Journal of Classification 20:257-261 (2003)
DOI: 10.1007/s00357-003-0014-4

A Note on K-modes Clustering

Zhexue Huang

The University of Hong Kong

Michael K. Ng

The University of Hong Kong

Abstract: Recently, Chaturvedi, Green and Carroll (2001) presented a nonparametric approach to deriving clusters from categorical data using a new clustering procedure called K-modes. Huang (1998) proposed the K-modes clustering algorithm. In this note, we demonstrate the equivalence of the two K-modes procedures.

Keywords: Clustering; K-means algorithm; Fuzzy partitioning; Categorical data.

The authors thank the anonymous referees for recommending the materials on the Chaturvedi, Green and Carroll early work of their K-modes methodology and providing valuable comments on the previous version.

Authors' Addresses: Zhexue Huang, E-Business Technology Institute, The University of Hong Kong, Pokfulam Road, Hong Kong; e-mail: jhuang@eti.hku.hk; Michael K. Ng, Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong. Research supported in part by Hong Kong Research Grants Council Grant Nos. HKU 7130/02P, HKU 7046/03P and HKU CRCG Grant Nos. 10203501, 10203907 and 10204437; e-mail: mng@maths.hku.hk.

1. Introduction

The K-means algorithm (e.g. Ball and Hall 1967; MacQueen 1967; Anderberg 1973; Jain and Dubes 1988) is well known for its efficiency in clustering large data sets. However, working only on numeric data limits the use of these K-means algorithms in such areas as data mining where large categorical data sets are frequently encountered.

A commonly used approach for clustering categorical data using the K-means algorithm is to convert nominal variables into binary variables, each representing presence or absence of a category in a nominal variable (e.g., Ralambondrainy 1995). If a data set has nominal variables with many categories, this approach will result in a large number of binary variables. This will inevitably increase both computational cost and memory need of the K-means algorithm. The other drawback is that the cluster means given by real values between 0 and 1 do not indicate the characteristics of the clusters.

To tackle the problem of clustering large categorical data sets in data mining, the K-modes algorithm was proposed by Huang (1998). The K-modes algorithm is a modified version of the K-means algorithm that uses a simple matching dissimilarity measure for categorical variables, modes instead of means for clusters, and a frequency-based method to update modes in the clustering process to minimize the clustering cost function with the matching dissimilarity measure for categorical variables. These modifications to the K-means algorithm enable the K-modes algorithm to efficiently cluster large categorical data sets from real world databases.

Recently, Chaturvedi, Green and Carroll (2001) presented a nonparametric approach to deriving clusters from categorical data using a new clustering procedure called K-modes, which is also analogous to the traditional K-means procedure (MacQueen, 1967) for clustering numerical data. Prior to this publication in Journal of Classification in 2001, the three authors had presented their methodology at three meetings of Classification Society of North America (CSNA) and American Statistical Association in 1994, 1996 and 1997. The aim of this note is to demonstrate the equivalence of the two independently developed K-modes algorithms given in two papers (Huang 1998; Chaturvedi, Green and Carroll 2001).

2. The Huang Algorithm

Let $\mathbf{X}$ be an n-by-m matrix of categorical values. The ith row vector X_i of $\mathbf{X}$ refers to the ith object. The simple matching dissimilarity measure between two vectors X_i and X_l is defined as follows:

$$d(X_i, X_l) \equiv \sum_{j=1}^{m} \delta(x_{i,j}, x_{l,j}) \tag{1}$$

where

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

It is easy to verify that the function d defines a metric space. The K-modes algorithm (Huang, 1998) uses the k-means paradigm to cluster row vectors of $\mathbf{X}$. Let $\mathbf{W} = [w_{i,l}]$ be an n-by-k matrix representing a partitioning of n vectors into k clusters, and

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_k \end{pmatrix}$$

be an k-by-m matrix representing a set of k modes for k clusters. The objective of clustering n vectors into k clusters is to find $\mathbf{W}$ and $\mathbf{Z}$ that minimize

$$F(\mathbf{W}, \mathbf{Z}) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{i,l} d(X_i, Z_l) \tag{2}$$

subject to

$$w_{i,l} \in \{0, 1\}, \quad 1 \leq l \leq k, \ 1 \leq i \leq n, \tag{3}$$

and

$$\sum_{l=1}^{k} w_{i,l} = 1, \quad 1 \leq i \leq n. \tag{4}$$

where $k(\leq n)$ is a known number of clusters,

Minimization of F in (2) with the constraints in (3) and (4) forms a class of constrained nonlinear optimization problems whose solutions are unknown. The usual method towards optimization of F in (2) is to use partial optimization for $\mathbf{Z}$ and $\mathbf{W}$. In this method we first fix $\mathbf{Z}$ and find necessary conditions on $\mathbf{W}$ to minimize F. Then we fix $\mathbf{W}$ and minimize F with respect to $\mathbf{Z}$. Huang (1998) has presented a frequency-based method to update $\mathbf{Z}$, and calculate $\mathbf{W}$ for a given $\mathbf{Z}$. Huang and Ng (1999) have shown that the K-modes algorithm converges to a merely local minimum in a finite number of iterations.

