MASTER'S SEMINAR ANNOUNCEMENT
Department of Statistics

Model-Based Clustering Via the Bayesian Information Criterion with Applications and Comparison with K-means Algorithm

Tuesday, November 11, 2003, 2:00 pm
Eckhart Hall, Room 110, 5734 S. University Avenue

Yuan Li
Department of Statistics, University of Chicago

ABSTRACT

Data mining aims to discover unknown knowledge in data sets. An important step in data mining is the discovery of structural features in a data set without using prior knowledge. Here we consider the situation in which the training data is not available, known as the clustering analysis. One difficult question we are often faced with in clustering analysis is how to choose the number of clusters.

In this paper, we review the approach first introduced by Banfield and Raftery (1993). Based on parsimonious geometric modeling of the within-group covariance matrices in a mixture of multivariate normal distributions, the approach uses hierarchical and iterative relocations. In addition, the Bayesian Information Criterion (BIC), a model selection criterion in the statistics literature, is proposed to estimate the number of clusters. Unlike significance tests, this allows multiple comparisons and removes the restriction that the models compared must be nested. The problems of determining the number of clusters and the clustering method are solved simultaneously by choosing the best model. Partitions are determined by the EM (expectation-maximization) algorithm for maximum likelihood, with initial values obtained from agglomerative hierarchical clustering.

We apply our method to the geographical location of the Lansing Woods hickories, and present the Gaussian mixture modeling, and shows that the BIC is able to choose the number of clusters according to the intrinsic complexity present in the data. Given the absence of the training data, we compare the model-based clustering with K-means method, the traditional relocation clustering method, by assessing the pairing preserving error rate computed from the simulated bootstrapping data sets. In the second example, we apply our method to classify the Italian olive oil based on its eight fatty acids. Like in the first example, we use the pairing preserving error rate to compare the model-based clustering and the K-means algorithm.