Cluster Analysis in Fitting Mixtures of Curves

Tom Burr
Los Alamos National Laboratory, USA

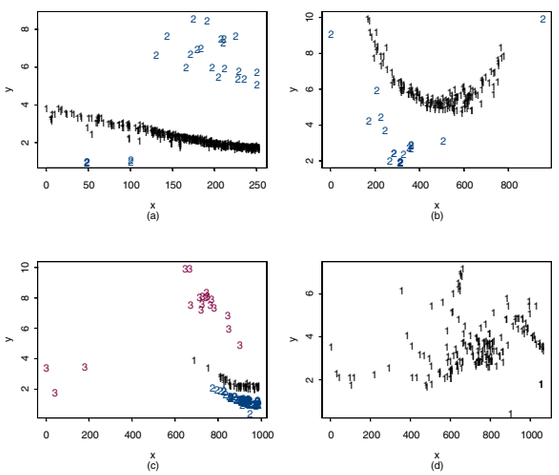
INTRODUCTION

One data mining activity is cluster analysis, of which there are several types. One type deserving special attention is clustering that arises due to a mixture of curves. A mixture distribution is a combination of two or more distributions. For example, a bimodal distribution could be a mix with 30% of the values generated from one unimodal distribution and 70% of the values generated from a second unimodal distribution. The special type of mixture we consider here is a mixture of curves in a two-dimensional scatter plot. Imagine a collection of hundreds or thousands of scatter plots, each containing a few hundred points, including background noise, but also containing from zero to four or five bands of points, each having a curved shape. In a recent application (Burr et al., 2001), each curved band of points was a potential thunderstorm event (see Figure 1), as observed from a distant satellite, and the goal was to cluster the points into groups associated with thunderstorm events. Each curve has its own shape, length, and location, with varying degrees of curve overlap, point density, and noise magnitude. The scatter plots of points from curves having small noise resemble a smooth curve with very little vertical variation from the curve, but there can be a wide range in noise magnitude so that some events have large vertical variation from the center of the band. In this context, each curve is a cluster and the challenge is to use only the observations to estimate how many curves comprise the mixture, plus their shapes and locations. To achieve that goal, the human eye could train a classifier by providing cluster labels to all points in example scatter plots. Each point either would belong to a curved region or to a catch-all noise category, and a specialized cluster analysis would be used to develop an approach for labeling (clustering) the points generated according to the same mechanism in future scatter plots.

BACKGROUND

Two key features that distinguish various types of clustering approaches are the assumed mechanism for how the data is generated and the dimension of the data. The data-generation mechanism includes deterministic and stochastic components and often involves deterministic mean shifts between clusters in high dimensions. But there are other settings for cluster analysis. The particular one discussed here (see Figure 1) is a mixture of curves, where any notion of a cluster mean would be quite different from that in more typical clustering applications. Furthermore, although finding clusters in a two-dimensional scatter plot seems less challenging than in higher-dimensions (the trained human eye is likely to perform as well as any machine-automated method, although the eye would be slower), complications include overlapping clusters; varying noise magnitude; varying feature and noise and density; varying feature shape, locations, and length; and varying types of noise (scene-wide and event-specific). Any one of these complications would justify treating the fitting of curve mixtures as an important special case of cluster analysis. Although as in pattern recognition, the following methods discussed require training scatter plots with points labeled according to their cluster memberships, we regard this as cluster analysis rather than pattern recognition, because all scatter plots have from zero to four or five clusters whose shape, length, location, and extent of overlap with other clusters vary among scatter plots. The training data can be used both to train clustering methods and then to judge their quality. Fitting mixtures of curves is an important special case that has received relatively little attention to date. Fitting mixtures of probability distributions dates to Titterington et al. (1985), and several model-based clustering schemes have been developed (Banfield & Raftery, 1993; Bensmail et al., 1997; Dasgupta & Raftery, 1998), along with associated theory (Leroux, 1992). However, these models assume that the mixture is a mixture of probability distributions (often Gaussian, which can be long and thin, ellipsoidal, or more circular) rather than curves. More recently, methods for mixtures of curves have been introduced, including a mixture of principal curves model (Stanford & Raftery, 2000), a mixture of regression models (Gaffney & Smyth 2003; Hurn, Justel, & Robert, 2003; Turner, 2000), and mixtures of local regression models (i.e., smooth curves obtained using splines or nonparametric kernel smoothers for example).

