Classification based on data depth

Richard Hoberg
Karl Mosler
Universität zu Köln, Seminar für Wirtschafts- und Sozialstatistik
Albertus-Magnus-Platz
50923 Köln, Germany
mosler@statistik.uni-koeln.de

Consider a multivariate data set C that is partioned into given classes $C_1, \ldots, C_k$. An additional data point y has to be assigned to the class to which, in some sense, it 'fits best'. In other words, a new 'object' is assigned to one of several given classes of 'objects'.

To solve this *classification problem*, many rules have been proposed in the literature and successfully used in applications (e.g. Hand, 1997). They differ in their notion of 'best fit' and in the structure imposed on the data. This paper introduces a new approach to the construction of classification rules, which is based on notions of data depth.

A *data depth* is a real-valued function $y \mapsto d(y|S), y \in \mathbb{R}^d$, that measures the centrality of a point y with respect to a given finite set S. Points lying close to the 'center' of the set have larger depth values than points that lie more outside, and the center has maximum depth. For formal definitions of data depth, see Dyckerhoff (2002), Zuo and Serfling (2000), and Liu (1992).

Any depth function d provides a *depth classification rule*

$$classd(y) = argmax_j d(y|C_j)\,,$$

which assigns y to that class C_j in which y is deepest. Denote $C_j = \{x_{j1}, \ldots, x_{j_j}\}$. Especially, with the *Euclidean depth*

$$d_{Euc}(y|C_j) = \frac{1}{1 + ||y - \overline{x}_j||^2}\,, \quad \text{where } \overline{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ji}\,,$$

the classic *centroid classification rule* is obtained. Alternatively, the *Mahalanobis depth*

$$d_{Mah}(y|C_j) = \frac{1}{1 + (y - \overline{x}_j)'\Sigma_j^{-1}(y - \overline{x}_j)}$$

with Σ_j denoting the covariance matrix of group C_j, yields the well known *Mahalanobis classification rule*. While with the Euclidean depth the points of equal given depth form a sphere around $\overline{x}_j$, with the Mahalanobis depth they form the border of an ellipsoid. Thus, the distance from class C_j is measured in an elliptically symmetric way, which appears to be a natural distance if the data in the class have an elliptically symmetric distribution, like normal data, but not if they are distributed in an asymmetric way.

In our paper we develop classification rules that are based on other notions of data depth and reflect eventual asymmetries of the data. For example with the *zonoid depth* d_Z, the points y of depth $d_Z(y|C_j) \geq \alpha$ build a convex set $D_Z(\alpha)$,

$$D_Z(\alpha) = conv\left\{y : y = \sum_{i=1}^{n_j} \lambda_i x_{ji},\ 0 \leq \lambda_i,\ \alpha\lambda_i \leq \frac{1}{n},\ \sum_{i=1}^{n_j} \lambda_i = 1\right\},$$

if $\alpha > 0$. Other special notions of depth are the halfspace, simplicial, and convex-hull peeling depths, among others. All these depths are affine invariant. They differ in their measure

their practical computability; see Mosler (2002), Hoberg (2003). The halfspace, simplicial and peeling depths are also combinatorial invariant, while the Mahalanobis and the zonoid depths are not combinatorial invariant and refer to the metric structure of the data. The latter two show important continuity properties. As the Mahalanobis depth involves just the mean vector and the covariance matrix, it is very easy to calculate. The computational complexity of the combinatorial invariant depths impedes – or even prevents – their application in dimensions $d \geq 4$, but the zonoid depth of a point can be efficiently calculated also in higher dimensions.

One drawback of most depths is that they vanish outside the convex hull of the set C_j. By this, a point y lying outside the convex hulls of all classes cannot be classified by the depth. The paper introduces a new depth to be used for classification, the *zonoid-Mahalanobis depth*

$$d_{ZMah}(y|C_j) = \max\left\{ d_Z(y|C_j), \frac{1}{\max_j n_j} d_{Mah}(y|C_j) \right\} .$$

This maximum, again, is a depth and everywhere positive. It equals the zonoid depth inside the convex hull of C_j and is a multiple of the Mahalanobis depth outside. Thus it extends the zonoid depth beyond the convex hull. Classification by this depth is called the *zonoid-Mahalanobis rule.*

This new classification rule is applied to several small benchmark data sets from the literature and compared with known classification rules such as nearest neighbour, hypervolume, histogram and other rules. The classification rules are evaluated with respect to their *apparent error rates* as well as to their *leave-one-out error rates* (Hand, 1997). In the result, the zonoid-Mahalanobis classification rule comes out to be a good alternative to the existing rules, provided the convex hulls of the groups do not intersect. Compared with the commonly employed rules based on density estimators, the new rule avoids the – often problematic – choice of bandwidth.

REFERENCES

Dyckerhoff, R. (2002) "Inference based on data depth". Chapter 5 in K. Mosler (2002).

Hand, D.J. (1997) *Construction and Assessment of Classification Rules.* John Wiley, Chichester.

Hoberg, R. (2003) *Clusteranalyse, Klassifikation und Datentiefe.* Josef Eul, Lohmar.

Liu, R. (1992) "Data depth and multivariate rank tests". In Y. Dodge: L_1*-Statistics Analysis and Related Methods,* 279–294. North Holland, Amsterdam.

Mosler, K. (2002) *Multivariate Dispersion, Central Regions and Data Depth: The Lift Zonoid Approach.* Springer, New York.

Zuo, Y., Serfling, R. (2000) "General notions of statistical depth functions", *Annals of Statistics*, **28**, 461–482.

RÉSUMÉ

Une approche nouvelle est proposée du problème de classification. Cette approche se fonde sur la notion de profondeur, en particulier la profondeur derivée de 'lift zonoid'. Nous comparons notre méthode a quelques autres règles populaires de classification.