K-modes Clustering

Anil Chaturvedi
Kraft Foods

Paul E. Green
The Wharton School, University of Pennsylvania

J. Douglas Carroll
Rutgers University

Authors’ Addresses: Anil Chaturvedi, GV529, 1 Kraft Court, Glenview, IL 60025, USA; Paul E. Green, Suite 1450, Dietrich Hall, Philadelphia, PA, 19104, USA; J. Douglas Carroll, Rutgers University, Graduate School of Management, MEC 125, 111 Washington Street, Newark NJ 07102-3027, USA.
Abstract: We present a nonparametric approach to deriving clusters from categorical (nominal scale) data using a new clustering procedure called K-modes, which is analogous to the traditional K-Means procedure (MacQueen 1967) for clustering interval scale data. Unlike most existing methods for clustering nominal scale data, the K-modes procedure explicitly optimizes a loss function based on the $L_p$ norm (defined as the limit of an $L_p$ norm as $p$ approaches zero).

In Monte Carlo simulations, both K-modes and latent class procedures (e.g., Goodman 1974) performed with equal efficiency in recovering a known underlying cluster structure. However, K-modes is an order of magnitude faster than the latent class procedure in speed and suffers from fewer problems of local optima than do latent class procedures. For data sets involving a large number of categorical variables, latent class procedures become computationally extremely slow and hence infeasible.

We conjecture that, although in some cases latent class procedures might perform better than K-modes, it could out-perform latent class procedures in other cases. Hence, we recommend that these two approaches be used as "complementary" procedures in performing cluster analysis. We also present an empirical comparison of K-modes and latent class, where the former method prevails.

Keywords: Categorical data; Cluster analysis; Groups; Modes; Latent class analysis

1. Introduction

This paper presents a simple procedure for clustering of nominal scale data. The procedure, which we call K-modes clustering, is analogous to MacQueen's (1967) K-means clustering procedure. Input data for K-means clustering procedures must generally have either interval or ratio scale properties. In contrast, researchers frequently need clusters based on nominal scale data. K-means procedures are generally inappropriate for such categorical data.

Five techniques commonly used for finding clusters from categorical data entail the following: (a) dummy code the categorical variables, compute intersubject distances from the dummy coded data, and use such hierarchical clustering procedures as single, complete, or average linkage on the derived intersubject distances; (b) dummy code the categorical variables, and use K-means on these dummy variables; (c) use correspondence analysis to derive spatial coordinates (e.g., see Carroll, Green, and Schaffner 1986) for each subject, and then use K-means on the derived spatial coordinates; (d) use latent class procedures (e.g., Goodman 1974) available for contingency table analysis; and (e) use Hartigan's Ditto Algorithm (1975, pp. 143-154) for categorical data.

The first three procedures have some drawbacks. In (a), one needs to select a distance measure from among many candidate choices. In addition,