amcs

AN ALTERNATIVE EXTENSION OF THE *k*-MEANS ALGORITHM FOR CLUSTERING CATEGORICAL DATA

OHN MAR SAN*, VAN-NAM HUYNH**, YOSHITERU NAKAMORI**

* Mathematics and Statistics Department
Co-Operative Degree College, Sagaing, Myanmar

** Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
e-mail: huynh@jaist.ac.jp

Most of the earlier work on clustering has mainly been focused on numerical data whose inherent geometric properties can be exploited to naturally define distance functions between data points. Recently, the problem of clustering categorical data has started drawing interest. However, the computational cost makes most of the previous algorithms unacceptable for clustering very large databases. The k-means algorithm is well known for its efficiency in this respect. At the same time, working only on numerical data prohibits them from being used for clustering categorical data. The main contribution of this paper is to show how to apply the notion of "cluster centers" on a dataset of categorical objects and how to use this notion for formulating the clustering problem of categorical objects as a partitioning problem. Finally, a k-means-like algorithm for clustering categorical data is introduced. The clustering performance of the algorithm is demonstrated with two well-known data sets, namely, *soybean disease* and *nursery* databases.

Keywords: cluster analysis, categorical data, data mining

1. Introduction

During the last decade, data-mining has emerged as a rapidly growing interdisciplinary field which merges together databases, statistics, machine learning and related areas in order to extract useful knowledge from data (Han and Kamber, 2001). Clustering is one of fundamental operations in data mining.

Clustering can be defined as the process of organizing objects in a database into clusters/groups such that objects within the same cluster have a high degree of similarity, while objects belonging to different clusters have a high degree of dissimilarity (Anderberg, 1973; Jain and Dubes, 1988; Kaufman and Rousseeuw, 1990). Traditionally, numerical clustering methods have been viewed in opposition to conceptual clustering methods developed in Artificial Intelligence. Numerical techniques emphasize the determination of homogeneous clusters according to some similarity measures but providing low-level descriptions of clusters (Anderberg, 1973; Kaufman and Rousseeuw, 1990), while a conceptual approach is more concerned with high-level (i.e. more understandable) descriptions of classes (Fisher, 1987; Michalski and Stepp, 1983). Most of the earlier work on clustering has been focused on numerical data whose inherent geometric properties can be exploited to naturally define distance functions between data points. However, data mining applications frequently involve many datasets that also consist of categorical attributes on which distance functions are not naturally defined (Ganti *et al.*, 1999). Recently, clustering data with categorical attributes have drawn some attention (Ganti *et al.*, 1999; Gibson *et al.*, 1998; Guha *et al.*, 2000; Huang, 1998).

As is well known, k-means clustering (MacQueen, 1967) has been a very popular technique for partitioning large data sets with numerical attributes. Ralambondrainy (1995) proposed a hybrid numeric-symbolic method that integrates an extended version of the k-means algorithm for cluster determination and a complementary conceptual characterization algorithm for cluster description. However, converting categorical attributes into binary attributes in Ralambondrainy's approach makes the proposed technique face the increasing of both computational and space costs if categorical attributes have many categories. Further, real values between 0 and 1 representing the cluster means do not indicate the characteristics of the clusters. Recently, Huang (1997; 1998) proposed the k-modes algorithm to tackle the problem of clustering large categorical data sets in data mining. 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. Further, Huang also combined the k-modes algorithm with the k-means algorithm resulting in the so-called k-prototypes algorithm for clustering objects described by mixed numerical and categorical attributes. These extensions have removed the numeric-only limitation of the k-means algorithm and enabled it to be used for efficient clustering of very large data sets from real-world databases. However, the k-modes algorithm is unstable due to non-uniqueness of the modes. That is, the clustering results depend strongly on the selection of modes during the clustering process.

This paper aims at eliminating the above drawback in the k-modes algorithm by introducting a new notion of “cluster centers” called representatives for categorical objects. As arithmetic operations are completely absent in the setting of categorical objects, we apply the notion of fuzziness in defining representatives instead of means for clusters. With this notion we can also formulate the clustering problem of categorical objects as a partitioning problem in the fashion similar to k-means clustering. The remainder of this paper is organized as follows: In the next section, we introduce the definition and notation for categorical objects. In Section 3, we briefly outline the k-means-type clustering for numerical and categorical data. Section 4 formulates the clustering problem of categorical objects as a partitioning problem in the fashion similar to k-means clustering, and describes the proposed algorithm. Experimental results with two well-known data sets, namely, *soybean disease* and *nursery* databases are reported in Section 5. Finally, Section 6 presents some concluding remarks and suggestions for further work.

