Stat Comput (2011) 21:295–308
DOI 10.1007/s11222-010-9169-0

Density-based Silhouette diagnostics for clustering methods

Giovanna Menardi

Received: 23 February 2009 / Accepted: 6 January 2010 / Published online: 4 February 2010

Abstract Silhouette information evaluates the quality of the partition detected by a clustering technique. Since it is based on a measure of distance between the clustered observations, its standard formulation is not adequate when a density-based clustering technique is used. In this work we propose a suitable modification of the Silhouette information aimed at evaluating the quality of clusters in a density-based framework. It is based on the estimation of the data posterior probabilities of belonging to the clusters and may be used to measure our confidence about data allocation to the clusters as well as to choose the best partition among different ones.

Keywords Cluster analysis · Density estimation · Diagnostics · Silhouette information

1 Introduction

Cluster analysis refers to a widespread class of methods for exploring data with the aim of finding groups of objects that are similar to each other but different from the objects in other groups. These goals of similarity within groups and dissimilarity between groups are typically achieved according to a traditional approach based on some measures of distance (dissimilarity) or, alternatively, by evaluating the density underlying the data.

Unlike supervised techniques of data mining, when dealing with a clustering problem, there is no prior information about the existence of interesting partitions of the data in groups. Therefore, it is not known to what extent the quality of the clustering is due to the real structure of the data or to the performance of the clustering technique adopted. Hence, a tool for evaluating the quality of the results can be useful to assess the ability of the clustering procedure in finding partitions or to choose the best partition.

Several methods have addressed this issue in the past, many of them based on some measure of distance between objects and clusters. One exploratory tool hinging on this idea is the *Silhouette information* (Rousseeuw 1987). However, it can be argued that the diagnostics used for evaluating the goodness of a partition should be consistent with the clustering method adopted to produce that partition and, in particular, that distance-based diagnostics are inadequate to assess the groups identified by using a technique based on density estimation.

In this work we propose a suitable modification of the Silhouette index aimed at appraising the clustering quality in a density-based framework.

The rest of the paper is organized as follows: in Sect. 2 a review of clustering methods based on density measures is presented. Section 3 discusses some diagnostics for evaluating the quality of clusters and focuses on Silhouette information specifically. Section 4 introduces the proposed modification. Its diagnostic ability is evaluated on some simulated and real data in Sect. 5. Some concluding remarks are reported in Sect. 6.

2 Density-based clustering methods

Among the many techniques for cluster analysis, a quite recent approach expresses the concept of intra-group similarity and inter-group dissimilarity in terms of the density of data. This approach goes back to Hartigan's (1975, p. 205) definition of clusters, which *may be thought of as regions of*

G. Menardi (✉)
Department of Economics and Statistics, University of Trieste, P.le Europa, 1, Trieste, Italy
e-mail: giovanna.menardi@econ.units.it

high density separated from other such regions by regions of low density, but it has been developed only recently with current computational advances.

This idea has an explicitly inferential motivation: the observed data $\mathcal{X} = (x_1, \ldots, x_i, \ldots, x_n)'$, $x_i \in \Re^d$, are, in fact, supposed to be a sample of independent and identically distributed realizations of a $d-$dimensional random vector with an unknown probability density function f. Density estimation allows for the detection of high density regions (empirical clusters) which approximate the population clusters.

Two main classes of density-based methods arise in the statistical literature about cluster analysis. The model-based approach (see, for a review, McLachlan and Peel 1998; Fraley and Raftery 2002) rests on the idea that each cluster $\mathcal{U}_m$, $m = 1, \ldots, M$ corresponds to a subpopulation f_m typically belonging to some parametric family. The overall population f is then modeled as a finite mixture of these subpopulations:

$$f(x) = \sum_{m=1}^{M} \pi_m f_m(x),$$

where $f_m(\cdot)$ is the density of the mth component of the mixture, corresponding to the group $\mathcal{U}_m$ and usually depending on some parameter vector θ_m and π_m is the mixing proportion of that component. Estimation of θ_m is usually carried out by the expectation-maximization (EM) algorithm which determines the maximum likelihood estimate of the mixture model parameters. The maximization step of EM allocates the data to the clusters according to the Bayes rule for classification, namely the observation x is classified to the cluster $\mathcal{U}_{m_0}$ if

$$\hat{\pi}_{m_0} \hat{f}_{m_0}(x) > \hat{\pi}_m \hat{f}_m(x), \quad m \neq m_0.$$

A second class of clustering methods, which is more directly related to the Hartigan definition, links the clusters with the high density level sets. Basically, any section of the density f underlying data, at level k, induces a partition of the sample space into two sets, one having density less than k, one having density greater than k. The clusters correspond to the maximum connected regions of the latter set. As k varies, these clusters may be represented according to a hierarchical structure in the form of a tree. Density estimation is usually performed by a nonparametric method and allows for the detection of high density regions. However, the associated connected regions are not, in general, explicitly defined.

The clustering methods of this class mainly differ from each other in the way of detecting such regions. For instance, in Cuevas et al. (2001), the identification of the connected sets is pursued by a technique based on the smoothed bootstrap data resampling. Essentially, the population clusters are approximated by the union of closed balls centered at the sampled data with estimated density above a fixed threshold k. Such a constant depends on some user-defined related parameters. Stuetzle (2003) takes advantage of a link between the minimum spanning tree and the nearest neighbor density estimate of data, that allows for an easy detection of the connected components of the level sets. Azzalini and Torelli (2007) approximate the groups with the polyhedrons formed by applying a Delaunay triangulation on the data with estimated density greater than k. The remaining data, with lower density, are assigned to the clusters by following a logic typical of supervised classification. A notable advantage of the procedure is that the number of clusters is automatically selected.

