, the assigned memberships of data to a cluster are fuzzy, but the centroid itself is not fuzzy. However, the use of hard centroids can give rise to artifacts. For example, although the fuzzy k-modes algorithm efficiently handles categorical data sets, it uses a hard centroid representation for categorical data in a cluster. This use of hard rejection of data can lead to misclassification in the region of doubt.

To address the problems caused by using hard centroids, in the present study we developed a fuzzy clustering algorithm with fuzzy centroids for clustering categorical data. The use of fuzzy centroids allows the user to fully exploit the power of fuzzy sets in representing uncertainty and imprecision. The proposed approach preserves the uncertainty inherent in data sets for longer before decisions are made, and is therefore less prone to falling into local optima in comparison to other clustering algorithms.

2. Fuzzy k-modes algorithm

Let $X = \{X_1, X_2, \ldots, X_n\}$ be a set of n categorical data. Let data X_j $(1 \leqslant j \leqslant n)$ be defined by a set of categorical attributes $A_1, A_2, \ldots, A_p$. Each A_l describes a domain of values denoted by $\mathrm{DOM}(A_l) = \{a_l^{(1)}, a_l^{(2)}, \ldots, a_l^{(n_l)}\}$ where n_l is the number of category values of attribute A_l for $1 \leqslant l \leqslant p$. A domain $\mathrm{DOM}(A_l)$ is defined as categorical if it is finite and unordered. Let X_j be denoted by $[x_{j,1}, x_{j,2}, \ldots, x_{j,p}]$. Thus X_j can be logically represented as a conjunction of attribute-value pairs $[A_1 = x_{j,1}] \wedge [A_2 = x_{j,2}] \wedge \cdots \wedge [A_p = x_{j,p}]$, where $x_{j,l} \in \mathrm{DOM}(A_l)$ for $1 \leqslant l \leqslant p$.

The objective of the fuzzy k-modes algorithm is to cluster the data X into k clusters by minimizing the function (Huang and Ng, 1999)

$$J_m(U, V : X) = \sum_{i=1}^{k} \sum_{j=1}^{n} (\mu_{ij})^m d_c(V_i, X_j) \tag{1}$$

$$\text{subject to} \quad 0 \leqslant \mu_{ij} \leqslant 1, \quad 1 \leqslant i \leqslant k, \quad 1 \leqslant j \leqslant n \tag{2}$$

$$\sum_{i=1}^{k} \mu_{ij} = 1, \quad 1 \leqslant j \leqslant n \tag{3}$$

$$0 < \sum_{j=1}^{n} \mu_{ij} < n, \quad 1 \leqslant i \leqslant k \tag{4}$$

where μ_{ij} is the membership degree of data X_j to the ith cluster, and is additionally an element of a $(k \times n)$ pattern matrix $U = [\mu_{ij}]$. $V = (V_1, V_2, \ldots, V_k)$ consists of the centroids of the fuzzy clusters. Centroid V_i is represented as $[v_{i,1}, v_{i,2}, \ldots, v_{i,p}]$. The parameter m controls the fuzziness of membership of each datum.

To cluster categorical data, the fuzzy k-modes algorithm extends the hard k-modes algorithm based on the fuzzy c-means-type procedure. First, the method for measuring the distance between a cluster centroid and a datum is proposed, along with the method for updating the cluster centroid at each iteration.

The distance measure $d_c(V_i, X_j)$ between a centroid V_i and a categorical data point X_j is defined as

$$d_c(V_i, X_j) = \sum_{l=1}^{p} \delta(v_{i,l}, x_{j,l}) \tag{5}$$

where

$$\delta(v_{i,l}, x_{j,l}) = \begin{cases} 0, & v_{i,l} = x_{j,l} \\ 1, & v_{i,l} \neq x_{j,l} \end{cases} \tag{6}$$

The measure d_c satisfies a metric space on the set of categorical objects, and is also a kind of generalized Hamming distance (Kohonen, 1980).

