Ann Inst Stat Math (2010) 62:11–35
DOI 10.1007/s10463-009-0258-9

Latent class analysis variable selection

Nema Dean · Adrian E. Raftery

Received: 19 July 2008 / Revised: 22 April 2009 / Published online: 24 July 2009

Abstract We propose a method for selecting variables in latent class analysis, which is the most common model-based clustering method for discrete data. The method assesses a variable's usefulness for clustering by comparing two models, given the clustering variables already selected. In one model the variable contributes information about cluster allocation beyond that contained in the already selected variables, and in the other model it does not. A headlong search algorithm is used to explore the model space and select clustering variables. In simulated datasets we found that the method selected the correct clustering variables, and also led to improvements in classification performance and in accuracy of the choice of the number of classes. In two real datasets, our method discovered the same group structure with fewer variables. In a dataset from the International HapMap Project consisting of 639 single nucleotide polymorphisms (SNPs) from 210 members of different groups, our method discovered the same group structure with a much smaller number of SNPs.

Keywords Bayes factor · BIC · Categorical data · Feature selection · Model-based clustering · Single nucleotide polymorphism (SNP)