A Mixture-based Approach to Latent Class Discovery

Jeffrey L. Solka \hspace{1cm} Wendy L. Martinez
Code B10 \hspace{1cm} Code 311
NSWCDD \hspace{1cm} Office of Naval Research
Dahlgren, VA 22448 \hspace{1cm} Arlington, VA 22217

Abstract

Given a set of pre-categorized n-dimensional observations, one is often interested in discovering hidden categories within the pre-defined categories. This paper will present new work that takes a Gaussian-mixtures-based approach to the problem. This approach uses Gaussian mixture models to subcategorize observations based on their relationship to the mixture-induced discriminant boundary. The methodology is illustrated using two synthetic datasets.

Keywords

adaptive mixtures, mixture-based latent class discovery

Introduction

The proposed methodology is aimed at the discovery of latent classes that might reside within a set of pre-categorized observations. Given a set of \( n \) observations residing in \( \mathbb{R}^p \) and a set of associated class labels \( C \in \{0,1\} \) one is interested in discovering latent structure within the class 0 and class 1 observations. The discovered latent class structure may be relevant to some unknown characterization of the data.

There are a number of different ways that this process can be facilitated. We are particularly interested in those latent classes that reveal their class structure based on their relationship to a discriminant boundary that separates the two classes. So our philosophy is that if we are going to build a discriminant boundary that separates the two classes then why don’t we proceed a little further and use the information regarding the relationship of the discriminant boundary to the observation as an “enabling technology” in order to facilitate the discovery of latent classes within the observations. This process of discovery is accomplished by merely clustering the observations based on their relationship to the discriminant boundary as measured with a surrogate, namely the posterior probability that the observation comes from the particular class.

We still have not discussed how best to ascertain the discriminant boundary separating the observations. One could certainly proceed forward with some sort of tree-based or other method of estimating the discriminant boundary directly. We prefer to proceed in a classical Bayes-based manner and estimate the class conditional probability density describing each of the classes. We can then easily
compute the posterior probability of class membership, for class 0 and class 1 by simply using Bayes Theorem.

\[
P(c_j | \tilde{x}) = \frac{p(\tilde{x} | c_j) P(c_j)}{p(\tilde{x})}
\]

(1)

where

\[
p(\tilde{x}) = \sum_{j=1}^{2} p(\tilde{x} | c_j) P(c_j).
\]

(2)

We note that \(P(c_j)\) is the prior probability of membership in class \(c_j\) and \(p(\tilde{x} | c_j)\) is the class conditional probability density function for those observations belonging to class \(c_j\).

We have still not indicated how we are going to estimate the class conditional probabilities. We choose to use a semi-parametric mixture-based density estimation procedure. This procedure, which was originally invented by Priebe [Priebe, 1994], is termed the adaptive mixtures density estimation (AMDE) method. The AMDE is a hybrid of a kernel-based and mixtures-based density estimator. It is based on a recursive mixture of normals model formulation that allows for automatic data driven term creation as dictated by the complexity of the data. We propose to use the AMDE as a means to estimate the class conditional probability density functions. Priebe previously had demonstrated the \(L^1\) convergence of the AMDE procedure. This makes it an ideal candidate for our mixtures-based latent class discovery [Priebe, 1994].

**Approach**

We will describe our proposed methodology in the case of a dataset drawn from one of the two unknown class conditional distributions \(f(\tilde{x} | c_0)\). We can think of the AMDE procedure as a methodology wherein as we are presented with a new observation we either update the existing mixture model using the recursive update equations or else choose to create a new term in the model specifically to accommodate this new observation. The choice to create a new term is made with a probability \(B\) that may be a function of the current observation and state of the model \(\Psi_t\). A is defined simply as \(1 - B\). The AMDE methodology is given below.

\[
\hat{\Psi}_{t+1} = \hat{\Psi}_{t} + A(U_t(\tilde{x}_{t+1}; \hat{\Psi}_{t})) + B(C_t(\tilde{x}_{t+1}; \hat{\Psi}_{t}))
\]

(3)

When we do not create a new term then we update our existing model using the recursive form of the update equations.