The cluster centroids are updated as follows. When the cluster centroid $V_i = [v_{i,1}, v_{i,2}, \ldots, v_{i,p}]$ is given, each $v_{i,l} \in V$ for $1 \leqslant l \leqslant p$ is updated as

$$v_{i,l} = a_l^{(r)} \in \mathrm{DOM}(A_l) \tag{7}$$

where

$$\sum_{x_{j,l}=a_l^{(r)}} \mu_{ij}^m \geqslant \sum_{x_{j,l}=a_l^{(t)}} \mu_{ij}^m, \quad 1 \leqslant t \leqslant n_l \tag{8}$$

The category of attribute A_l of the cluster centroid V_i is given by the category value that achieves the highest value of the summation of μ_{il} (the degrees of membership to the ith cluster) over all categories.

3. Fuzzy clustering with fuzzy centroids

3.1. Intuition and approach

In the present work, we introduce the notion of fuzzy centroids into the fuzzy clustering algorithm. This approach makes it possible to exploit the power of fuzzy sets when representing the cluster centroid. In the fuzzy k-modes algorithm, the centroids of the categorical attributes are determined through hard decisions based on the membership degrees. Thus, this representation does not keep information on the current centroids for the next iteration. For instance, consider the following example.

Example 1. Let $\mathrm{DOM}(A_l) = \{\text{high}, \text{low}\}$ and let us consider three data X_1, X_2, and X_3 whose degrees of membership to the ith cluster are $\mu_{i1} = 0.70$, $\mu_{i2} = 0.80$, and $\mu_{i3} = 0.15$, respectively.

$$X_1 = [x_{1,1}, \ldots, x_{1,l-1}, \text{“high”}, x_{1,l+1}, \ldots, x_{1,p}]$$

$$X_2 = [x_{2,1}, \ldots, x_{2,l-1}, \text{“low”}, x_{2,l+1}, \ldots, x_{2,p}]$$

$$X_3 = [x_{3,1}, \ldots, x_{3,l-1}, \text{“high”}, x_{3,l+1}, \ldots, x_{3,p}]$$

Consider the lth attribute, $v_{i,l}$, of the cluster centroid $V = [v_{i,1}, \ldots, v_{i,l}, \ldots, v_{i,p}]$. By Eqs. (7) and (8), $v_{i,l}$ is assigned the value “high” or “low” depending on the calculations of $\sum_{x_{j,l}=\text{high}} \mu_{ij}^m = 0.70^m + 0.15^m$ and $\sum_{x_{j,l}=\text{low}} \mu_{ij}^m = 0.80^m$. For instance, $v_{i,l}$ is assigned “high” for $m = 1.0$, whereas $v_{i,l}$ is assigned “low” for $m = 2.0$. According to the decision, one of the two is rejected and, despite its potential, is not concerned with the computations of the membership degrees (μ_{ij}) of data in the next iteration. This can lead to the misclassifications of data, and therefore drive the algorithm to fall into a local minimum. To prevent this, we herein propose that a soft decision be made when selecting the cluster centroids for categorical attributes, thereby preserving the uncertainty for long as possible before actual decisions are made. To achieve this objective, we introduce the concept of a fuzzy centroid.

In a hard centroid, each attribute of the centroid has a single hard category value. In contrast, each attribute of a fuzzy centroid has a fuzzy category value to describe the information distributed in the cluster. For $\mathrm{DOM}(A_l) = \{a_l^{(1)}, a_l^{(2)}, \ldots, a_l^{(n_l)}\}$, the proposed fuzzy centroid, denoted by $\widetilde{V}$, is defined as

$$\widetilde{V} = [\tilde{v}_1, \ldots, \tilde{v}_l, \ldots, \tilde{v}_p] \tag{9}$$

where

$$\tilde{v}_l = a_l^{(1)}/\omega_l^{(1)} + a_l^{(2)}/\omega_l^{(2)} + \cdots + a_l^{(n_l)}/\omega_l^{(n_l)} \tag{10}$$

$$\text{subject to} \quad 0 \leqslant \omega_l^{(t)} \leqslant 1, \quad 1 \leqslant t \leqslant n_l \tag{11}$$

$$\sum_{t=1}^{n_l} \omega_l^{(t)} = 1, \quad 1 \leqslant l \leqslant p \tag{12}$$