It is worthwhile noting that the two classes of methods differ not only because of their approach to density estimation (parametric and, respectively, nonparametric) but the definition of a cluster is also conceptually different in the two approaches. While the nonparametric methods associate the clusters to the regions around the modes of the probability distribution of data, clusters in the model-based approach correspond to the components of a mixture of distributions. Since the number of the modes in a mixture of distributions does not necessarily match the number of components, the difference between the two approaches emerges quite clearly.

The idea of defining groups as regions associated with the high density connected components is also the cornerstone on which several methods, proposed by the machine learning community, rest. However, it would be more appropriate to place these methods halfway between the density-based and the distance-based clustering procedures. In fact the data lie on a metric space and the density of the data is usually expressed as a function of the distance between the objects. DBSCAN (Ester et al. 1996) is one of the most representative procedures of this class. Here the concepts of both density and connectivity are based on the local distribution of the data's nearest neighbors. GDBSCAN (Sander et al. 1998) is a suitable generalization of DBSCAN which allows for clustering point objects as well as spatially extended objects, according to their spatial and non-spatial attributes. An attempt to overcome some arbitrariness in the choice of the input parameters inherent to these methods is given in Ankerst et al. (1999).

3 Evaluating the quality of clusters

Several recent and earlier methods have addressed the issue of evaluating the partition produced by a clustering method, in order to give a measure of its quality, choose the best partition among different ones and select the optimal number of clusters.

Classical distance-based indexes have been proposed by Dunn (1974) as well as by Davies and Bouldin (1979). Both the indexes compare, for each cluster, a measure of compactness given by an average distance between the objects in the cluster and the cluster centroid, with a measure of separation from the other clusters (distance between the centroids). Hubert and Schultz (1976) provide a general means for measuring the association between two proximity matrices. More recent distance-based cluster validity indexes have been proposed by Xie and Beni (1991), Bezdek and Pal (1995), Maulik and Bandyopadhyay (2002).

Other methods are based on probabilistic schemes, such as likelihood ratio tests (Duda and Hart 1973) or information-based criteria (Cutler and Windham 1994). Bayesian inference provides an alternative to likelihood ratio tests for the number of groups in a model-based clustering, both for normal mixtures and other types of distributions (Binder 1978, 1981; Banfield and Raftery 1993; Bensmail et al. 1997).

A traditional distance-based exploratory method aimed at appraising the quality of clusters is the so called *Silhouette information* (Rousseeuw 1987). The idea on which Silhouette information hinges, arises from the comparison of a measure of closeness of each observation to the cluster where it has been allocated and a measure of separation from the closest alternative cluster.

Let $\mathcal{X} = (x_1, \ldots, x_n)'$ be the matrix of the observations, $x_i' \in \Re^d$, and $\mathcal{U}_1, \ldots, \mathcal{U}_M$ a partition of $\mathcal{X}$ produced by a clustering procedure. For each x_i one computes $\overline{d}(x_i; \mathcal{U}_m)$, the average distance between x_i and the elements of $\mathcal{X}$ belonging to $\mathcal{U}_m, m \in 1, \ldots, M$. Moreover, let us suppose that the clustering procedure has assigned x_i to the cluster $\mathcal{U}_{m_0}$ and that $\mathcal{U}_{m_1}$ is the cluster which minimizes the average distance $\overline{d}(x_i; \mathcal{U}_m), m \neq m_0$.

The Silhouette Information for the elements x_i is:

$$s_i = \frac{\overline{d}(x_i, \mathcal{U}_{m_1}) - \overline{d}(x_i, \mathcal{U}_{m_0})}{\max\{\overline{d}(x_i, \mathcal{U}_{m_1}), \overline{d}(x_i, \mathcal{U}_{m_0})\}}. \tag{1}$$

Observations with a large s_i (near 1) are supposed to be well clustered, a small s_i (near 0) means that the observation lies between two clusters, and observations with a negative s_i are probably placed in the wrong cluster. The clustering structure can be displayed after splitting the observations between the groups, sorted according to Silhouette information.

An average s provides a global measure of quality of clusters and allows for the comparison of partitions (the partition with maximum average Silhouette is taken as the optimal one). See Kaufman and Rousseeuw (1990) for details.

4 Cluster validation in a density-based framework

4.1 A density-based Silhouette information

Density-based clustering methods include natural information about the degree of confidence we give to the cluster membership of the observations. High density data points are given maximum confidence because they lie just around the modes of the density function. In contrast, a lower confidence is given to the data points which lie on the tails or at the valleys of the density function. This feature may be developed in order to produce a measure for evaluating the quality of the partition detected by the clustering procedure. In particular, we propose an adaptation of the Silhouette information suitable for density-based clustering procedures. The main difference between the existing Silhouette information and the new tool is that the former is based on the distance between clusters, whereas the latter is built in a density-based framework.

Recalling the notation introduced in Sect. 3, since $x_i \in \mathcal{X}$ is drawn from a probability density function f, one can evaluate the posterior probability that it belongs to group $\mathcal{U}_m, m = 1, \ldots, M$, as:

$$\tau_m(x_i) = \frac{\pi_m f_m(x_i)}{\sum_{m=1}^{M} \pi_m f_m(x_i)}, \tag{2}$$

where π_m is a prior probability of $\mathcal{U}_m$ and f_m is the density of group $\mathcal{U}_m$ at x_i.

