Using Nominal Variables to Classifying College Enrollments

Arnold L Hook, University of Louisville, Louisville, KY

Abstract
SAS PROC FASTCLUS is a clustering procedure available in the SAS Statistical package and in SAS Enterprise Miner. It is an efficient classification scheme when the data set is large and there are one or more quantitative variables. In this example, a clustering of college students with undecided majors was done in order to observe similarity with those students already in a decided major and possibly predict which majors the undecided students would eventually choose. Here students are clustered into one of several possible clusters, each representing a particular major, using various student attributes that consist of both interval and nominal data. The problem of using nominal variables in any clustering scheme is that these variables need to be recoded into units that are useable in a clustering algorithm, in that they need to be defined in metric terms so geometric distances can be defined. This example demonstrates how using nominal data in PROC FASTCLUS can be overcome.