Each attribute $\tilde{v}_l \in \widetilde{V}$ is a fuzzy category value represented as a fuzzy set $\{(a_l^{(t)}, \omega_l^{(t)})\}$, which is a convenient notation for a fuzzy set proposed by Zadeh (1972), for $1 \leqslant t \leqslant n_l$. This is determined by the category distribution of attribute A_l for data belonging to the cluster. $\omega_l^{(t)}$ indicates the confidence degree with which $a_l^{(t)}$ contributes to $\tilde{v}_l$.

3.2. Distance measure and centroid's update

Let $\widetilde{V}$ and X be a fuzzy centroid and a data point represented as $[\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_p]$ and $[x_1, x_2, \ldots, x_p]$, respectively. The distance measure between $\widetilde{V}$ and X is defined as

$$d(\widetilde{V}, X) = \sum_{l=1}^{p} \delta(\tilde{v}_l, x_l) \tag{13}$$

where

$$\delta(\tilde{v}_l, x_l) = \sum_{t=1}^{n_l} \tau(a_l^{(t)}, x_l) \tag{14}$$

and

$$\tau(a_l^{(t)}, x_l) = \begin{cases} 0, & a_l^{(t)} = x_l \\ \omega_l^{(t)}, & a_l^{(t)} \neq x_l \end{cases} \tag{15}$$

The function δ is obtained by summing the dissimilarity between $a_l^{(t)} \in \mathrm{DOM}(A_l)$ and x_l. The function τ is assigned a value of 0.0 when two values are equal; otherwise it is assigned a value of 1.0 multiplied by its confidence degree.

Let us consider the method for updating fuzzy centroids $\widetilde{V}_i = [\tilde{v}_{i,1}, \ldots, \tilde{v}_{i,l}, \ldots, \tilde{v}_{i,p}]$ when the partition matrix $U = [u_{ij}]$ is determined based on the distance measure in Eq. (13). The attribute $\tilde{v}_l$, shown in Eq. (10), is then updated by determining $\omega_l^{(t)}$ for $1 \leqslant t \leqslant n_l$ as follows.

$$\omega_l^{(t)} = \sum_{j=1}^{n} \gamma(x_{j,l}) \tag{16}$$

where

$$\gamma(x_{j,l}) = \begin{cases} \mu_{ij}^m, & a_l^{(t)} = x_{j,l} \\ 0, & a_l^{(t)} \neq x_{j,l} \end{cases} \tag{17}$$

Each $\tilde{v}_l \in \widetilde{V}$ contains the distributions of category values of $\mathrm{DOM}(A_l)$. It is easy to verify that the conditions in Eqs. (11) and (12) are satisfied. Let us consider the Example 1 again. For $m = 1.0$, the attribute $\tilde{v}_{i,l} \in \widetilde{V}_i$ is given by

$$\tilde{v}_{i,l} = \text{“high”}/0.85 + \text{“low”}/0.80 \tag{18}$$

$\tilde{v}_l$ stores the category values and their contributions to the cluster, and is updated by the $\omega_l^{(t)}$ at each iteration before an actual decision is required.

3.3. The proposed clustering algorithm for categorical data

To minimize the objective function with fuzzy centroids,

$$J_m(U, \widetilde{V} : X) = \sum_{i=1}^{k} \sum_{j=1}^{n} (\mu_{ij})^m \mathrm{d}(\widetilde{V}_i, X_j) \tag{19}$$

the proposed algorithm uses the fuzzy c-means-type paradigm to cluster categorical data.