The *density-based Silhouette information* (*dbs*) of x_i is then defined as follows:

$$dbs_i = \frac{\log(\frac{\tau_{m_0}(x_i)}{\tau_{m_1}(x_i)})}{\max_{j=1,\ldots,n} |\log(\frac{\tau_{m_0}(x_j)}{\tau_{m_1}(x_j)})|}, \tag{3}$$

where m_0 is such that x_i has been classified to $\mathcal{U}_{m_0}$ and m_1 is the group index for which τ_m is maximum, $m \neq m_0$.

The normalization factor in (3) does not correspond exactly to its counterpart in (1) because the maximum is taken with respect to the observations (instead of the two competing groups). This is a discretionary choice but some preliminary analysis has shown better results using the formulation above (see the next section for further details).

The density-based Silhouette information of x_i is, therefore, proportional to the log ratio between the posterior probability that it belongs to the group to which it has been allocated and the maximum posterior probability that it belongs to another group. Large values of *dbs* are evidence of a well clustered data point while small values of *dbs* mean a low confidence in the classification. Negative values of *dbs* are possible and occur when $\hat{\tau}_{m_0}(x_i) < \hat{\tau}_{m_1}(x_i)$ that is, the observation x_i allocated to the cluster $\mathcal{U}_{m_0}$ has an higher posterior probability of belonging to a different cluster. Hence,

a negative value of *dbs* is usually evidence of an incorrect allocation of the observation.

After evaluating the *dbs* index for all the observations, they are partitioned into the clusters, sorted in a decreasing order with respect to *dbs* and displayed on a bar graph (*density-based Silhouette plot*). A location index could then be calculated to obtain summarizing information about the quality of the clusters.

4.2 Computation of *dbs*

The practical evaluation of the described diagnostic index requires the specification of both the density of data and the prior probabilities of groups. Since the distribution underlying the data is not known, the empirical *dbs* is used:

$$\hat{dbs}_i = \frac{\log(\frac{\hat{\tau}_{m_0}(x_i)}{\hat{\tau}_{m_1}(x_i)})}{\max_{j=1,\ldots,n}|\log(\frac{\hat{\tau}_{m_0}(x_j)}{\hat{\tau}_{m_1}(x_j)})|}, \tag{4}$$

where

$$\hat{\tau}_m(x_i) = \frac{\pi_m \hat{f}_m(x_i)}{\sum_{m=1}^{M} \pi_m \hat{f}_m(x_i)}, \quad m = 1, \ldots, M, \tag{5}$$

and $\hat{f}_m(x_i)$ is a density estimate at x_i obtained, after clustering the data, by using only the data points in $\mathcal{U}_m$.

The illustrated procedure is not linked to any specific technique of density estimation and, provided that a density estimator with good properties is used, parametric models as well as nonparametric ones can be chosen. Among the many possible choices, in the subsequent examples a kernel estimator with Gaussian kernel and diagonal smoothing matrix $h = (h_1, \ldots, h_d)$ has been used. In order to reduce the computational effort, the diagonal smoothing parameters have been selected as asymptotically optimal for estimating a normal density function. This approach usually tends to induce oversmoothing when applied to non-normal data. However, we are confident that our choice looks well advised because the f_m are at least unimodal by definition.

From a computational point of view, the use of only the data points allocated to $\mathcal{U}_m$ to estimate the f_m has the effect of pushing the clusters apart, with two main consequences. First, most x_i values will have $\tau_{m_0}(x_i)$ close to one. Hence, normalizing the density-based Silhouette by using the exact counterpart of (1) would result in most *dbs* values close to one. Instead, better performance derives from the use of (3), because the difference between the dbs_i emerges more clearly. A further consequence is that the observations lying at the valley of the density underlying the data have a higher chance of getting a negative value of the *dbs*. This effect turns out to be desirable because the confidence we give to the cluster membership decreases as we move away from one mode of the density to another mode.

With regard to the specification of π_m, the choice depends on the prior knowledge about the composition of the clusters and a lack of information would imply the choice of a uniform distribution of the π_m over the groups. However, information derived from the detected partition can also be used. In a model-based clustering, for example, the mixing proportions would seem to be a natural choice. When using the AT procedure (Azzalini and Torelli 2007) a first detection of some *cluster cores* is performed, and prior probabilities can be chosen as proportional to the cardinalities of the cluster cores.

It is worth noting that a further connection between the original Silhouette information and its density-based version exists. If $f_m(\cdot)$ is chosen as proportional to $\exp(-\lambda d(\cdot, \theta_m))$, where $d(\cdot,\cdot)$ is a distance and θ_m, λ some location and scale parameters characterizing the groups (*i.e.* a distance-based model is used for estimating the cluster densities), it is easy to show that the numerator of the (3) becomes:

$$\log\left(\frac{\pi_{m_0}}{\pi_{m_1}}\right) + \lambda(d(x_i, \theta_{m_1}) - d(x_i, \theta_{m_0})).$$

Hence, if it is reasonable to assume a uniform distribution of the π_m over the groups, the *dbs* relates to the original distance-based version of the Silhouette index even more closely than the general case, being different only in the way that the distances are averaged. The former computes the distance between x_i and the average points of $\mathcal{U}_{m_0}, \mathcal{U}_{m_1}$, while the latter measures the average distances between x_i and the elements of $\mathcal{U}_{m_0}, \mathcal{U}_{m_1}$.

5 Validation of density-based Silhouette information

The ability of the density-based Silhouette information in evaluating the quality of a clustering and in helping to choose the best partition between different ones is assessed on some real and simulated data sets.

Since one considers a cluster analysis which produces meaningful groups as successful, some clustering procedures have been applied on data with a known clustering structure, with the purpose of reconstructing the original groups. Then, the detected clustering has been evaluated by using the density-based Silhouette information. The analysis has been carried out by applying the AT procedure based on nonparametric density estimation (Azzalini and Torelli 2007) and the MCLUST model-based clustering method (Fraley and Raftery 2006).

