Local Modelling in Classification

Gero Szepannek\textsuperscript{1}, Julia Schiffner\textsuperscript{1}, Julie Wilson\textsuperscript{2}, and Claus Weihs\textsuperscript{1}

\textsuperscript{1} Department of Statistics, University of Dortmund
44227 Dortmund, Germany
\textsuperscript{2} Department of Mathematics and Chemistry, University of York
Heslington, York YO1 5DD, UK
szepannek@statistik.uni-dortmund.de

Abstract. In classification tasks it may sometimes not be meaningful to build single rules on the whole data. This may especially be the case if the classes are composed of several subclasses. Several common as well as recent issues are presented to solve this problem. As it can e.g. be seen in Weihs et al. \textsuperscript{2006} there may result strong benefit from such local modelling. All presented methods are evaluated and compared on four real-world classification problems in order to obtain some overall ranking of their performance following an idea of Hornik and Meyer \textsuperscript{2007}.

1 Introduction

In the context of local modelling it has to be distinguished between global classification models generating one single classification rule that holds for the entire data and local classification models that do not hold for the entire population but rather for a subsample (see e.g. Morik et al., \textsuperscript{2004}). To give an idea of the problem, three simple examples of different artificial data sets in the two dimensional space are shown in Figure 1. Both data sets on the left hand side can be correctly classified using single discrimination rules as they result from e.g. linear or polynomial svms or linear or quadratic discriminant analysis. Contrariwise, the right plot shows data where a combination of two rules is necessary to obtain a good classification model.

Proposed classification methods in literature that perform local modelling may be split into several groups according to their way of performing the local modelling task. For some of the methods the classes are assumed to be composed of several subclasses each. A first distinction might be whether these subclasses have to be given in the training data or not.

Section \textsuperscript{2} describes different approaches to local modelling: methods that assume the subclasses to be specified (section \textsuperscript{2.1}) as well as methods that perform an implicit subclass detection (section \textsuperscript{2.2}). Furthermore, in section \textsuperscript{2.3} some common classification methods are mentioned that implicitly perform some sort of local modelling.

In section \textsuperscript{3} all methods are evaluated. Four real-world data sets are introduced in this paper. Methods to evaluate the performance of the different methods are presented in section \textsuperscript{3.2} a statistical test to judge significant dominance of some
Fig. 1. Examples of classification rules for different data sets. Left: rule generated by a linear svm, middle: rule resulting from quadratic discriminant analysis, right: example of data where one single rule is not sufficient for discrimination.

of the methods as well as a method to derive a consensus ranking of them as a combination from their results on the four different data sets. The results of the study are shown in section 3.3.

2 Local Modelling in Classification

2.1 Known Subclasses in the Training Data

Combining local hypotheses
To motivate the task of combining several local hypotheses in case of known subclasses of the training set let us introduce some first example:

Example 1: music data
The classification problem concerning the music data set consists in register classification by means of timbre characteristics, i.e. without any information about the underlying fundamental frequency (for details see e.g. Weihs et al., 2006). The variables are formed by masses and widths of the fundamental frequency and the first twelve harmonics (see Figure 2), summing up to 26 variables in total. The data set consists of 432 tones (= observations) played by 9 different instruments/voices. The goal was to predict the correct register (i.e. high or low). Using the well-known linear discriminant analysis (Fisher, 1936) which is known to perform well and robust under a lot of data situations (see e.g. Hastie et al. 2001, p.89 or Michie et al., 1994) resulted in a (cross validated) unsatisfyingly high error rate of 0.352.

The idea of local modelling may result here in building local classification models for each instrument (denoted by \( l \)) separately. We can consider the population to be the union \( \Omega = \bigcup_{l=1}^{L} \Omega_l \) of subpopulations. The problem consists in prediction of a new observation if the instrument (and thus the choice of the local model) is not known. The resulting task can be formulated as some globalization of local classification rules.