Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition 38 (2005) 1131–1134
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Rapid and brief communication

A *k*-populations algorithm for clustering categorical data

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

[a]*Department of BioSystems and Advanced Information Technology Research Center, Korea Advanced Institute of Science and Technology, Guseong-dong, Yuseong-gu 305-701, Daejeon, Republic of Korea*

[b]*Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology, Guseong-dong, Yuseong-gu 305-701, Daejeon, Republic of Korea*

Received 13 October 2004; accepted 1 November 2004

Abstract

In this paper, the conventional *k*-modes-type algorithms for clustering categorical data are extended by representing the clusters of categorical data with *k*-populations instead of the hard-type centroids used in the conventional algorithms. Use of a population-based centroid representation makes it possible to preserve the uncertainty inherent in data sets as long as possible before actual decisions are made. The *k*-populations algorithm was found to give markedly better clustering results through various experiments.

Keywords: Clustering; Categorical data; Hierarchical algorithm; *k*-Modes algorithm; Fuzzy *k*-modes algorithm

1. Introduction

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 [1]. However, as Huang and Ng pointed out [2], these algorithms become prohibitively inefficient when applied to large data sets containing only categorical data. 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 [2]. This extended method 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 [3], which assigns membership degrees to data in different clusters.

Although the *k*-modes-type algorithms efficiently handle categorical data sets, they use a hard centroid for categorical attributes in a cluster. This use of hard centroids and a simple distance measure compromise its precision and its ability to classify categorical data, leading to misclassification. To address these problems, in the present study, we developed a *k*-populations algorithm for clustering categorical data. The notion of population minimizes the uncertainty and imprecision in the representation of cluster centroids. 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. The *k*-modes and fuzzy *k*-modes algorithms

Let $X=\{x_1, x_2, \ldots, x_n\}$ be a set of n categorical data. Let x_j $(1 \leqslant j \leqslant n)$ be defined by a set of categorical attributes

* Corresponding author. Tel.: +82 42 869 4353; fax: +82 42 869 8680.
E-mail address: dwkim@bisl.kaist.ac.kr (D.-W. Kim).

0031-3203/$30.00
doi:10.1016/j.patcog.2004.11.017

$A_1, A_2, \ldots, A_p$. Each A_l $(1 \leqslant l \leqslant p)$ describes a domain of values denoted by $DOM(A_l) = \{a_l^{(1)}, a_l^{(2)}, \ldots, a_l^{(n_l)}\}$ where n_l is the number of category values of attribute A_l. 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 $x_{j,l} \in DOM(A_l)$ for $1 \leqslant l \leqslant p$.

The objective of the *k*-modes-type algorithms is to cluster X into k clusters by minimizing the function

$$J_m(V : X) = \sum_{i=1}^{k} \sum_{j=1}^{n} (\mu_{i,j})^m d_c(v_i, x_j), \tag{1}$$

where $\mu_{i,j}$ indicates whether x_j belongs to the ith cluster for the *k*-modes algorithm; $\mu_{i,j} = 1$ if x_j belongs to the ith cluster and 0 otherwise, and for the fuzzy *k*-modes algorithm, $\mu_{i,j}$ is the membership degree of x_j to the ith cluster. $V = (v_1, v_2, \ldots, v_k)$ consists of the cluster centroids. Centroid v_i is represented as $[v_{i,1}, v_{i,2}, \ldots, v_{i,p}]$. The parameter m is a positive coefficient for controlling the membership of each datum.