In order to understand if the density-based Silhouette information can be used also for choosing between different partitions, distinct clustering structures have been produced by each procedure for the considered data sets. This has been possible by suitably varying the input parameters of the clustering procedures. More specifically, MCLUST has been run

by setting, in turn, a different number of components (*i.e.* groups) in the mixture of densities. The AT procedure, instead, automatically determines the number of groups by counting the modes of the density estimate. Hence, partitions having different number of groups have been obtained, by alternatively scaling the bandwidth matrix which governs the smoothing of the density estimate so that different modal structures emerged.

The sensitivity of the *dbs* in detecting misclassified observations has been measured through a ROC analysis.

Moreover, the behaviour of the mean and median *dbs* has been analysed when the number of clusters varies, in order to evaluate the opportunity of using a location index as a summarizing *dbs* for measuring the global quality of the clustering.

Finally, we have also computed the (distance-based) Silhouette information on the clusters returned by the classical Ward clustering method and applied a ROC analysis to compare the diagnostic ability of the original and the density-based Silhouette.

We present here the results from the use of four data sets. The first data set was originally presented by Forina et al. (1983) and subsequently analyzed by various authors to illustrate classification and clustering techniques. See, for instance, Stuetzle (2003). The data represent eight chemical measurements on $n = 572$ specimens of olive oil produced in various areas of Italy: Centre-North, South, and Sardinia. The clustering algorithms have been applied to reconstruct the geographical origin of the oils. Some preliminary analysis has been conducted on this set of data, following Azzalini and Torelli (2007). Since the resulting data matrix had a dimensionality too large to be handled by the AT procedure, all the clustering methods have been applied on the first five principal components. To see the distribution of the clusters, we have displayed the first two principal components in Fig. 1. Despite the possibility that the reduced dimension could lead to misleading considerations, the figure suggests that the clustering methods have difficulty clustering the Sardinia group.

The second example data set describes 13 chemical characteristics of 178 wines grown in the same region in Italy but derived from three different cultivars (groups). Also this data set was introduced by Forina et al. (1986) and widely used for testing new classifiers or clustering methods (see, for example, Dy and Brodley 2004). All the clustering methods were applied on the first three principal components. Here, the clustering structure is more evident than the previous example, even when looking at the first two principal dimensions only (Fig. 2).

The choice of using PCA to reduce the data dimension might, admittedly, be risky, because there is no guarantee that the clustering structure is preserved in the reduced space (for further details see, for example, Chang 1983). However,

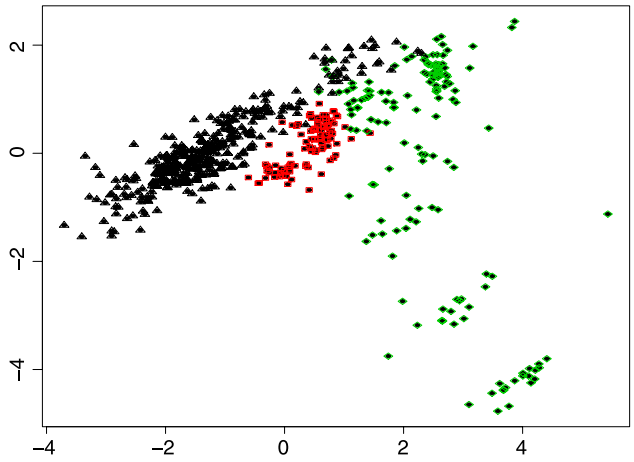

Fig. 1 Distribution of the clusters in the first two principal components of the olive oil data. *Symbols* indicate the geographic areas: *triangles* for the South, *rumbles* for the Center-North, *squares* for Sardinia

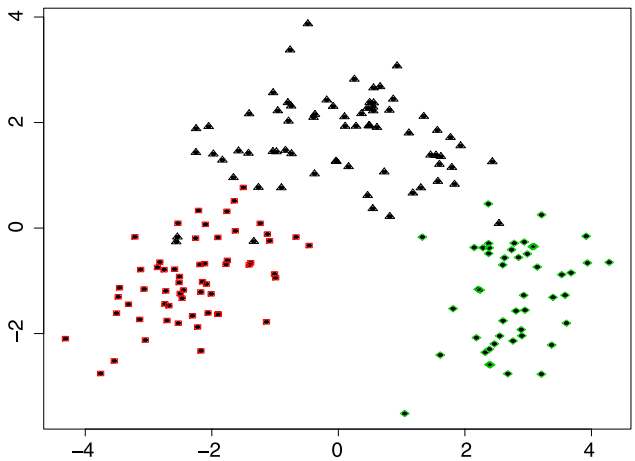

Fig. 2 Distribution of the clusters in the first two principal components of the wine data. *Symbols* indicate the different wine cultivars

the issue is beyond the scope of this paper and it is not addressed here. In fact, it should be stressed that the purpose of the analysis is not to appraise the quality of the partition generated by the clustering procedures but to test the ability of *dbs* in understanding such quality.

As a first synthetic example, a sample of $n = 100$ observations has been generated from the mixture of bivariate Gaussian distributions $1/2N(\mu_1, \Sigma) + 1/2N(\mu_2, \Sigma)$ having the same covariance matrix.[1]

The second artificial data set contains $n = 100$ realizations from a bivariate standard Normal distribution. It does not present any clustering structure and has been used to test the ability of the density-based Silhouette in recognizing the absence of groups.

[1] $\mu_1 = (1/2, -1)$, $\mu_2 = (-1/2, 1)$, $\Sigma = \begin{pmatrix} 0.4 & 1 \\ 1 & 3 \end{pmatrix}$.

