Clustering Categorical Data with k-Modes

Joshua Zhexue Huang
The University of Hong Kong, Hong Kong

INTRODUCTION

A lot of data in real world databases are categorical. For example, gender, profession, position, and hobby of customers are usually defined as categorical attributes in the CUSTOMER table. Each categorical attribute is represented with a small set of unique categorical values such as {Female, Male} for the gender attribute. Unlike numeric data, categorical values are discrete and unordered. Therefore, the clustering algorithms for numeric data cannot be used to cluster categorical data that exists in many real world applications.

In data mining research, much effort has been put on development of new techniques for clustering categorical data (Huang, 1997b; Huang, 1998; Gibson, Kleinberg, & Raghavan, 1998; Ganti, Gehrke, & Ramakrishnan, 1999; Guha, Rastogi, & Shim, 1999; Chaturvedi, Green, Carroll, & Foods, 2001; Barbara, Li, & Couto, 2002; Andritsos, Tsaparas, Miller, & Sevcik, 2003; Li, Ma, & Ogihara, 2004; Chen, & Liu, 2005; Parmar, Wu, & Blackhurst, 2007). The k-modes clustering algorithm (Huang, 1997b; Huang, 1998) is one of the first algorithms for clustering large categorical data. In the past decade, this algorithm has been well studied and widely used in various applications. It is also adopted in commercial software (e.g., Daylight Chemical Information Systems, Inc, http://www.daylight.com/).

BACKGROUND

In data mining, k-means is the mostly used algorithm for clustering data because of its efficiency in clustering very large data. However, the standard k-means clustering process cannot be applied to categorical data due to the Euclidean distance function and use of means to represent cluster centers. To use k-means to cluster categorical data, Ralambondrainy (1995) converted each unique category to a dummy binary attribute and used 0 or 1 to indicate the categorical value either absent or present in a data record. This approach is not suitable for high dimensional categorical data.

The k-modes approach modifies the standard k-means process for clustering categorical data by replacing the Euclidean distance function with the simple matching dissimilarity measure, using modes to represent cluster centers and updating modes with the most frequent categorical values in each of iterations of the clustering process. These modifications guarantee that the clustering process converges to a local minimal result. Since the k-means clustering process is essentially not changed, the efficiency of the clustering process is maintained. The k-modes clustering algorithm was first published and made publicly available in 1997 (Huang, 1997b). An equivalent approach was reported independently in (Chaturvedi, Green, Carroll, & Foods, 2001). The relationship of the two k-modes methods is described in (Huang & Ng, 2003).

In the past decade, a few other methods for clustering categorical data were also proposed, including the dynamic system approach (Gibson, Kleinberg, & Raghavan, 1998), Cactus (Ganti, Gehrke, & Ramakrishnan, 1999), ROCK (Guha, Rastogi, & Shim, 1999), Coolcat (Barbara, Li, & Couto, 2002), and LIMBO (Andritsos, Tsaparas, Miller, & Sevcik, 2003). However, these methods have largely stayed in research stage and not been widely applied to real world applications.

MAIN FOCUS

The k-modes clustering algorithm is an extension to the standard k-means clustering algorithm for clustering categorical data. The major modifications to k-means include distance function, cluster center representation and the iterative clustering process (Huang, 1998).

Distance Function

To calculate the distance (or dissimilarity) between two objects X and Y described by m categorical attributes, the distance function in k-modes is defined as

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

where

$$\delta(x_j,y_j)=\begin{cases}0, & x_j=y_j\\ 1, & x_j\neq y_j.\end{cases} \tag{2}$$

Here, x_j and y_j are the values of attribute j in X and Y. This function is often referred to as simple matching dissimilarity measure or Hemming distance. The larger the number of mismatches of categorical values between X and Y is, the more dissimilar the two objects.

A new distance function for k-modes defines the dissimilarity measure between an object X and a cluster center Z_l as (Ng, Li, Huang, & He, 2007):

$$\phi(x_j,z_j)=\begin{cases}1-\dfrac{n_j^r}{n_l}, & x_j=z_j\\ 1, & x_j\neq z_j\end{cases} \tag{3}$$

where z_j is the categorical value of attribute j in Z_l, n_l is the number of objects in cluster l and n_j^r is the number of objects whose attribute value is r. In this function, when the categorical value of an object is same as the value of cluster center, its distance depends on the frequency of the categorical value in the cluster.

Cluster Modes

In k-modes clustering, the cluster centers are represented by the vectors of modes of categorical attributes. In statistics, the mode of a set of values is the most frequent occurring value. For example, the mode of set $\{a, b, a, a, c, b\}$ is the most frequent value a. There can be more than one mode in a set of values. If a data set has m categorical attributes, the mode vector Z consists of m categorical values $(z_1, z_2, \ldots, z_m)$, each being the mode of an attribute. The mode vector of a cluster minimizes the sum of the distances between each object in the cluster and the cluster center (Huang, 1998).

Clustering Process

To cluster a categorical data set $\boldsymbol{X}$ into k clusters, the k-modes clustering process consists of the following steps:

Step 1: Randomly select k unique objects as the initial cluster centers (modes).

Step 2: Calculate the distances between each object and the cluster mode; assign the object to the cluster whose center has the shortest distance to the object; repeat this step until all objects are assigned to clusters.

Step 3: Select a new mode for each cluster and compare it with the previous mode. If different, go back to Step 2; otherwise, stop.

This clustering process minimizes the following k-modes objective function

$$F(U,Z)=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m}u_{i,l}d\left(x_{i,j},z_{l,j}\right)$$

where $U=[u_{i,j}]$ is an $n\times k$ partition matrix, $Z=\{Z_1, Z_2, \ldots, Z_k\}$ is a set of mode vectors and the distance function $d(.,.)$ is defined as either (2) or (3). Since it is essentially same as k-means, the k-modes clustering algorithm is efficient in clustering large categorical data and also produces locally minimal clustering results.

Fuzzy k-Modes and Other Variants

The fuzzy k-modes clustering algorithm is an extension to k-modes (Huang & Ng, 1999). Instead of assigning each object to one cluster, the fuzzy k-modes clustering algorithm calculates a cluster membership degree value for each object to each cluster. Similar to the fuzzy k-means, this is achieved by introducing the fuzziness factor in the objective function (Huang & Ng, 1999). The fuzzy k-modes clustering algorithm has found new applications in bioinformatics (Thornton-Wells, Moore, & Haines, 2006). It can improve the clustering result whenever the inherent clusters overlap in a data set.

The k-prototypes clustering algorithm combines k-means and k-modes to cluster data with mixed numeric and categorical values (Huang, 1997a). This is achieved