2. Notation

Similar to (Huang, 1998), we assume that the set of objects to be clustered is stored in a dataset $\mathbf{D}$ defined by a set of attributes $A_1, \ldots, A_m$ with domains $\mathcal{D}_1, \ldots, \mathcal{D}_m$, respectively. Each object in $\mathbf{D}$ is represented by a tuple $t \in \mathcal{D}_1 \times \ldots \times \mathcal{D}_m$. In terms of the clustering problem discussed in this paper, we only consider two general data types, namely, *numeric* and *categorical*. The domains of attributes associated with these two types are called numerical and categorical, respectively. A numerical domain consists of continuous real values. As such, each numerical data object is considered as a point in a multidimentional metric space adopting a distance metric such as the Euclidean or the Mahalanobis measure (Jain and Dubes, 1988). Following the lines of (Huang, 1998), a domain $\mathcal{D}_i$ is defined as categorical if it is finite and unordered, e.g., that only a comparison operation is allowed in $\mathcal{D}_i$. That is, for any $a, b \in \mathcal{D}_i$ either $a = b$ or $a \neq b$. Symbolic data objects as considered in (Gowda and Diday, 1991) are not discussed in the present paper.

Logically, each data object X in the dataset is also represented as a conjunction of attribute-value pairs

$$[A_1 = x_1] \wedge \ldots \wedge [A_m = x_m],$$

where $x_i \in \mathcal{D}_i$ for $1 \leq i \leq m$. For simplicity, we represent X as a tuple

$$(x_1, \ldots, x_m) \in \mathcal{D}_1 \times \cdots \times \mathcal{D}_m.$$

If all $\mathcal{D}_i$'s are categorical domains, then objects in $\mathbf{D}$ are called categorical objects. Huang (1997; 1998) also considered the clustering problem for mixed-type data objects where some domains are numeric, while others are categorical.

3. *k*-Means Clustering

The general algorithm was introduced by Cox (1957), and (Ball and Hall, 1967; MacQueen, 1967) first named it k-means. Since then it has become widely popular and is classified as a *partitional* or *non-hierarchical* clustering method (Jain and Dubes, 1988). It is defined as follows: given a set $\mathbf{D} = \{X_1, \ldots, X_n\}$ of n numerical data objects, a natural number $k \leq n$, and a distance measure d, the k-means algorithm aims at finding a parition $\mathcal{C}$ of $\mathbf{D}$ into k non-empty disjoint clusters $C_1, \ldots, C_k$ with $C_i \cap C_j = \emptyset$ and $\bigcup_{i=1}^{k} C_i = \mathbf{D}$ such that the overall sum of the squared distances between data objects and their cluster centers is minimized. Mathematically, if we use indicator variables $w_{i,l}$ which take value 1 if object X_i is in cluster C_l, and 0 otherwise, then the problem can be stated in terms of a constrained non-linear optimization problem as follows: Minimize

$$P(W, \mathcal{Q}) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{i,l} d(X_i, Q_l) \tag{1}$$

subject to

$$\sum_{l=1}^{k} w_{i,l} = 1, \quad 1 \leq i \leq n, \tag{2}$$

$$w_{i,l} \in \{0, 1\}, \quad 1 \leq i \leq n, 1 \leq l \leq k,$$

where $W = [w_{i,l}]_{n \times k}$ is a partition matrix, $\mathcal{Q} = \{Q_1, \ldots, Q_k\}$ is a set of cluster centers, and $d(\cdot, \cdot)$ is the squared Euclidean distance between two objects.