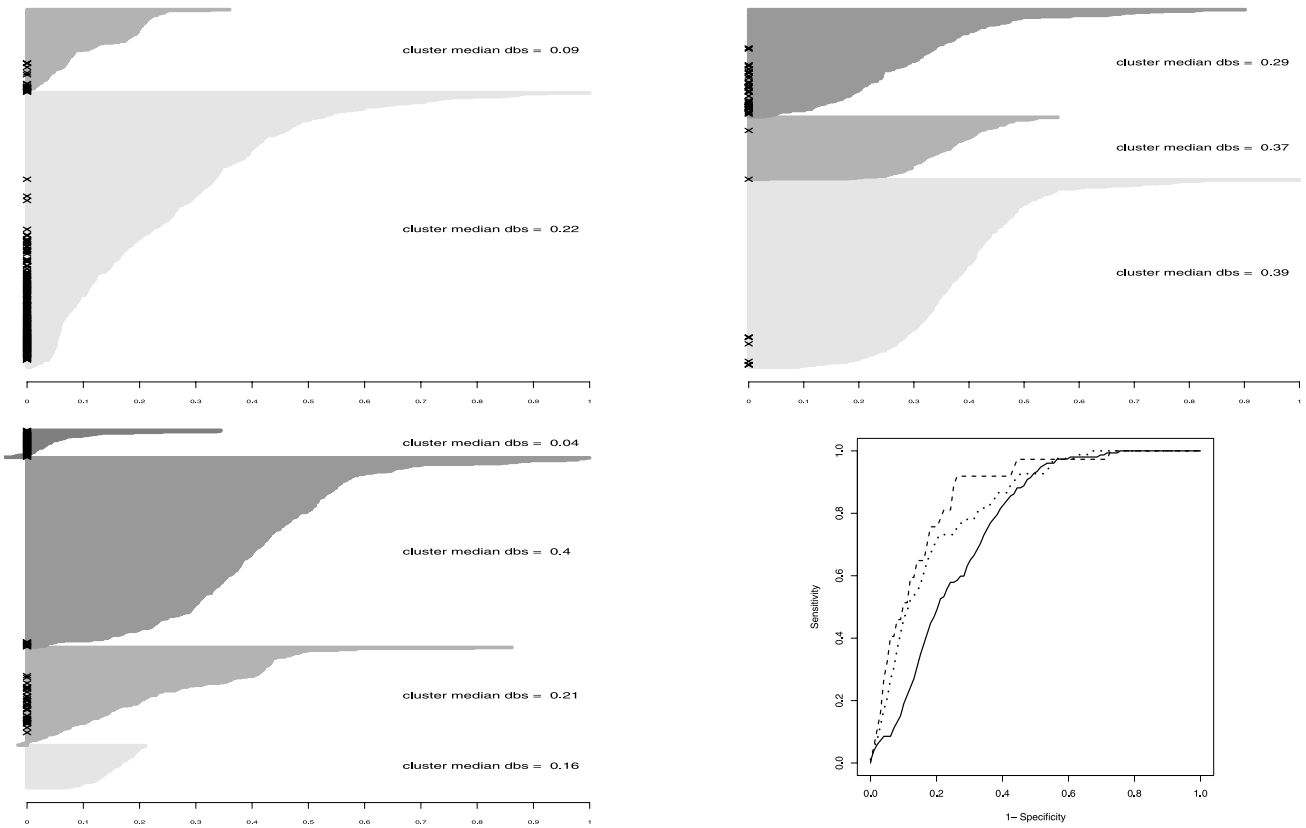

Fig. 3 *dbs* plots of three partitions produced by the AT method on the first five principal components of the olive oil data. The plots refer to differently user selected parameters of the clustering methods, leading to a distinct number of clusters. On each plot different blocks correspond to different clusters. The *black crosses* (normally not present in the plot) identify the misclassified units and help us in evaluating the diagnostic power of *dbs*. In the *first panel* the largest group mainly correspond to the olive oils from Sardinia and from the South of Italy. In the *top right plot*, the groups are associated to the Centre-North, Sardinia and Southern areas (from the *top to* the *bottom*). The three largest groups in the *dbs* plot on the *bottom* may be labeled as the South of Italy, the Center North and Sardinia. The *bottom right panel* displays the ROC curves corresponding to the three *dbs* plots: the *solid*, *dashed* and *dotted lines* refer to the partitions in two, three and four groups, respectively

In Figs. 3 to 10 some results are reported. On each figure the *dbs* plots corresponding to partitions of the data in two, three and four groups are displayed. In the bar plots we have highlighted the misclassified observations to better understand the diagnostic ability of the *dbs*. The bottom right panel of each figure compares the diagnostic abilities of the three partitions in terms of ROC curve, as will be explained in the next paragraphs.

Concerning the olive oil data, when AT and MCLUST are forced to return two groups (top left panel of Figs. 3 and 4) the Sardinia cluster is entirely assigned to the Southern area and to the Centre-North respectively. Additionally, AT exchanges a few labels from these two areas. However, misclassified data lie on the margins of the groups, corresponding to the valley of the density estimate, thus having the smallest *dbs*. The tripartition of the olive oil data leads to good performance of both the clustering methods (especially MCLUST) which correctly detect the olive oils from the Sardinia group, but misclassifies a few values from the remaining groups. Again, the misclassified labels have a small value of the *dbs*. The same behaviour emerges when four groups are formed by MCLUST. The additional group includes some data belonging to the Southern and Centre-Northern regions. Instead, when four groups are returned by AT, the additional group is formed by some Centre-Northern and Sardinian oils, while some Southern olive oils are incorrectly labeled as coming from the Centre-North area. The most notable feature of the *dbs* plots corresponding to four groups is that, when using both the AT and the MCLUST method, a very small median *dbs* in the fourth group is evidence of a wrong number of clusters.

