Comparison of Local Classification Methods

Julia Schiffner and Claus Weihs

Department of Statistics, University of Dortmund,
44221 Dortmund, Germany
schiffner@statistik.uni-dortmund.de

Abstract. In this paper four local classification methods are described and their statistical properties in the case of local data generating processes (LDGPs) are compared. In order to systematically compare the local methods and LDA as global standard technique, they are applied to a variety of situations which are simulated by experimental design. This way, it is possible to identify characteristics of the data that influence the classification performances of individual methods. For the simulated data sets the local methods on the average yield lower error rates than LDA. Additionally, based on the estimated effects of the influencing factors, groups of similar methods are found and the differences between these groups are revealed. Furthermore, it is possible to recommend certain methods for special data structures.

1 Introduction

We consider four local classification methods that all use the Bayes decision rule. The Common Components and the Hierarchical Mixture Classifiers, as well as Mixture Discriminant Analysis (MDA), are based on mixture models. In contrast, the Localized LDA (LLDA) relies on locally adaptive weighting of observations. Application of these methods can be beneficial in case of LDGPs. That is, there is a finite number of sources where each one can produce data of several classes. The local data generation by individual processes can be described by local models. The LDGPs may cause, for example, a division of the data set at hand into several clusters containing data of one or more classes. For such data structures global standard methods may lead to poor results. One way to obtain more adequate methods is localization, which means to extend global methods for the purpose of local modeling. Both MDA and LLDA can be considered as localized versions of Linear Discriminant Analysis (LDA).

In this paper we want to examine and compare some of the statistical properties of the four methods. These are questions of interest: Are the local methods appropriate to classification in case of LDGPs and do they perform better than global methods? Which data characteristics have a large impact on the classification performances and which methods are favorable to special data structures? For this purpose, in a
simulation study the local methods and LDA as widely-used global technique are applied systematically to a large variety of situations generated and simulated by experimental design. This paper is organized as follows: First the four local classification methods are described and compared. In section 3 the simulation study and its results are presented. Finally, in section 4 a summary is given.

2 Local classification methods

2.1 Common Components Classifier – CC Classifier

The CC Classifier (Titsias and Likas (2001)) constitutes an adaptation of a radial basis function (RBF) network for class conditional density estimation with full sharing of kernels among classes. Miller and Uyar (1998) showed that the decision function of this RBF Classifier is equivalent to the Bayes decision function of a classifier where class conditional densities are modeled by mixtures with common mixture components.

Assume that there are $K$ given classes denoted by $c_1, \ldots, c_K$. Then in the common components model the conditional density for class $c_k$ is

$$f_\theta(x|c_k) = \sum_{j=1}^{G_{CC}} \pi_{jk} f_{\theta_j}(x|j) \quad \text{for } k = 1, \ldots, K,$$

where $\theta$ denotes the set of all parameters and $\pi_{jk}$ represents the probability $P(j|c_k)$. The densities $f_{\theta_j}(x|j), j = 1, \ldots, G_{CC}$, with $\theta_j$ denoting the corresponding parameters, do not depend on $c_k$. Therefore all class conditional densities are explained by the same $G_{CC}$ mixture components.

This implicates that the data consist of $G_{CC}$ groups that can contain observations of all $K$ classes. Because all data points in group $j$ are explained by the same density $f_{\theta_j}(x|j)$ classes in single groups are badly separable. The CC Classifier can only perform well if individual groups mainly contain data of a unique class. This is more likely if the parameter $G_{CC}$ is large. Therefore the classification performance depends heavily on the choice of $G_{CC}$.

In order to calculate the class posterior probabilities the parameters $\theta_j$ and the priors $\pi_{jk}$ and $P_k := P(c_k)$ are estimated based on maximum likelihood and the EM algorithm. Typically, $f_{\theta_j}(x|j)$ is a normal density with parameters $\theta_j = \{\mu_j, \Sigma_j\}$. A derivation of the EM steps for the gaussian case is given in Titsias and Likas (2001), p. 989.

2.2 Hierarchical Mixture Classifier – HM Classifier

The HM Classifier (Titsias and Likas (2002)) can be considered as extension of the CC Classifier. We assume again that the data consist of $G_{HM}$ groups. But additionally, we suppose that within each group $j$, $j = 1, \ldots, G_{HM}$, there are class-labeled