As is well known, the usual method toward the optimization of P in (1) subject to the constraint (2) is to use partial optimization for $\mathcal{Q}$ and W. That is, we first fix $\mathcal{Q}$ and find necessary conditions for W to minimize P. Then we fix W and minimize P according to $\mathcal{Q}$. Basically, the k-means algorithm iterates through a three-step process until $P(W, \mathcal{Q})$ converges to some local minimum (Selim and Ismail, 1984):

1. Select an initial $\mathcal{Q}^{(0)} = \{Q_1^{(0)}, \ldots, Q_k^{(0)}\}$, and set $t = 0$.
2. Keep $\mathcal{Q}^{(t)}$ fixed and solve $P(W, \mathcal{Q}^{(t)})$ to obtain $W^{(t)}$, i.e., regarding $\mathcal{Q}^{(t)}$ as the cluster centers, assign each object to the cluster of its nearest cluster center.
3. Keep $W^{(t)}$ fixed and generate $\mathcal{Q}^{(t+1)}$ such that $P(W^{(t)}, \mathcal{Q}^{(t+1)})$ is minimized, i.e., construct new cluster centers according to the current distribution of objects.
4. In the case of convergence or if a given stopping criterion is fulfilled, output the result and stop. Otherwise, set $t = t + 1$ and go to Step 2.

In the setting of numerical data clustering, the Euclidean norm

$$d(X, Y) = \sqrt{\sum_{j=1}^{m} |x_j - y_j|^2} \tag{3}$$

is often chosen as a natural distance measure in the k-means algorithm. With this distance measure, the computation of the mean of a cluster's objects returns the cluster's center, fulfilling the minimization condition of Step 3 above. Namely, $Q_l^{(t+1)} = (q_{l,1}^{(t+1)}, \ldots, q_{l,m}^{(t+1)})$ for $l = 1, \ldots, k$, and

$$q_{l,j}^{(t+1)} = \frac{\sum_{i=1}^{n} w_{i,l}^{(t)} x_{i,j}}{\sum_{i=1}^{n} w_{i,l}^{(t)}}. \tag{4}$$

To deal with the problem of not well-defined boundaries between clusters, the notion of fuzzy partitions has been applied successfully to the clustering problem resulting in the so-called *fuzzy clustering* (Ruspini, 1969; Bezdek, 1980; Ismail and Selim, 1986). However, we do not consider this topic in the present paper.

As is shown in (Huang, 1998), the k-means algorithm has the following characteristics:

- It is efficient in processing large data sets.
- It often terminates at a local optimum.
- It works only on numerical data.
- The clusters have convex shapes.

It was also shown in (Huang, 1997; Huang, 1998) that the k-means method can be extended to categorical data by using a simple matching distance measure for categorical objects with a majority-vote strategy to define the "cluster centers" called *modes*. Specifically, the distance between two categorical objects $X_1, X_2 \in \mathbf{D}$, with $X_1 = (x_{1,1}, \ldots, x_{1,m})$ and $X_2 = (x_{2,1}, \ldots, x_{2,m})$ consisting of categorical values only, can be defined as follows:

$$d(X_1, X_2) = \sum_{j=1}^{m} \delta(x_{1,j}, x_{2,j}), \tag{5}$$

where

$$\delta(x_{1,j}, x_{2,j}) = \begin{cases} 0 & \text{if } x_{1,j} = x_{2,j}, \\ 1 & \text{if } x_{1,j} \neq x_{2,j}. \end{cases}$$

Given a cluster $\{X_1, \ldots, X_p\}$ of categorical objects, with $X_i = (x_{i,1}, \ldots, x_{i,m})$, $1 \leq i \leq p$, its mode $Q = (q_1, \ldots, q_m)$ is defined by assigning q_j, $1 \leq j \leq m$, the category most frequently encountered in $\{x_{1,j}, \ldots, x_{p,j}\}$.

With these modifications, Huang (1998) proposed a k-modes algorithm that mimics the k-means method to cluster categorical data. However, it should be also emphasized that, by definition, the mode of a cluster is not generally unique. This makes the algorithm unstable depending on mode selection during the clustering process. Huang (1998), presented two mode selection methods and tested the k-modes algorithm with these methods (the result will be discussed in Section 5).

4. Proposed Algorithm

In this section we discuss how to avoid the drawback of the k-modes algorithm, and propose a new alternative algorithm that also mimics the k-means method in clustering categorical data.

As we have seen, in applying the k-means method to categorical objects, two main problems are encountered, namely, the formation of cluster centers and the calculation of dissimilarity between objects and cluster centers. These problems have been completely solved in the k-modes algorithm by using the simple matching dissimilarity measure for categorical data instead of the Euclidean distance measure, and replacing the means of clusters by the modes. These modifications also meet the minimization condition, as was shown in (Huang, 1998). In the following, we address these two problems before introducing the proposed algorithm.