The recursive EM update equations are given below.

\[
\hat{\pi}^{(i)}_{n+1} = \frac{\hat{\pi}^{(i)}_{n} f^{(i)}(\tilde{x}_{n+1}; \hat{\theta}_n)}{\sum_{t=1}^{g} \hat{\pi}^{(i)}_{n} f^{(i)}(\tilde{x}_{n+1}; \hat{\theta}_n)}
\]

(4)

\[
\hat{\pi}^{(i)}_{n+1} = \hat{\pi}^{(i)}_{n} + \frac{1}{n} \left( \hat{\pi}^{(i)}_{n+1} - \hat{\pi}^{(i)}_{n} \right)
\]

(5)

\[
\hat{\mu}^{(i)}_{n+1} = \hat{\mu}^{(i)}_{n} + \frac{\hat{\pi}^{(i)}_{n}}{n \hat{\pi}^{(i)}_{n}} \left( \tilde{x}_{n+1} - \hat{\mu}^{(i)}_{n} \right)
\]

(6)

\[
\hat{\Sigma}^{(i)}_{n+1} = \hat{\Sigma}^{(i)}_{n} + \frac{\hat{\pi}^{(i)}_{n}}{n \hat{\pi}^{(i)}_{n}} \left[ (\tilde{x}_{n+1} - \hat{\mu}^{(i)}_{n}) (\tilde{x}_{n+1} - \hat{\mu}^{(i)}_{n})^T - \hat{\Sigma}^{(i)}_{n} \right].
\]

(7)
\( \hat{\pi}^{(i)}_{n+1} \) is the posterior probability that the \( n+1 \) observation comes from the \( i \)-th term of the mixture, \( \hat{\pi}^{(i)}_{n+1} \) is the updated mixing coefficient associated with the \( i \)-th term of the mixture, \( \hat{\mu}^{(i)}_{n+1} \) is the updated mean of the \( i \)-th term of the mixture, and \( \hat{\Sigma}^{(i)}_{n+1} \) is the updated covariance matrix associated with the \( i \)-th term of the mixture.

A new term is created based on the Mahalanobis distance from the observations to existing terms in the model. The Mahalanobis distance is given below

\[
MHD(i, \vec{x}_t) = (\vec{x}_t - \hat{\mu}^{(i)}_{t})^T \Sigma^{-1}(i)(\vec{x}_t - \hat{\mu}^{(i)}_{t}).
\] (8)

If \( MHD(i, \vec{x}_t) > \tau_c \) for every term then a new term is created with \( \vec{\mu}^{(new)} = \vec{x}_t \), \( \Sigma^{(new)} = W(\Sigma^{(i)}) \), and \( \hat{\pi}^{(new)} = \frac{1}{n} \). The new terms covariance is based on a weighted sum, \( W(\Sigma^{(i)}) \), of the covariances for the existing terms in the model. The explicit form of the weighted sum is given below

\[
\Sigma^{(new)} = \sum_{i=1}^{g} \left[ \frac{MHD(i, \vec{x}_t)}{\sum_{j=1}^{g} MHD(j, \vec{x}_t) \Sigma^{(i)}} \right] \Sigma^{(i)}
\] (9)

where we have assumed that there are currently \( g \) terms in the model.

We have after the application of the AMDE an overdetermined mixture model that “fits” the class 0 data. This process is repeated for the class 1 data.

So at this point in time we have our two class conditional probability functions \( \hat{f}(\vec{x}|c_0) \) and \( \hat{f}(\vec{x}|c_1) \). Let’s now assume for the sake of our discussions that we wish to ascertain the latent class structure that might be resident in the class 0 observations. We first compute the posterior probability of membership in class 0 for each of the observations, \( \vec{x} \) in class 0. For example the calculation of said posterior probability at the point \( \vec{x}_0 \) is given below.

\[
\hat{P}_r(c_0|\vec{x}_0) = \frac{P(c_0)\hat{f}(\vec{x}_0|c_0)}{P(c_0)\hat{f}(\vec{x}_0|c_0) + P(c_1)\hat{f}(\vec{x}_0|c_1)}
\] (10)

So we now have \( \hat{P}_r(c_0|\vec{x}) \) for every \( \vec{x} \) in class 0. We now wish to cluster these values. We will then use the clusters obtained from these values as our indicator of latent class structure within the observations. There are of course a number of different ways that one could cluster the posterior probabilities. We have decided for this particular research to proceed forward using a hierarchical clustering methodology. In this manner one can mine the resultant set of cluster labeled observations looking for interesting latent classes.