To cluster categorical data, the *k*-modes-type algorithms measure the distance between a cluster centroid and a categorical data point, and update the cluster centroid at each iteration as follows:

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

$$d_c(v_i, x_j) = \sum_{l=1}^{p} \delta(v_{i,l}, x_{j,l}), \tag{2}$$

where $\delta(v_{i,l}, x_{j,l}) = 0$ if $v_{i,l} = x_{j,l}$ and 1 if $v_{i,l} \neq x_{j,l}$. The ith cluster centroid $v_i = [v_{i,1}, \ldots, v_{i,p}]$, referred to as the ith mode, are updated as follows. Each $v_{i,l} \in v_i$ for $1 \leqslant l \leqslant p$ is updated as

$$v_{i,l} = a_l^{(r)} \in DOM(A_l), \tag{3}$$

where $a_l^{(r)}$ satisfies the following criterion for the *k*-modes algorithm:

$$|\{\mu_{i,j} \mid x_{j,l} = a_l^{(r)}, \mu_{i,j} = 1\}| \geqslant |\{\mu_{i,j} \mid x_{j,l} = a_l^{(t)}, \mu_{i,j} = 1\}|, \quad 1 \leqslant t \leqslant n_l, \tag{4}$$

and satisfies the following criterion for the fuzzy *k*-modes algorithm:

$$\sum_{x_{j,l} = a_l^{(r)}} \mu_{i,j}^m \geqslant \sum_{x_{j,l} = a_l^{(t)}} \mu_{i,j}^m, \quad 1 \leqslant t \leqslant n_l. \tag{5}$$

For the *k*-modes algorithm, the category of attribute $v_{i,l}$ of the cluster centroid v_i is determined by the frequency mode of categories of attribute A_l in the set of data belonging to the ith cluster. For the fuzzy *k*-modes algorithm, $v_{i,l}$ is given by the category value that achieves the highest value of the summation of $\mu_{i,j}$ to the ith cluster over all categories.

3. The *k*-populations algorithm

3.1. Definition of k-population

In the fuzzy *k*-modes algorithm, the centroids of the categorical attributes are determined through hard decisions based on membership degrees. Thus, this representation does not keep information on the current centroids for the next iteration. For example, let $DOM(A_l) = \{yes, no\}$ 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. The lth attribute value of each data point is given as $x_{1,l} = yes$, $x_{2,l} = no$, and $x_{3,l} = yes$.

Consider the lth attribute, $v_{i,l}$, of the ith cluster centroid. By Eqs. (3) and (5), $v_{i,l}$ is assigned the value "*yes*" or "*no*" depending on the calculations of $\sum_{x_{j,l}=yes} \mu_{ij}^m = 0.70^m + 0.15^m$ and $\sum_{x_{j,l}=no} \mu_{ij}^m = 0.80^m$. Therefore, $v_{i,l}$ is assigned "*yes*" for $m = 1.0$, whereas $v_{i,l}$ is assigned "*no*" 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 thus drive the algorithm to fall into a local minimum. A similar problem also arises for the *k*-modes algorithm; a single attribute value $a_l^{(r)}$ with the highest frequency is not sufficient to effectively represent the distribution of attribute A_l in a cluster. 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 the actual decisions are made. To achieve this objective, we introduce the notion of a population.

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

$$v_i = [v_{i,1}, \ldots, v_{i,l}, \ldots, v_{i,p}], \tag{6}$$

where

$$v_{i,l} = \{(a_l^{(t)}, \omega_l^{(t)}) \mid a_l^{(t)} \in DOM(A_l), 1 \leqslant t \leqslant n_l\}, \tag{7}$$

subject to

$$0 \leqslant \omega_l^{(t)} \leqslant 1, \quad 0 < \sum_{t=1}^{n_l} \omega_l^{(t)} < n. \tag{8}$$

Thus, $v_{i,l}$ describes the category distribution of attribute A_l for data belonging to the ith cluster. $\omega_l^{(t)}$ indicates the confidence degree with which $a_l^{(t)}$ contributes to $v_{i,l}$.

