`kernel.SDA (symbolicDA)`

Kernel discriminant analysis for symbolic data

In most real-data discrimination tasks we can't assume anything about density function. We have to estimate such a function by:

a) approximating unknown density by applying one of known densities,

b) applying one of 12 functions proposed by Pearson as the estimator and solving a integral equation,

c) estimating the unknown density by applying kernel estimators.

The general form of kernel density estimator can be defined as follows (see Hand, Mannila and Smyth [2001], p. 170; Härdle and Simar [2003], p. 27):

$$\hat{f}_k(A_i) = \frac{1}{n_k (2h_k)^S} \sum_{i=1}^{n_k} K\left(\frac{A_i - A_{jk}}{h_k}\right), \quad A_i \in R^S \tag{1}$$

where: $\hat{f}_k(A_i)$ – kernel density estimator for *i*-th object and *k*-th cluster; $k = 1,\ldots,g$ – cluster number; A_{jk} – *j*-th object from *k*-th cluster; S – dimension; $i = 1,\ldots,n_k$ – number of objects in *k*-th cluster; h_k – bandwidth parameter; $K(\bullet)$ – uniform kernel.

In case of symbolic data we can't apply the well-known kernel density estimator, due to the fact for these object integral operator can't be defined and symbolic data space is not a Euclidean subspace too. Instead of density kernel estimator the kernel intensity estimator is applied (see Bock and Diday [2000], p. 242):

$$\hat{I}_k(A_i) = \frac{1}{n_k} \sum_{i=1}^{n_k} \prod_{l=1}^{b} K_{A_i,h_l}(A_{jk}), \tag{2}$$

where: $\hat{I}_k(A_i)$ – kernel intensity estimator for *i*-th object and *k*-th cluster; $i = 1,\ldots,n_k$ – number of objects in *k*-th cluster; $k = 1,\ldots,g$ – cluster number; $l = 1,\ldots,b$ – number of distance measures applied; A_{jk} – *j*-th object from *k*-th cluster; h_l – bandwidth parameter for *l*-th distance measure; $K_{A_i,h_l}(A_{jk})$ – uniform kernel based on *l*-th distance measure for *i*-th symbolic object and *j*-th symbolic object from *k*-th cluster.

For symbolic data uniform kernel is defined as (Bock and Diday [2000], p. 242):

$$K_{A_i,h_l}(A_{jk}) = \begin{cases} 1 & \text{for } d_{ij} < h \\ 0 & \text{for } d_{ij} \geq h \end{cases}, \tag{3}$$

where: d_{ij} – distance measure for *i*-th and *j*-th symbolic object; h – bandwidth parameter.

Calculation of posterior probabilities requires to determine prior probabilities for each cluster. The prior probabilities can be (Bock and Diday [2000], p, 242-243):

a) equal for each cluster: $\hat{p}_k(A_i) = \frac{1}{g}$, where g – number of clusters,

b) dependent on the number of the objects in the cluster: $\hat{p}_k(A_i) = \frac{n_k}{n}$, where n_k – number of objects in k-th cluster; n – total number of objects in the dataset,

c) calculated as:

$$\hat{p}_k(t+1) = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{\hat{p}_k(t) \hat{I}_k(A_i)}{\sum_{k=1}^{g} \hat{p}_k(t) \hat{I}_k(A_i)} \right), \tag{4}$$

where: $k = 1,\ldots,g$ – cluster number; n – number of objects; t – t-th iteration step; $\hat{p}_k(0) = \frac{1}{k}$ – probability at the starting point of the algorithm; $\hat{I}_k(A_i)$ – intensity estimators for i-th object and k-th cluster that are constant.

Bock and Diday [2000], p. 241 suggest that ten iteration steps are enough to determine prior probabilities.

Posterior probabilities are calculated as (Bock and Diday [2000], p. 244):

$$q_k(A_i) = \frac{\hat{p}_k \hat{I}_k(A_i)}{\sum_{k=1}^{g} \hat{p}_k \hat{I}_k(A_i)}, \tag{5}$$

where: $k = 1,\ldots,g$ – cluster number; $q_k(A_i)$ – posterior probability for i-th symbolic object and k-th cluster; $\hat{p}_k$ – prior probabilities; $\hat{I}_k(A_i)$ – intensity estymator for i-th symbolic object and k-th cluster.

References:

1. Billard L., Diday E. (Eds.) (2006), *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley & Sons Ltd, Chichester.
2. Bock H.-H., Diday E. (Eds.) (2000), *Analysis of symbolic data. Explanatory methods for extracting statistical information from complex data*. Springer Verlag, Berlin-Heidelberg.
3. Hand D., Mannila H., Smyth P. (2001), *Principles of data mining*, MIT Press, Cambridge.
4. Härdle W., Simar L. (2003), *Applied multivariate data analysis*, Springer-Verlag, Berlin-Heidelberg.