Figure 1. Four mixture examples containing (a) one, (b) one, (c) two, and (d) zero thunderstorm events plus background noise. The label "1" is for the first thunderstorm in the scene, "2" for the second, and so forth., and the highest integer label is reserved for the catch-all noise class. Therefore, in (d), because the highest integer is 1, there is no thunderstorm present (the mixture is all noise).



MAIN THRUST

Several methods have been proposed for fitting mixtures of curves. In method 1 (Burr et al., 2001), first use density estimation to reject the background noise points such as those labeled as 2 in Figure 1a. For example, each point in the scatter plot has a distance to its *k*th nearest neighbor, which can be used as a local density estimate (Silverman, 1986) to reject noise points. Next, use a distance measure that favors long, thin clusters (e.g., let the distance between clusters be the minimum distance between any a point in the first cluster and a point in the second cluster), together with standard hierarchical clustering to identify at least the central portion of each cluster. Alternatively, model-based clustering favoring long, thin Gaussian shapes (Banfield & Raftery, 1993) or the fitting straight lines method in Murtagh and Raftery (1984) or Campbell et al. (1997) are effective for finding the central portion of each cluster. A curve fitted to this central portion can be extrapolated and then used to accept other points as members of the cluster. Because hierarchical clustering cannot accommodate overlapping clusters, this method assumes that the central portions of each cluster are non-overlapping. Points away from the central portion from one cluster that lie close to the curve fitted to the central portion of the cluster can overlap with points from another cluster. The noise points are identified initially as those having low local density (away from the central portion of any cluster) but, during the extrapolation, can be judged to be a cluster member, if they lie near the extrapolated curve. To increase robustness, method 1 can be applied twice, each time using slightly different inputs (such as the decision threshold for the initial noise rejection and the criteria for accepting points into a cluster that are close to the extrapolated region of the cluster's curve). Then, only clusters that are identified both times are accepted.

Method 2 uses the minimized, integrated squared error (ISE, or L_2 distance) (Scott, 2002; Scott & Szewczyk, 2002) and appears to be a good approach for fitting mixture models, including mixtures of regression models, as is our focus here. Qualitatively, the minimum L_2 distance method tries to find the largest portion of the data that matches the model. In our context, at each stage, the model is all the points belonging to a single curve plus everything else. Therefore, we first seek cluster 1 having the most points, regard the remaining points as noise, remove the cluster, then repeat the procedure in search of feature 2, and so on, until a stop criterion is reached. It also should be possible to estimate the number of components in the mixture in the first evaluation of the data, but that approach has not yet been attempted. Scott (2002) has shown that in the parametric setting with model $f(x|q\theta)$, we estimate θ using $\hat{\theta} = \arg\min_{\theta} \int [f(x|\theta) - f(x|\hat{\theta}_0)]^2 dx$ where the true parameter θ_0 is unknown. It follows that a reasonable estimator minimizing the parametric ISE criterion is $\hat{\theta}_{L_2E} = \arg\min_{\theta}[\int f(x|\theta)^2 dx - \frac{2}{n}\sum_{i=1}^{n} f(x_i|\theta)]$. This assumes that the correct parametric family is used; the concept can be extended to include the case in which the assumed parametric form is incorrect in order to achieve robustness.

Method 3 (principal curve clustering with noise) was developed by Stanford and Raftery (2000) to locate principal curves in noisy spatial point process data. Principal curves were introduced by Hastie and Stuetzle (1989). A principal curve is a smooth curvilinear summary of *p*-dimensional data. It is a nonlinear generalization of the first principal component line that uses a local averaging method. Stanford and Raftery (2000) developed an algorithm that first uses hierarchical principal curve clustering (HPCC, which is a hierarchical and agglomerative clustering method) and next uses iterative relocation (reassign points to new clusters) based on the classification estimation-maximization (CEM) algorithm. A probability model included the principal curve probability model for the feature clusters and a homogeneous Poisson process model for the noise cluster. More specifically, let X denote the set of