With regard to the wine data set, when AT and MCLUST are forced to partition the data into the actual number of groups (top right panel of Figs. 5 and 6) the overall misclassification error is very low. The median *dbs* of each group is relatively large, still taking values close to zero when data are incorrectly labeled. Clustering the data into two groups mainly corresponds to partitioning one of the original clus-

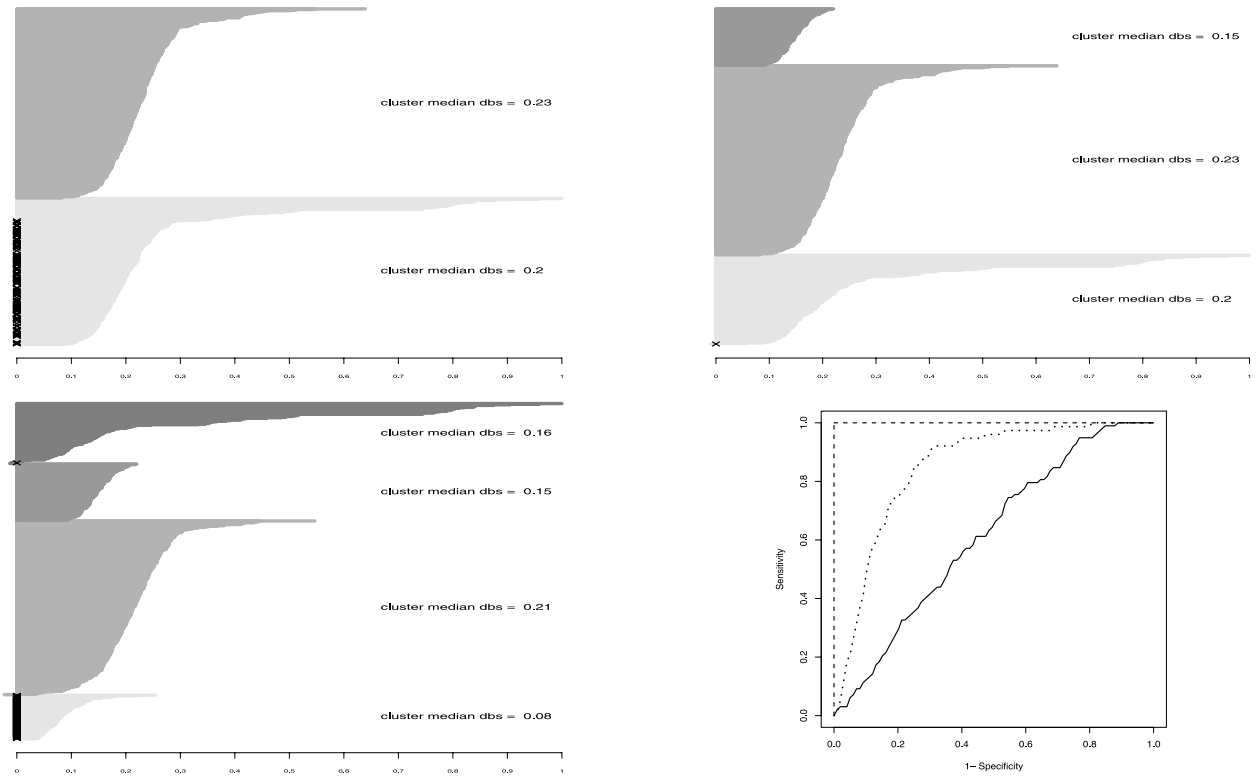

Fig. 4 Cf. Fig. 3. In this example the first five principal components of the olive oil data have been clustered by the MCLUST method. The *dbs* behaviour is optimum when three groups are found because the minimum *dbs* corresponds to the only one misclassified point