Step 1. Given the number of clusters, k, and a chosen value of m, choose initial centroids $\widetilde{V}(0)(t = 0)$. Each $\tilde{v}_{i,l} \in \widetilde{V}_i$ is assigned random membership values for $\omega_l^{(t)}$.

Step 2. Compute the ith fuzzy cluster for $i = 1, 2, \ldots, k$. For each x_j:

$$\mu_{ij}(t) = \left(\sum_{z=1}^{k} \left(\frac{\mathrm{d}(\widetilde{V}_i, X_j)}{\mathrm{d}(\widetilde{V}_z, X_j)} \right)^{\frac{1}{m-1}} \right)^{-1} \tag{20}$$

Step 3. Update the fuzzy cluster centroid $\widetilde{V}_i(t+1) = [\tilde{v}_{i,1}, \ldots, \tilde{v}_{i,l}, \ldots, \tilde{v}_{i,p}]$ for $i = 1, 2, \ldots, k$. For each $\tilde{v}_{i,l} = \{(a_l^{(t)}, \omega_l^{(t)})\}$ for $1 \leqslant l \leqslant p$

$$\omega_l^{(t)} = \sum_{j=1}^{n} \gamma(x_{j,l}), \quad 1 \leqslant t \leqslant n_l \tag{21}$$

where

$$\gamma(x_{j,l}) = \begin{cases} \mu_{ij}^m, & a_l^{(t)} = x_{j,l} \\ 0, & a_l^{(t)} \neq x_{j,l} \end{cases} \tag{22}$$

Step 4. If there is no improvement in J_m, then stop; otherwise, set $t \leftarrow t+1$ and go to Step 2.

Let us consider the time and space complexities of the proposed algorithm. The time complexity required mainly depends on the updates of the fuzzy centroids and partition matrix in an each iteration. The computational costs of updating the fuzzy centroids and partition matrix are $\mathrm{O}(kpn)$ and $\mathrm{O}(kNpn)$, respectively, where k is the number of clusters, p is the number of attributes, $N(= \max(n_l))$ is the maximum number of categories for $1 \leqslant l \leqslant p$, and n is the number of data. Therefore, the overall time complexity is $\mathrm{O}(kpn(N+1)s)$, where s is the number of iterations required for the algorithm to converge. The time complexity of the fuzzy k-modes algorithm is $\mathrm{O}(kn(p+M)s)$, where $M(= \sum_{l=1}^{p} n_l)$ is the total number of categories of all attributes (Huang and Ng, 1999). Typically, k and s are fixed in advance; thus the two algorithms have linear time complexities with respect to the size of the data and its attributes. When $n \gg k, p, s$, these are faster than the hierarchical clustering algorithms whose time complexity is generally $\mathrm{O}(n^2)$. For space complexity, it requires $\mathrm{O}(p(kN+n))$ to store the fuzzy centroids V and the set of data X, and it requires an additional $\mathrm{O}(kn)$ to store the pattern matrix U.

Thus, the overall space complexity of the algorithm is $\mathrm{O}(p(kN+n)+kn)$.

4. Experimental results

To test the effectiveness with which the proposed algorithm clusters categorical data, we applied the proposed algorithm and two conventional methods (the k-modes algorithm and the fuzzy k-modes algorithm) to real categorical data sets and compared the performances of the algorithms. The initial centroids of the k-modes and fuzzy k-modes algorithms were k distinct data randomly selected from the data set. Three data sets were used to evaluate the performance of each method, specifically, the SOYBEAN (Huang and Ng, 1999), CREDIT (Quilan and Quilan, 1992), and ZOO (Blake and Merz, 1989) data sets. The clustering results were assessed using Huang's accuracy measure (r) (Huang and Ng, 1999).

$$r=\frac{\sum_{i=1}^{k}a_i}{n} \tag{23}$$

where a_i is the number of data occurring in both the ith cluster and its corresponding true class, and n is the number of data 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 $r=1.0$.