4.1. Formation of "Cluster Centers"

As arithmetic operations are completely absent in the setting of categorical objects, we use the Cartesian product and union operations for the formation of "cluster centers" based on the notion of means in the numerical setting. Particularly, we replace addition and multiplication in (4) by the union and the Cartesian product, respectively, for categorical data in defining the notion of *representatives* for clusters.

Given a cluster $C = \{X_1, \ldots, X_p\}$ of categorical objects, with

$$X_i = (x_{i,1}, \ldots, x_{i,m}), \quad 1 \le i \le p,$$

denote by D_j the set formed from categorical values $x_{1,j}, \ldots, x_{p,j}$. For example, the set formed from values a, b, a, c is $\{a, b, c\}$.

Then the representative of C is defined by $Q = (q_1, \ldots, q_m)$, with

$$q_j = \{(c_j, f_{c_j}) \,|\, c_j \in D_j\}, \tag{6}$$

where f_{c_j} is the relative frequency of category c_j within C, i.e., $f_{c_j} = n_{c_j}/p$, where n_{c_j} is the number of objects in C having category c_j at attribute A_j. Formally, each q_j can be seen as a fuzzy set on D_j with membership grades of elements to be defined by their relative frequencies within the cluster.

It would be worthwhile to note that if we apply the above definition of representatives (cf. (6)) for numerical data objects with replacing the union and Cartesian product by addition and multiplication, respectively, we will obtain the notion of cluster centers by means (cf. (4)).

4.2. Dissimilarity Measure

Due to the modification proposed in forming representatives for clusters of categorical objects, the dissimilarity between a categorical object and the representative of a cluster is defined based on simple matching as follows.

Let $C = \{X_1, \ldots, X_p\}$ be a cluster of categorical objects, with

$$X_i = (x_{i,1}, \ldots, x_{i,m}), \quad 1 \le i \le p,$$

and $X = (x_1, \ldots, x_m)$ be a categorical object. Note that X may or may not belong to C. Assume that $Q = (q_1, \ldots, q_m)$, with

$$q_j = \{(c_j, f_{c_j}) \,|\, c_j \in D_j\},$$

is a representative of cluster C. Now we define the dissimilarity between object X and representative Q by

$$d(X, Q) = \sum_{j=1}^{m} \sum_{c_j \in D_j} f_{c_j} \cdot \delta(x_j, c_j). \tag{7}$$

Under such a definition, the dissimilarity $d(X, Q)$ is mainly dependent on the relative frequencies of categorical values within the cluster and simple matching between categorical values. It is also of interest to note that the simple matching dissimilarity measure between categorical objects can be considered as a categorical counterpart of the squared Euclidean distance measure.

It is easily seen that

$$\begin{aligned} d(X, Q) &= \sum_{j=1}^{m} \sum_{c_j \in D_j} f_{c_j} \delta(x_j, c_j) \\ &= \sum_{j=1}^{m} \sum_{c_j \in D_j, c_j \ne x_j} f_{c_j} \\ &= \sum_{j=1}^{m} (1 - f_{x_j}), \end{aligned} \tag{8}$$

where f_{x_j} is the relative frequency of category x_j within C.

4.3. *k*-Representatives Algorithm

With the modifications just made above, we are now ready to formulate the problem of clustering categorical data as a partitioning problem in a fashion similar to k-means clustering.

Assume that we have a data set $\mathbf{D} = \{X_1, \ldots, X_n\}$ of categorical objects to be clustered, where each object $X_i = (x_{i,1}, \ldots, x_{i,m})$, $1 \le i \le n$ is described by m categorical attributes. Then the problem can be mathematically stated as follows: Minimize

$$P(W, \mathcal{Q}) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{i,l} d(X_i, Q_l), \tag{9}$$

subject to

$$\sum_{l=1}^{k} w_{i,l} = 1, \quad 1 \le i \le n, \tag{10}$$

$$w_{i,l} \in \{0, 1\}, \quad 1 \le i \le n, 1 \le l \le k,$$

where $W = [w_{i,l}]_{n \times k}$ is a partition matrix, $\mathcal{Q} = \{Q_1, \ldots, Q_k\}$ is the set of representatives, and $d(X_i, Q_l)$ is the dissimilarity between object X_i and representative Q_l defined by (7).