**Results**

Our first test case consists of 100 observations drawn from a bivariate normal distribution with a mean of 0 and a covariance matrix given by the 2 x 2 identity matrix. A scatterplot for the first test case where the class 0 observations have been colored red and the class 1 observations have been colored green is shown in Figure 1.

Figure 2 presents the adaptive mixture model for the first class for the first test case. Each term in the mixture model has been plotted centered at the mean of the associated term with a shape determined by the covariance matrix associated with the term. The height or z-coordinate of the term is determined by the mixing proportion associated with the term. The reader is referred to [Solka et al., 1995] for a full treatment of this visualization methodology. This model forms our \( \hat{f}(\vec{x}|c_0) \).
Figure 3 presents the adaptive mixture model for the second class for the first test case using an identical visualization strategy.

Figure 4 presents a scatterplot where the class 0 observations have been colored based on a hierarchical clustering [Duda and Hart, 1973] of the class 0 posterior probability values using the adaptive mixture models. We have rendered the class 0 observations using circles with the colors chosen from the set {green, red, cyan, and blue} and the class 1 observations using black plus signs. The 4 colors reflect the fact that we have utilized a 4 cluster solution obtained from the hierarchical clustering tree. The 4 term solution was based on a complete linkage agglomerative clustering solution. So each set of colored observations represents a potential latent class that has been discovered using the procedure.

Figure 5 presents a scatterplot for the second test case. The class 0 sample, consisting of 100 observations, was drawn from a mixture of two terms, each with a mixing coefficient of .5, identity covariance matrices and a first term mean of (-2,0) and a second term mean of (0,-2). The class 0 observations have been colored red. The class 1 sample, consisting of 100 observations, was drawn from a mixture of two terms, each with a mixing coefficient of .5, identity covariances and a first term mean of (2,0) and a second term mean of (0,2). The class 1 observations are plotted in green.

Figure 6 presents the adaptive mixture model for class 0 for the second test case. Figure 7 presents the adaptive mixture model for class 1 for the second test case.

Figure 8 presents a scatter plot where the class 0 observations have been colored based on a hierarchical clustering of the class 0 posterior probability values from the adaptive mixture models. The complexity of the underlying mixture models has led to some more interesting structure in the latent classes discovered via the posterior probability clustering.

**Discussions and Conclusions**

We have discussed a new method of latent class discovery based on the use of adaptive mixtures models. This method utilizes estimated posterior probability of class membership in order to discover latent classes that are resident within a particular known class. In this manner latent classes are discovered based on the relationship of the observations to the Bayes-based discriminant boundary as computed using adaptive mixtures estimates of the class conditional probability densities. The latent classes are obtained based on hierarchically clustering the estimated posterior probability values associated with a particular class’s observations. We have illustrated the application of our technique on two artificial datasets.

Future work will include additional testing of the procedure using artificial and “real world” datasets. Our hope is that we will be able to demonstrate the efficacy of the procedure using datasets from the areas of image-based region of interest identification, gene expression analysis, and text data mining.

**Acknowledgments**

The first author (JLS) would like to acknowledge the support of the Office of Naval Research (ONR) In-house Laboratory Independent Research (ILIR) Program while the second author (WLM) would like to acknowledge the support of the ONR Research Opportunities for Program Officers (ROPO) program.
References


Figure 1: Scatterplot of the first test case. Class 1 observations have been colored red and class 2 observations have been colored green.
Figure 2: Adaptive mixture model for the first test case class 1.
Figure 3: Adaptive mixture model for the first test case class 2.
Figure 4: A scatter plot where the class 0 observations have been colored based on a hierarchical clustering of the posterior probabilities based on the adaptive mixtures model. The class 0 observations have been plotted using circles. The colors of the circles are obtained based on a 4 cluster complete linkage agglomerative clustering solution.
Figure 5: Scatterplot of the second test case.
Figure 6: Adaptive mixture model for the second test case class 1.
Figure 7: Adaptive mixture model for the second test case class 2.
Figure 8: A scatter plot where the class 0 observations have been colored based on a hierarchical clustering of the posterior probabilities based on the adaptive mixtures model. The class 0 observations have been plotted using circles. The colors of the circles are obtained based on a 4 cluster complete linkage agglomerative clustering solution.