4.1. Clustering performance

The first data set, referred to as the SOYBEAN data set (Huang and Ng, 1999), contains 47 data points on diseases in soybeans. Each data point has 35 categorical attributes and is classified as one of the following four diseases: Diaporthe Stem Canker, Charcoal Rot, Rhizoctonia Root Rot, and Phytophthora Rot. Phytophthora Rot has 17 data, and the three other diseases have 10 data each. The three clustering algorithms were used to cluster this data set into four clusters ($k=4$). Each algorithms was run 100 times. Fig. 1 shows the distributions of the number of runs with respect to the number of data correctly classified by each algorithm. It is evident from this figure that the proposed algorithm, which successfully classified the data set in 87 of the 100 runs, gives markedly better clustering performance in comparison to the k-modes and fuzzy k-modes algorithms. Table 1 lists the average accuracy of clustering achieved by each algorithm over 100 runs for varying $m\in[1.1,2.0]$. In agreement with the work of Huang and Ng (1999), the fuzzy k-modes algorithm provides the best result for a weighting exponent of $m=1.1$ ($r=0.772$), and is more accurate than the k-modes algorithm ($r=0.685$). We see that the execution time of the fuzzy k-modes algorithm (0.04 s) was almost the same as that of the proposed algorithm (0.06 s), but the proposed algorithm gave an accuracy of $r=0.972$ at $m=1.8$, making it 20% more accurate than the fuzzy k-modes algorithm.

Table 2 summarizes the clustering accuracies of each algorithm for the CREDIT and ZOO data sets. As for the SOYBEAN data set, each algorithm was run 100 times with varying $m\in[1.1,2.0]$. Distribution histograms similar to Fig. 1 are not presented for these data sets due to space considerations. The CREDIT data set contains 202 applicants data for credit-approval. Each application is described by nine attributes and classified as approved or rejected ($k=2$). Over the 100 runs, the proposed algorithm was most accurate at $m=1.8$, yielding an accuracy of $r=0.800$. The k-modes and fuzzy k-modes algorithms showed the lower classification accuracies of $r=0.658$ and $r=0.744$ (at $m=1.1$) respectively. In this case, there was 5.6% increase of accuracy by the proposed algorithm. The third data set, the ZOO set, contains 101 data, where each data represents an animal with 18 categorical attributes. Each animal data point is classified into seven classes ($k=7$) according to its type (e.g., mammal or bird). Over 100 runs, the k-modes and fuzzy k-modes algorithms gave accuracies of $r=0.602$ and $r=0.642$ ($m=1.1$) respectively. In contrast, the proposed algorithm gave the superior accuracy of $r=0.751$ at $m=1.8$. Thus, in this case, the proposed algorithm was 10.9% more accurate than the fuzzy k-modes algorithm.

The fourth data set, the ADULT set, contains 32,561 data, where each data represents a personal income with eight categorical attributes ($k=2$) (Blake and Merz, 1989). This set was employed to