3.2. Distance measure and centroid's update

Let v_i and x_j be the ith cluster and a data point represented as $[v_{i,1}, v_{i,2}, \ldots, v_{i,p}]$ and $[x_{j,1}, x_{j,2}, \ldots, x_{j,p}]$,

respectively. The distance measure between v_i and x_j is defined as

$$d_c(v_i, x_j) = \sum_{l=1}^{p} \delta(v_{i,l}, x_{j,l}), \tag{9}$$

where

$$\delta(v_{i,l}, x_{j,l}) = \frac{1}{\eta_i} \sum_{t=1}^{n_l} \tau(a_l^{(t)}, x_{j,l}) \tag{10}$$

and

$$\tau(a_l^{(t)}, x_{j,l}) = \begin{cases} 0, & a_l^{(t)} = x_{j,l}, \\ \omega_l^{(t)}, & a_l^{(t)} \neq x_{j,l}. \end{cases} \tag{11}$$

The function δ is obtained by summing the dissimilarity between $a_l^{(t)} \in DOM(A_l)$ and $x_{j,l}$. The function τ is assigned a value of 0.0 if two values are equal; otherwise it is assigned a value of its confidence degree. $\eta_i (=\sqrt{\sum_{t=1}^{n_l} (\omega_l^{(t)})^2})$ is a normalization factor. Then, the membership degree of x_j to v_i is calculated as

$$\mu_{i,j} = \left(\sum_{z=1}^{k} \left(\frac{d_c(v_i, x_j)}{d_c(v_z, x_j)} \right)^{1/(m-1)} \right)^{-1}. \tag{12}$$

Let us consider the method for updating the population of the ith centroid v_i. The attribute $v_{i,l}$, shown in Eq. (7), is then updated by determining $\omega_l^{(t)}$ for $1 \leqslant t \leqslant n_l$ as follows:

$$\omega_l^{(t)} = \frac{1}{\lambda_i} \sum_{j=1}^{n} \gamma(x_{j,l}), \tag{13}$$

where

$$\gamma(x_{j,l}) = \begin{cases} \mu_{i,j}^m, & a_l^{(t)} = x_{j,l}, \\ 0, & a_l^{(t)} \neq x_{j,l}. \end{cases} \tag{14}$$

$\lambda_i (=\sqrt{\sum_{j=1}^{n} \mu_{i,j}^{2m}})$ is a normalization factor. $v_{i,l}$ 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. In this way, the *k*-populations algorithm iteratively improves a set of clusters until no further improvement in $J_m(V : X)$ is possible.

4. Experimental results

To test the effectiveness of the *k*-populations algorithm, we applied the proposed algorithm and three conventional methods (the hierarchical, *k*-modes, 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, fuzzy *k*-modes, and *k*-populations algorithms were *k* distinct data randomly selected from the data set. For the fuzzy *k*-modes, and *k*-populations algorithms, *m* was varied from 1.1 to 2.0. Four data sets were used to evaluate the performance of each method, specifically, the Soybean, Zoo, Credit, and Hepatitis data sets from the UCI repository [4]. The clustering results were assessed using Huang's accuracy measure (*r*) [3]; a higher value of *r* indicates a better clustering result, with perfect clustering yielding a value of 100.0%.

Table 1
Average clustering accuracy (%) achieved by four clustering methods for the four data sets

Data set	Hierarchical	*k*-modes	Fuzzy *k*-modes	*k*-populations
Soybean database	85.11	79.90	78.26	100.00
Zoo database	69.31	67.66	68.65	86.58
Credit approval	53.73	72.00	73.30	84.83
Hepatitis domain	78.17	67.79	70.57	79.61

Table 1 lists the average accuracy of clustering achieved by each algorithm over 200 runs for the four data sets. The Soybean data set contains 47 data points on diseases in soybeans. Each data point has 35 categorical attributes and is classified as one of the four diseases ($k = 4$). It is evident from Table 1 that the *k*-populations algorithm gives markedly better clustering performance in comparison to other algorithms. The hierarchical algorithm provides an accuracy of 85.11%, and is more accurate than the *k*-modes (79.90%) and fuzzy *k*-modes (78.26%) algorithms. Notably, the *k*-populations algorithm gave an accuracy of 100.0%, making it 14.9% more accurate than the hierarchical algorithm. The second 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). The hierarchical algorithm yielded an accuracy of 69.31%. The *k*-modes and fuzzy *k*-modes algorithms gave accuracies of 67.66% and 68.65% respectively. In contrast, the *k*-populations algorithm gave the superior accuracy of 86.58%. Thus, in this case, the *k*-populations algorithm was 17.3% more accurate than the hierarchical algorithm. The Credit data set contains 202 applicants data for credit approval. Each application is described by nine attributes and classified as approved or rejected ($k = 2$). The *k*-populations algorithm was most accurate, yielding an accuracy of 84.83%. The *k*-modes and fuzzy *k*-modes algorithms gave accuracies of 72.00% and 73.30%, respectively. The hierarchical algorithm showed the lower classification accuracy of 53.73%. In this case, there was 11.5% increase of accuracy by the *k*-populations algorithm than the fuzzy *k*-modes algorithm. The fourth data set, the Hepatitis set, consists of 155 patients' data, where each patient described by 20 categorical attributes is classified as live or die ($k = 2$). The *k*-modes and fuzzy *k*-modes algorithms showed lower classification accuracies of 67.79% and 70.57%, respectively. In contrast,