In much the same way as in the k-mode algorithm proposed in (Huang, 1998), we introduce the following algorithm for clustering categorical data:

1. Initialize a *k*-partition of $\mathbf{D}$ randomly.
2. Calculate k representatives, one for each cluster.

3. For each X_i, calculate the dissimilarities
$$d(X_i, Q_l), \quad l = 1, \dots, k.$$
Reassign X_i to cluster C_l (from cluster $C_{l'}$, say) such that the dissimilarity between X_i and Q_l is least. Update both Q_l and $Q_{l'}$.
4. Repeat Step 3 until no object has changed clusters after a full cycle test of the whole data set.

We should also note that the definition of representatives of clusters in the proposed technique is based on the notion of means, i.e., the optimized solution of the corresponding numerical problem (cf. (4)). Thus the optimization problem (9) is reduced to the partial optimization problem, namely, P_1 (Huang, 1998), for W as specified in the above algorithm.

According to (8), we have

$$d(X_i, Q_l) = \sum_{j=1}^{m} (1 - f_{x_{i,j}}), \tag{11}$$

where $f_{x_{i,j}}$ is the relative frequency of category $x_{i,j}$ within cluster C_l. Thus, in the proposed algorithm object X_i will be allocated to cluster C_l so that the categories of X_i are most likely to constitute a mode[1] of C_l related to the other clusters. Note that, by definition, all possible modes of each cluster are taken into account in the proposed algorithm.

5. Experimental Results

In this section we present two experimental tests with the proposed algorithm on *soybean disease* and *nursery* databases taken from the UCI Repository of Machine Learning Databases and Domain Theorem (Blake and Merz, 1998).

As is well known, the primary aim of clustering algorithms is to discover classes that exist inherently in data. With this purpose in mind, we first assume that a structure may exist in a given dataset and then a clustering algorithm should be used to verify the assumption and recover the structure. In order to evaluate the k-modes algorithm, Huang (1998) adopted an *external criterion* which measures the degree of correspondence between the clusters obtained from the algorithm and the class assigned *a priori*.

In this paper we intended to develop the idea proposed by Huang for extending the *k*-means paradigm to clustering categorical data. Therefore, in the following, we also adopt the same criterion to analyze the clustering results of our algorithm.

5.1. Soybean Data Set

In much the same way as in (Huang, 1998), this data set is chosen to test our algorithm because of its public availability and since all its attributes can be treated as categorical ones.

The soybean dataset has 47 instances, each being described by 35 attributes. Also, we only selected 21 among 35 attributes in the present experiment as the others have only one category. Each instance is labeled as one of four diseases: Diaporthe Stem Canker, Charcoal Rot, Rhizoctonia Root Rot, and Phytophthora Rot. Except for Phytophthora Rot, which has 17 instances, all other diseases have 10 instances each. To study the effect of the record order, we created 1000 test data sets by randomly reordering the original records. We used the k-representatives algorithm to cluster each test data set of the soybean disease data into four clusters and produced 1000 clustering results. For each clustering result we also used a misclassification matrix to analyze the correspondence between the clusters and the disease classes of the instances. For instance, two misclassification matrices for the two test data sets are shown in Tables 1 and 2. The capital letters D, C, R, P in the first column of the matrices represent the four disease classes. In Table 1 there is a one-to-one correspondence between clusters and disease classes, which means the instances in the same disease classes were clustered into the same clusters. This represents a complete recovery of the four disease classes from the test data set, while in Table 2 two instances of the disease class P were misclassified into Cluster 1, which was dominated by the instances of the disease type R. However, the instances in the other two disease classes were correctly clustered into Clusters 3 and 4. The misclassification matrix of each

Table 1. Misclassification matrix for the first test data set.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
D			10	
C				10
R	10			
P		17		

Table 2. Misclassification matrix for the second test data set.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
D				10
C			10	
R	10			
P	2	15		

[1] in the sense of (Huang, 1998).

clustering result is also used to define the so-called *clustering accuracy* as below.