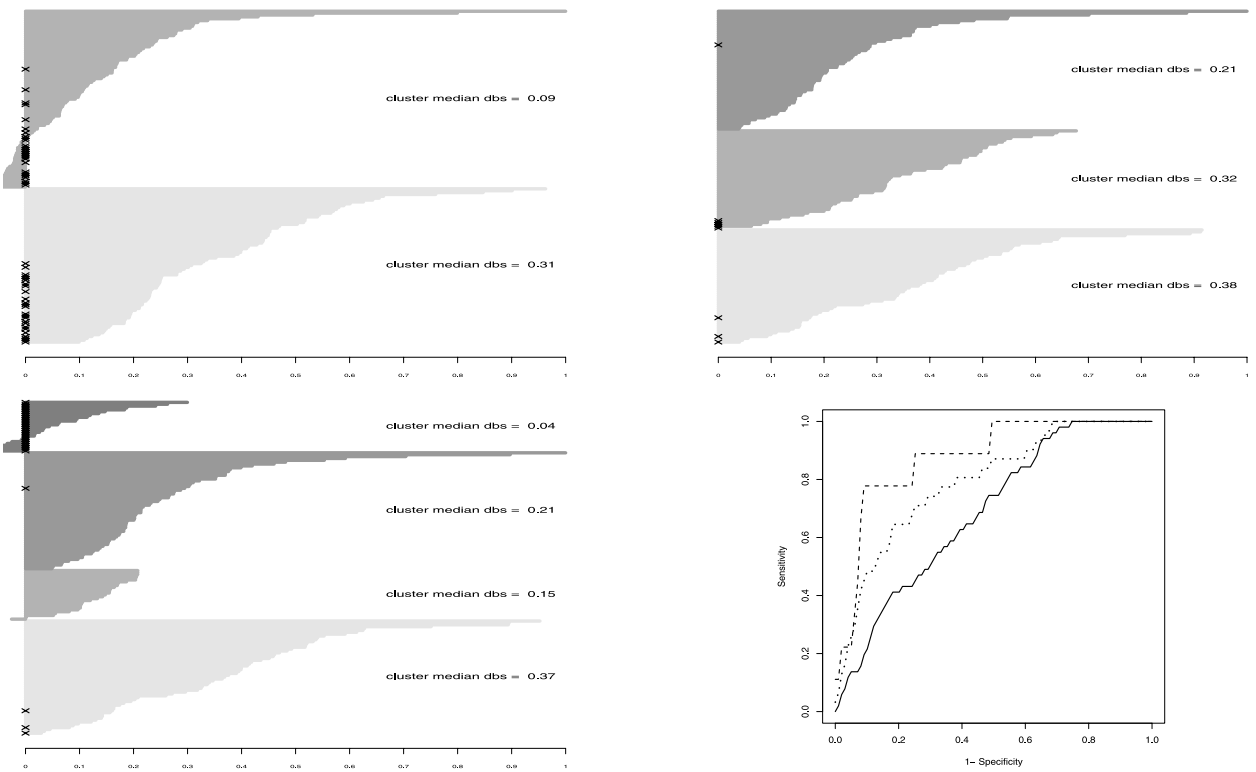

Fig. 5 Cf. Fig. 3. In this example the first three principal components of the wine data have been clustered by the AT method. Many negative values of the *dbs* in the bipartition and the close to zero value of one cluster's median *dbs* when four groups are found, suggest we should choose the top right clustering, corresponding to the actual structure of groups

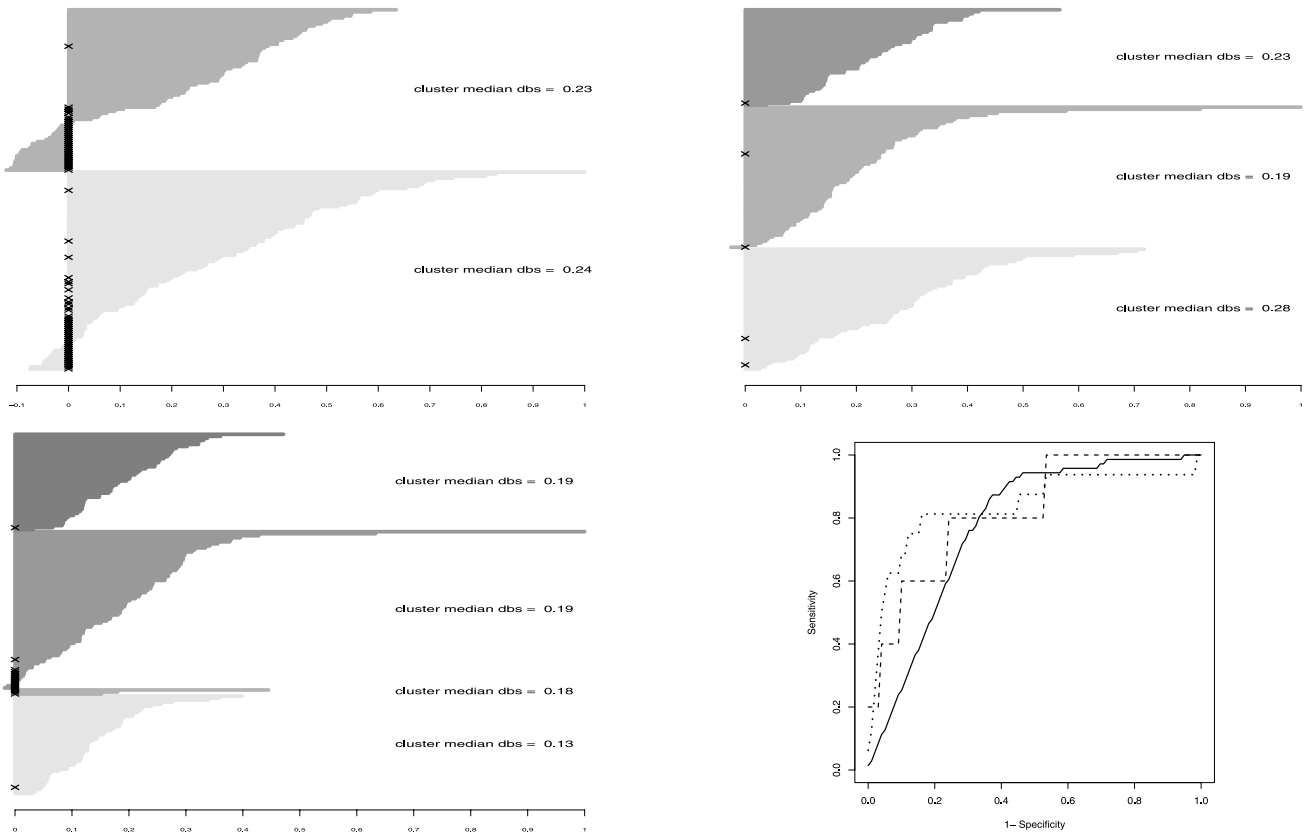

Fig. 6 Cf. Fig. 3. In this example the first three principal components of the wine data have been clustered by MCLUST. Good diagnostic abilities are shown by the *dbs* in all the partitions. When four groups are detected the small size of one group is evidence of a spurious cluster, even if the correspondent median *dbs* is not minimum

ters into the two detected ones. More specifically, MCLUST splits the data of the largest cultivar while AT splits the data of the smallest one. In both cases the incorrectly labelled data take a small or even a negative value of the *dbs* index. Good diagnostic abilities are shown by the *dbs* also when the data are partitioned into four groups. When using MCLUST the data assigned to the fourth group do not take the smallest values of the *dbs* but the small size of the group is evidence of a spurious cluster. Moreover, values of the *dbs* close to zero still correspond to misclassified data in the other groups. The fourth cluster detected by the AT procedure is generated by bipartitioning the third cultivar of the data set (corresponding to the smallest group) and it has the smallest median *dbs*.