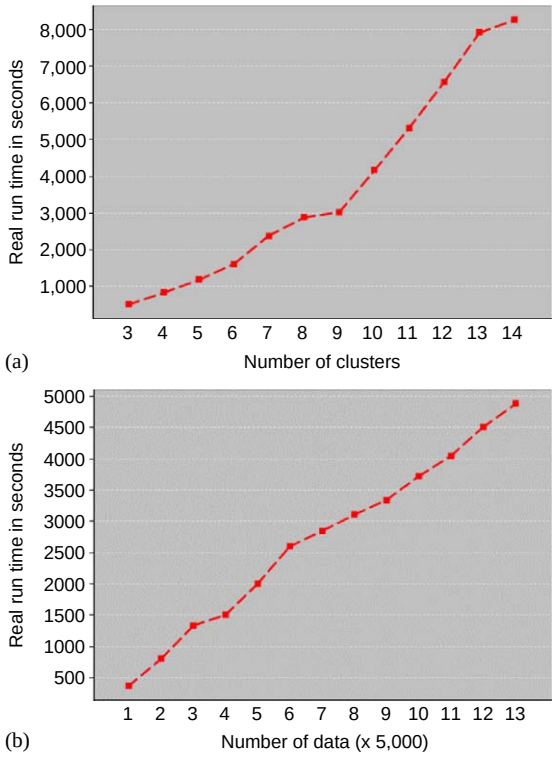

Fig. 1. The scalability of *k*-populations algorithm: (a) scalability to the number of clusters when clustering 65,000 data of the Connect data set; (b) scalability to the number of data when clustering the Connect data set into 10 clusters.

the hierarchical and *k*-populations algorithms gave accuracies of 78.17% and 79.61%, respectively.

Moreover, to test the scalability of the proposed algorithm on large data sets, we applied the *k*-populations algorithm to the Connect data set [4]. This set consists of 65,000 data, where each data point is composed of 42 categorical attributes. We tested two types of scalability: the scalability to the number of clusters for a given number of data and the scalability to the number of data for a given number of clusters. The test was performed on a IBM P690 server using a single processor. Fig. 1(a) shows the real run time to cluster 65,000 data into different numbers of clusters. Fig. 1(b) shows the real run time to cluster different numbers of data into 10 clusters. We see from these figures that the *k*-populations algorithm shows a linear increase in the run time as the number of clusters and the number of data are increased.

5. Conclusions

The conventional *k*-modes-type algorithms are 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 categorical data. Thus, we developed a *k*-populations algorithm in which the notion of population is used to represent the centroid of each cluster. The population is a set of pairs that contain category values and their confidence degrees for each attribute. The superiority of the *k*-populations algorithm over other clustering algorithms was clearly demonstrated through several experiments.

Acknowledgements

This work was supported by the Metabolites Analysis and Function Research Grant (2004-02145) 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

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

[2] Z. Huang, Extensions to the *k*-modes algorithm for clustering large data sets with categorical values, Data Min. Knowl. Disc. 2 (3) (1998) 283–304.

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

[4] C.L. Blake, C.J. Merz, UCI Repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html, 1989.