Huang (1998) proposed a measure of clustering results called the *clustering accuracy*, defined as follows:

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

where a_l is the number of data objects that occur in both cluster C_l and its corresponding labeled class, and n is the number of objects in the data set. Further, the clustering error is defined as $e = 1 - r$. For example, with the misclassification matrix in Table 2 we have $r = (10 + 15 + 10 + 10)/47 = 0.9574$ and $e = 0.0426$.

In the experiment with the k-representatives algorithm, we produced 1000 clustering results and if we also consider the clustering accuracy $r > 0.87$ as a 'good' clustering result, then 686 good results were produced. This means there is a 68.6% chance to obtain a good result by employing the k-representatives algorithm. The distribution of the clustering accuracies is shown in Table 3.

Table 3. Distribution of accuracies in the first experiment.

Clustering accuracies	Number of results	Good results
0.531 ∼ 0.595	2	
0.638 ∼ 0.659	7	
0.680 ∼ 0.702	8	
0.723 ∼ 0.744	28	
0.765 ∼ 0.787	94	
0.808 ∼ 0.829	86	
0.851 ∼ 0.872	89	
0.893 ∼ 0.914	80	good
0.936 ∼ 0.957	87	good
0.978 ∼ 1.000	519	good

5.2. Nursery Data Set

In this experiment we test the proposed algorithm with the nursery dataset donated by Marko Bohanec and Blaz Zupan, cf. Blake and Merz, 1998. The nursery dataset was derived from a hierarchical decision model originally developed to rank applications for nursery schools. It has in total 12960 instances, each being described by 8 categorical attributes, and there are no missing attribute values. The instances were classified into five classes. We also created 1000 test data sets from the nursery data set by randomly reordering the 12960 original instances. We produced 1000 clustering results with the algorithm.

With the same measure of clustering accuracy defined as above, we also obtained 691 good clustering results among 1000 tests with the accuracy of $r > 0.87$. This means that there is a 69.1% chance to obtain a good clustering result by employing the k-representatives algorithm for this data set. The distribution of clustering accuracies for this experiment is shown in Table 4.

Table 4. Distribution of accuracies in the second experiment.

Clustering accuracies	Number of results	Good results
0.459 ∼ 0.475	3	
0.506 ∼ 0.537	6	
0.614 ∼ 0.621	26	
0.675 ∼ 0.688	93	
0.764 ∼ 0.791	77	
0.857 ∼ 0.879	104	
0.891 ∼ 0.926	27	good
0.956 ∼ 0.964	97	good
0.978 ∼ 1.000	567	good

5.3. Discussion

We should mention that with the k-modes algorithm, Huang (1998) produced 200 clustering results for the soybean data set, among which each 100 for one selection method of modes. Also, considering the accuracy $r > 0.87$ as a good clustering result, Huang obtained 45 good results with the first selection method and 64 good results with the second one. This means that in applying the k-modes algorithm we have a 45% chance to obtain a good clustering result with the first selection method of modes and a 64% chance with the second one. This shows that clustering results of the k-modes algorithm strongly depend on the method of mode selection chosen.

The experimental result for the soybean data set has shown that the k-representatives algorithm is more stable than the k-modes algorithm. Because in the k-representative algorithm each object is allocated to a cluster, its categories are most likely to constitute a mode of the cluster in relation to the other clusters. That is, by definition, all possible modes of each cluster are taken into account in the proposed algorithm. We have not made a comparison with the k-modes algorithm in the case of the nursery data set as Huang did not present a test for the clustering performance with this data set in his papers.

6. Conclusions

In this paper we have made modifications in the k-means method while apply the method to the problem of clustering categorical data. Consequently, the so-called k-representatives algorithm for clustering categorical data has been proposed. The clustering performance of the proposed algorithm is demonstrated with two well-known data sets, namely, *soybean disease* and *nursery* databases.

The experimental results have shown that the proposed algorithm gives better results, and is more stable than the k-modes algorithm. Furthermore, it would be straightforward to combine the proposed algorithm with the k-means algorithm in a similar manner as is done in (Huang, 1997; Huang, 1998) for applying the k-means paradigm to clustering for mixed datasets. This problem, as well as the extention of the proposed technique to the problem of fuzzy clustering for categorical data (Huang and Ng, 1999), are the subject of our further work.

References

Anderberg M.R. (1973): *Cluster Analysis for Applications*. — New York: Academic Press.

