

Semi-parametric Clustering with Applications to Microarray Data Analysis

Ao Yuan, *National Human Genome Center, Howard University, Washington D.C., USA*

Wenqing HE, *University of Western Ontario, London, Ontario, Canada*, E-mail: whe@stats.uwo.ca

KEY WORDS: clustering, EM algorithm, microarray data, mixture model, nonparametric model, semi-parametric model

MATHEMATICAL SUBJECT CLASSIFICATION: 62H30, 62P10

Abstract: The existing clustering algorithms fall mainly into two categories, model-based (parametric) and non-model-based (nonparametric) methods. Parametric methods perform well when the specific model approximately fit data, but not so when there is non-negligible deviation between them. Nonparametric methods are robust, but efficiency loss may become an important issue in some situations. In this talk we discuss a semi-parametric mixture method in which the mixture proportions are specified as unknown parameters while the sub-distribution of each cluster is modeled nonparametric. The EM-algorithm along with a classification step is used to cluster the data, and the BIC is used to guide the determination of the optimal number of clusters. This method is formulated into clustering models for various types of microarray data analysis, and as an illustration, it is applied to a real gene expression data set. Simulation studies show the proposed method performances relatively well and are more robust than some of the commonly used parametric methods.