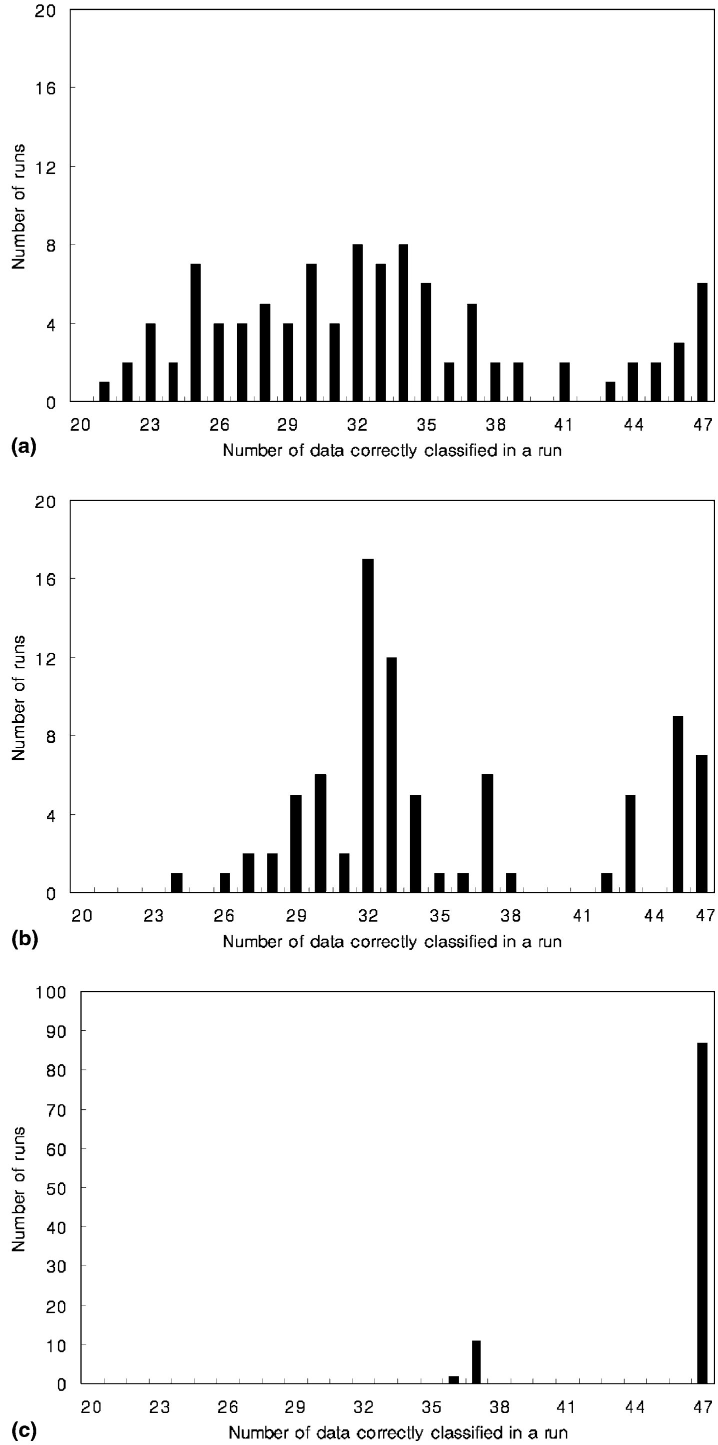

Fig. 1. Distributions of the number of runs with respect to the number of correctly classified records in each run. (a) The *k*-modes algorithm, (b) the fuzzy *k*-modes algorithm ($m = 1.1$), (c) the proposed algorithm ($m = 1.8$).

Table 1
Average clustering accuracy (r) achieved by three clustering methods for the SOYBEAN data set

m	k-modes	Fuzzy k-modes	Proposed
1.1		0.772	0.893
1.2		0.766	0.920
1.3		0.733	0.946
1.4		0.740	0.967
1.5		0.713	0.972
1.6	0.685	0.693	0.955
1.7		0.694	0.964
1.8		0.703	0.972
1.9		0.703	0.958
2.0		0.690	0.900

Table 2
Average clustering accuracy (r) achieved by three clustering methods for the CREDIT and ZOO data sets

Data set	k-modes	Fuzzy k-modes	Proposed
CREDIT	0.658	0.744	0.800
ZOO	0.602	0.642	0.751

see the differences in the computation time between the fuzzy k-modes and the proposed algorithm for clustering large data set. Both algorithms showed almost the same accuracy 75%. We see that the execution time of the proposed algorithm (4.57 s) was faster than that of the fuzzy k-modes algorithm (7.26 s). This is due to the fewer iterations of the proposed algorithm to converge than the fuzzy k-modes algorithm.

