Clustering and Outlier Detection by the EM Algorithm based on the Restriction Principle

Vedran Novoselac
Mechanical Engineering Faculty in Slavonski Brod, University of Osijek

Abstract. Cluster analysis and outlier detection present important topics in data mining, and in most cases they are studied separately. In this work the joint problem of clustering and outlier detection is considered. The problem is observed for data that are modeled by a random sample whose distribution is a mixture of Gaussians. Considering the form of statistical modeling in outlier analysis which is based on a level of statistical significance of the tails of a observed density function; the problem is resolved by the restriction of the hidden variable of the well known Expectation Maximization (EM) algorithm. In that sense the adaptive framework is developed which effectively preserve the cluster's structure, or in other senses detect outliers. The general problem is set as the optimization of the proposed algorithm in terms of the cluster validity criteria. For that purpose, new clustering quality measures are proposed. It is established by experminetal reserach, which are conducted on various numerical examples that the proposed method possesses the convergence property. This method is emphasized in digital image processing for pattern recognition.

Keywords: Outliers, Data clustering, Gaussian mixture, EM, Mahalanobis distance, Cluster Validation