3. The Chaturvedi, Green and Carroll Algorithm

Chaturvedi, Green and Carroll (2001) considered a bilinear clustering model:

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \mathbf{WZ} + \text{ error.} \tag{5}$$

They defined the parameter estimation problem via minimizing an L_0-norm based loss function:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\left[\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} - \mathbf{WZ}\right]_{i,j}\right)^0$$

or

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{i,j} - \sum_{l=1}^{k} w_{i,l}z_{l,j})^0.$$

Here $(\cdot)^0$ corresponds to the “counting metric”. The matrices $\mathbf{W}$ and $\mathbf{Z}$ are estimated iteratively until the L_0 loss function does not improve.

To demonstrate the equivalence of the two independently developed K-modes algorithms, we first rewrite the objective function in (2) as follows:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{l=1}^{k} w_{i,l}\delta(x_{i,j}, z_{l,j}).$$

It suffices to show that the term $\sum_{l=1}^{k} w_{i,l}\delta(x_{i,j}, z_{l,j})$ is equal to the term $(x_{i,j} - \sum_{l=1}^{k} w_{i,l}z_{l,j})^0$ for a given $\mathbf{W}$ with $w_{i,l}$ satisfying (3) and a given $\mathbf{Z}$. We note that for a given $\mathbf{W}$ with $w_{i,l}$ satisfying (3), there is one and only one $w_{i,l'} = 1$ and the others $w_{i,l}$ are zeros for $1 \leq l \leq k$ and $l \neq l'$ for each ith row. Therefore, we have

$$\sum_{l=1}^{k} w_{il}\delta(x_{i,j}, z_{l,j}) = \delta(x_{i,j}, z_{l',j}) = (x_{i,j} - z_{l',j})^0 = (x_{i,j} - \sum_{l=1}^{k} w_{i,l}z_{l,j})^0.$$

Based on this equality, we see that both K-modes clustering algorithms give the same minimizer of $\mathbf{W}$ for a given $\mathbf{Z}$ and the same minimizer of $\mathbf{Z}$ for a given $\mathbf{W}$. Hence the two K-modes clustering algorithms are equivalent. The type of convergence that discussed Section 2 is valid for the Chaturvedi, Green and Carroll Algorithm.

4. Final Remarks

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 algorithm was motivated to solve this problem. In this note, we have demonstrated the equivalence of the two K-modes algorithms recently published in two different journals by Huang (1998) and Chaturvedi, Green and Carroll (2001). Finally, we remark that Huang and Ng (1999) also introduced the fuzzy K-modes algorithm for clustering categorical data based on extensions to the fuzzy K-means algorithm. The fuzzy K-modes algorithm generates the fuzzy partition matrix that 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

ANDERBERG, M. (1973). *Cluster Analysis for Applications*, New York: Academic, 1973.

BALL, G., and HALL, D. (1967). "A Clustering Technique for Summarizing Multivariate Data," *Behavioral Science, 12,* 153-155.

CHATURVEDI, A., GREEN, P., and CARROLL, J. (1994). "K-Means, K-Medians and K-Modes: Special Cases of Partitioning Multiway Data," presented at The Classification Society of North America (CSNA) Meeting, Houston.

CHATURVEDI, A., GREEN, P., and CARROLL, J. (1996). "Market Segmentation via K-Modes Clustering," presented at The American Statistical Association Meeting, Chicago.

CHATURVEDI, A., GREEN, P., and CARROLL, J. (1997). "Empirical Findings Obtained from Evaluating K-Modes and Overlapping K-Centroids Clustering," presented at The Classification Society of North America (CSNA) Meeting, Washington D.C.

CHATURVEDI, A., GREEN, P., and CARROLL, J. (2001.) "K-modes Clustering," *Journal of Classification, 18,* 35-55.

HUANG, Z. (1998). "Extensions to the K-means Algorithm for Clustering Large Data Sets with Categorical Values," *Data Mining and Knowledge Discovery, 2,* 283-304.

HUANG, Z., and NG, M. (1999). "A Fuzzy K-modes Algorithm for Clustering Categorical Data," *IEEE Transactions on Fuzzy Systems, 7,* 446-452.

JAIN, A., and DUBES, R. (1988). *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall.

MACQUEEN, J. (1967). "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of 5th Symposium on Mathematical Statistics and Probability*, Berkeley, CA, Vol. 1, AD 669871. Berkeley, CA: University of California Press, 281-297.

RALAMBONDRAINY, H. (1995). "A Conceptual Version of the K-means Algorithm," *Pattern Recognition Letters, 16,* 1147-1157.