To test the scalability of the proposed algorithm for clustering very large scale data sets, we applied the proposed algorithm to the COVERTYPE data set (Blake and Merz, 1989). This set contains 300,000 data where each data point is composed of 44 categorical attributes. The scalability was tested by increasing the number of data for a fixed number of clusters ($k = 5$). Fig. 2 shows the execution times of clustering the number of data increased into five clusters. We see that the proposed algorithm shows a linear increase in execution time as the number of data increases.

From the test calculations, the k-modes and fuzzy k-modes algorithms showed similar performances for all data sets and, in comparison to these two algorithms, the proposed algorithm gave markedly better clustering performance. Thus, the test results highlight the effectiveness and potential of the proposed method for clustering categorical data.

4.2. Classification of boundary data

One of the most difficult problem in clustering is the classification of boundary data, that is, data located in the outer block of each cluster. Such data are more likely to be either misclassified to an incorrect cluster or assigned the same distance values to two neighboring clusters (Huang and Ng, 1999). To investigate the clustering of boundary data by the three algorithms considered above, we

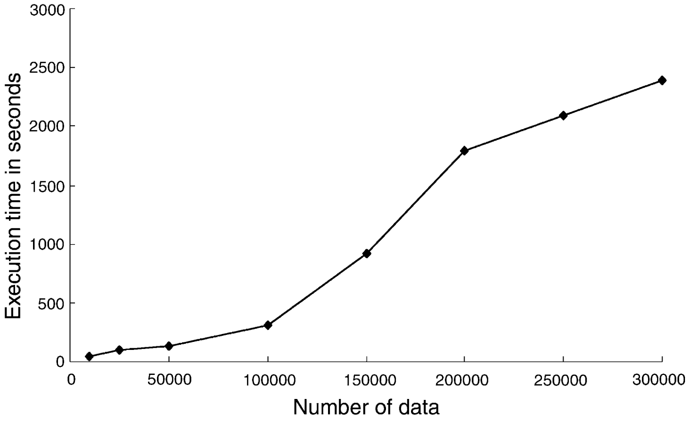

Fig. 2. Scalability to the number of data increased.

Table 3
The distance measure between misclassified data and cluster centroids, and the class no. of the assigned cluster and true cluster

Methods	Data (X_j)	Distance $d(V_i,X_j)$				Cluster no. assigned	True class
		V_1	V_2	V_3	V_4		
k-modes	$X_3(*)$	6	15	4	12	3	1
	$X_{23}(*)$	10	16	12	7	4	3
	$X_{25}(*)$	11	16	9	7	4	3
	$X_{29}(*)$	13	15	12	9	4	3
Fuzzy k-modes	X_3	6	15	6	12	1	1
	$X_{23}(*)$	10	16	11	10	1	3
	X_{25}	11	16	9	9	3	3
	$X_{29}(*)$	10	17	11	7	4	3
Proposed	X_3	6.86	12.94	11.43	11.43	1	1
	X_{23}	10.70	15.27	8.32	8.36	3	3
	X_{25}	9.99	14.34	7.64	7.67	3	3
	X_{29}	11.31	15.25	10.13	10.17	3	3

Here misclassified data are denoted by (∗).

examined four boundary data obtained from the clustering results of the SOYBEAN data set. The three algorithms were run with the same initial centroids. Table 3 shows the clustering results for the four data by the three algorithms. In this table, the distances between the data and centroids are listed, along with the differences between the cluster assigned by each algorithm and the true class. Misclassified data are denoted by (∗).

The k-modes algorithm misclassified all four boundary data, X_3, X_{23}, X_{25}, and X_{29}. Two of these misclassifications were corrected by the fuzzy k-modes algorithm. However, as pointed out by Huang and Ng (1999), the fuzzy k-modes algorithm may correctly classify boundary data simply by chance. This is observed in the present case, where X_3 was correctly classified only because it was assigned the same distances from the first and the third clusters, and was arbitrarily assigned to the first cluster (Huang and Ng, 1999). In addition, X_{25} was correctly assigned only after a similar arbitrary process. Furthermore, X_{23} and X_{29} were completely misclassified. In contrast, the proposed algorithm correctly classified all four data. From this test, we see that the boundary data, which conventional algorithms tend to misclassify or assign the same distances from two or more clusters, were correctly classified by the proposed algorithm. The success of the proposed algorithm stems from the fact that the distance measure between data and fuzzy centroids is more precise and therefore effective than that of the fuzzy k-modes algorithm.

