Atlas of Science
another view on science
http://atlasofscience.org

Classification based on quantiles

Supervised classification is about finding formal rules to classify observations into one of two or several known classes based on training data for which the classes are known. The observations are characterised by several variables or features. Classification tasks are ubiquitous; examples comprise medical diagnosis based on various measurements, object recognition in images, credit scoring, and the detection of chemical particles based on voltages measured on the various electrodes of a detector (such data are analysed as a real data example in the paper). Our method is for general classification tasks with potentially a large number of quantitative variables.

Hall et al. (2009) proposed the median classifier, a very simple classification rule according to which the variable-wise distances of an observation to the within-class medians are summed up for a new observation, which is then classified to the class for which this summed distance is minimum.

The median classifier treats the data as if the distributions of all variables are symmetric, which in practice is not often the case. We generalized this classifier by using a specific quantile (percentage point) of a distribution instead of the median, which is the 50%-quantile. Whereas the median lies in the middle of a distribution (i.e., half of the distribution is to its left and half to its right), general quantiles may lie elsewhere, and therefore we introduced a way to weight differently distances of an observation to a quantile depending on whether an observation is smaller or larger than the quantile. The quantile on which the quantile classifier is found by minimizing the classification error in the training sample. This produces a classifier that is almost as good as the median classifier in terms of misclassification rates on test data not from the training sample if all variables are symmetrically distributed, but often much better if this is not the case.

We derived some theory that makes sure that the classifier works well for large enough samples, and we carried out a comprehensive simulation study in which the quantile-based classifier outperforms many up to data classification methods.

There is a free software package quantileDA, developed by Cinzia Viroli, which can compute the quantile classifier, and which is an add-on to the statistical software system R. The work summarised here is published as Hennig and Viroli (2016).

Christian Hennig [1]***, Cinzia Viroli*** [2]
[1]*Department of Statistical Science, University College London, United Kingdom*
[2]*Department of Statistical Sciences, University of Bologna, Italy*

Publication

Quantile-based classifiers
Hennig C, Viroli C
Biometrika. 2016 Jun