Ball G.H. and Hall D.J. (1967): *A clustering technique for summarizing multivariate data*. — Behav. Sci., Vol. 12, No. 2, pp. 153–155.

Bezdek J.C. (1980): *A convergence theorem for the fuzzy ISODATA clustering algorithms*. — IEEE Trans. Pattern Anal. Mach. Intell., Vol. 2, No. 1, pp. 1–8.

Blake C.L. and Merz C.J. (1998): *UCI Repository of machine learning databases*. — Available at: http://www.ics.uci.edu/~mlearn/MLRepository.html, Irvine, CA: University of California, Department of Information and Computer Science.

Cox D. (1957): *Note on grouping*. — J. Amer. Stat. Assoc., Vol. 52, pp. 543–547.

Fisher D.H. (1987): *Knowledge acquisition via incremental conceptual clustering*. — Mach. Learn., Vol. 2, No. 2, pp. 139–172.

Ganti V., Gehrke J. and Ramakrishnan R. (1999): *CATUS - Clustering categorical data using summaries*. — Proc. Int. Conf. *Knowledge Discovery and Data Mining*, San Diego, USA, pp. 73–83.

Gibson D., Kleinberg J. and Raghavan P. (1998) *Clustering categorical data: An approach based on dynamic systems*. Proc. 24-th Int. Conf. *Very Large Databases*, New York, pp. 311–323.

Gowda K.C. and Diday E. (1991): *Symbolic clustering using a new dissimilarity measure*. — Pattern Recogn., Vol. 24, No. 6, pp. 567–578.

Guha S., Rastogi R. and Shim K. (2000): *ROCK: A robust clustering algorithm for categorical attributes*. — Inf. Syst., Vol. 25, No. 5, pp. 345–366.

Han J. and Kamber M. (2001): *Data Mining: Concepts and Techniques*. — San Francisco: Morgan Kaufmann Publishers.

Hathaway R.J. and Bezdek J.C. (1986): *Local convergence of the* c*-means algorithms*. — Pattern Recogn., Vol. 19, No. 6, pp. 477–480.

Huang Z. (1997): *Clustering large data sets with mixed numeric and categorical values*, In: *KDD: Techniques and Applications* (H. Lu, H. Motoda and H. Luu, Eds.). — Singapore: World Scientific, pp. 21–34.

Huang Z. (1998): *Extensions to the* k*-means algorithm for clustering large data sets with categorical values*. — Data Mining Knowl. Discov., Vol. 2, No. 2, pp. 283–304.

Huang Z. and Ng M.K. (1999): *A fuzzy* k*-modes algorithm for clustering categorical data*. — IEEE Trans. Fuzzy Syst., Vol. 7, No. 4, pp. 446–452.

Ismail M.A. and Selim S.Z. (1986): *Fuzzy* c*-means: Optimality of solutions and effective termination of the problem*. — Pattern Recogn., Vol. 19, No. 6, pp. 481–485.

Jain A.K. and Dubes R.C. (1988): *Algorithms for Clustering Data*. — Englewood Cliffs: Prentice Hall.

Kaufman L. and Rousseeuw P.J. (1990): *Finding Groups in Data*. — New York: Wiley.

MacQueen J.B. (1967): *Some methods for classification and analysis of multivariate observations*. — Proc. 5-th Symp. *Mathematical Statistics and Probability*, Berkelely, CA, Vol. 1, pp. 281–297.

Michalski R.S. and Stepp R.E. (1983): *Automated construction of classifications: Conceptual clustering versus numerical taxonomy*. — IEEE Trans. Pattern Anal. Mach. Intell., Vol. PAMI-5, No. 4, pp. 396–410.

Ralambondrainy H. (1995): *A conceptual version of the k-means algorithm*. — Pattern Recogn. Lett., Vol. 15, No. 11, pp. 1147–1157.

Ruspini E.R. (1969): *A new approach to clustering*. — Inf. Contr., Vol. 15, No. 1, pp. 22–32.

Selim S.Z. and Ismail M.A. (1984): k*-Means-type algorithms: A generalized convergence theorem and characterization of local optimality*. — IEEE Trans. Pattern Anal. Mach. Intell., Vol. PAMI-6, No. 1, pp. 81–87.

Received: 2 August 2003
Revised: 1 April 2004