5. Conclusions

The conventional fuzzy k-modes algorithm is capable of efficiently clustering categorical data; however, its use of hard centroids for categorical attributes and a simple distance measure compromise its precision and its ability to correctly classify boundary data. To address these shortcomings of the fuzzy k-modes algorithm, we developed a new fuzzy clustering algorithm that uses fuzzy centroids for clustering categorical data. Fuzzy centroids are a set of fuzzy values that contain category values and their confidence degrees for each attribute. In the formulation of the extended algorithm, the distance measure between data and fuzzy centroids was defined and the method for updating fuzzy centroids was presented. The proposed algorithm fully exploits the power of fuzzy sets in classifying vague data in a region of doubt such as boundary data. The superiority of the proposed algorithm over the fuzzy k-modes algorithm was clearly demonstrated through several experiments.

Besides the issues mentioned in the present work, we should tackle additional two issues in practice; (1) choosing the number of clusters and

(2) defining the categorical attributes. In this paper we assumed that the number of clusters is fixed and the categorical attributes are pre-defined by experts. Defining categorical attribute is an difficult and important task in cluster analysis. The attributes might be nominal, ordinal, or an interval scale. Different definitions of the attributes might lead to undesirable clustering results. As a future work, we would like to study this issue further.

Acknowledgements

This work was supported by the Korean Systems Biology Research Grant (M10309020000-03B5002-00000) from the Ministry of Science and Technology. We would like to thank CHUNG Moon Soul Center for Bioinformation and Bioelectronics and the IBM SUR program for providing research and computing facilities.

References

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York.

Bezdek, J.C. et al., 1999. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer Academy Publishers, Boston.

Blake, C.L., Merz, C.J., 1989. UCI Repository of machine learning databases. Available from <http://www.ics.uci.edu/mlearn/MLRepository.html>.

Dubois, D., Prade, H., 1980. Fuzzy Sets and Systems: Theory and Applications. Academic Press.

Gowda, K.C., Diday, E., 1991. Symbolic clustering using a new dissimilarity measure. Pattern Recognition 24 (6), 567–578.

Gower, J.C., 1971. A general coefficient of similarity and some of its properties. BioMetrics 27, 857–874.

Huang, Z., 1998. Extensions to the *k*-modes algorithm for clustering large data sets with categorical values. Data Min. Knowl. Disc. 2 (3).

Huang, Z., Ng, M.K., 1999. A fuzzy *k*-modes algorithm for clustering categorical data. IEEE Trans. Fuzzy Systems 7 (4).

Jain, A.K., Dubes, R.C., 1998. Algorithms for Clustering. Prentice-Hall, New Jersey.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: A review. ACM Comput. Surv. 31 (3), 264–323.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data—An Introduction to Cluster Analysis. Wiely Publishers, New York.

Kohonen, T., 1980. Content-Addressable Memories. Springer-Verlag, Berlin.

Michalski, R.S., Stepp, R.E., 1983. Automated construction of classification: Conceptual clustering versus numerical taxonomy. IEEE Trans. Pattern Anal. Machine Intell. PAMI-5, 396–410.

Quilan, J.R., Quilan, R., 1992. C4.5: Programs for Machine Learning. Morgan Kaufmann.

Woodbury, M.A., Clive, J.A., 1974. Clinical pure types as a fuzzy partition. J. Cybernet. 4 (3), 111–121.

Zadeh, L.A., 1972. A fuzzy set theoretic interpretation of linguistic hedges. J. Cybernet. 2 (3), 4–34.