The *dbs* behaviour on partitions of the simulated data sets (Figs. 7 to 10) shows the same tendency as the real data sets. The misclassified observations have a negative or a small value of the index and, on the other hand, negative and small values of the *dbs* correspond to incorrect labels. Some false negatives (misclassified data with large *dbs*) occur, but their presence is quite rare (see the ROC analysis below).

Moreover, the analysis suggests that clusters with a large median *dbs* correspond to a small misclassification error. Unlike the Silhouette plot which returns the mean *s* for each cluster, the median of the *dbs* values is used as a measure of a cluster's quality. Indeed, the mean is not the best representative of the overall cluster accuracy, due to the lack of robustness. See, for example, the bottom left panel of Figs. 3 and 10: the smallest clusters take a very close to zero (or even negative) value of the median *dbs*, which is consistent with their spuriousness. Instead, the mean *dbs* would be much higher. Further interesting features emerge from the observation of the density-based Silhouette on the data which do not exhibit any grouping structure (see Figs. 9 and 10). When the clustering methods are coerced to partition the data, a very small or negative *dbs* is given to the smallest groups, meaning a low confidence in the classification. Instead, a large group is formed, having the maximum *dbs* median. This behaviour, especially visible in the partitions returned by the AT method, is evidence of the absence of groups and reassures us about the diagnostic ability of the density-based Silhouette information even when partitioning data into homogeneous groups does not make sense.

These considerations can be confirmed through a ROC (Receiver Operating Characteristic) analysis (see, for example, Fawcett 2006). From a statistical point of view, ROC

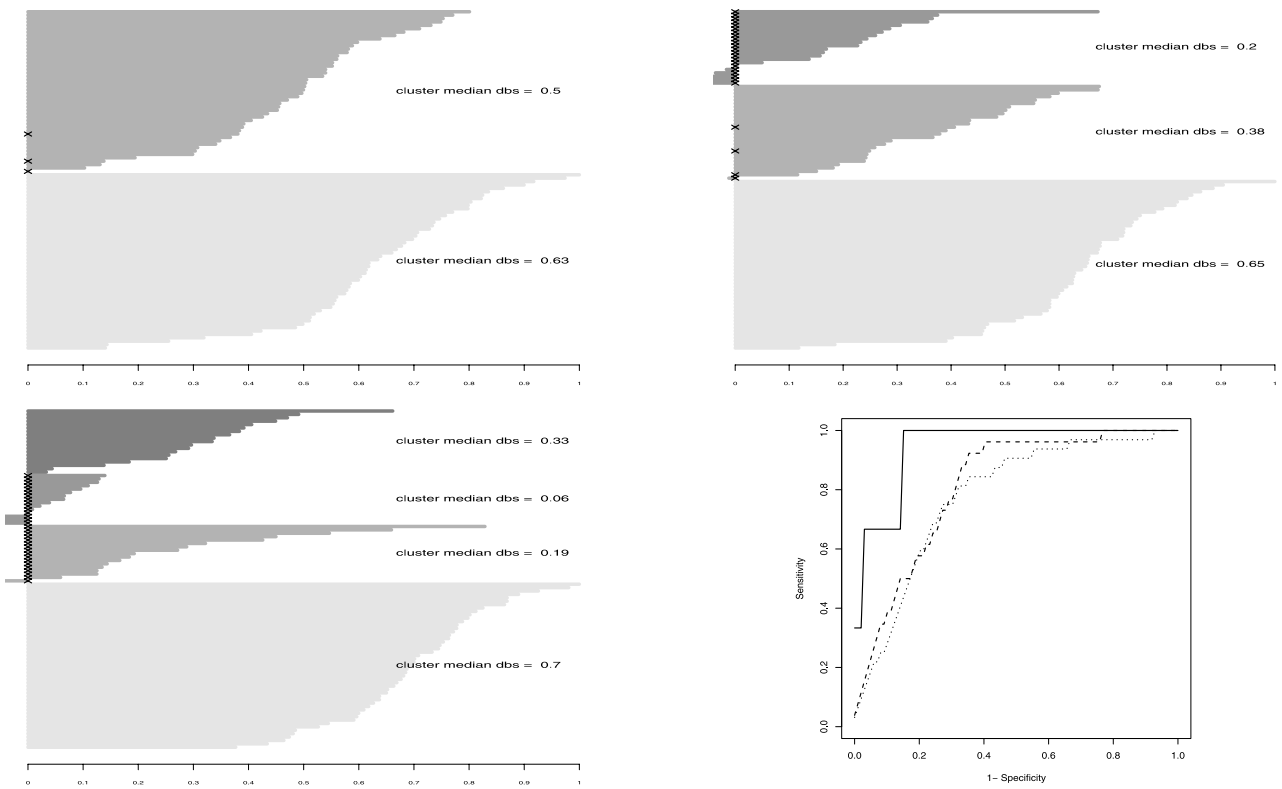

Fig. 7 Cf. Fig. 3. The first simulated data set has been clustered by AT. When the procedure is forced to return more than two groups, the additional clusters present several negative values of the *dbs* and a remarkably smaller median *dbs* than the actual groups

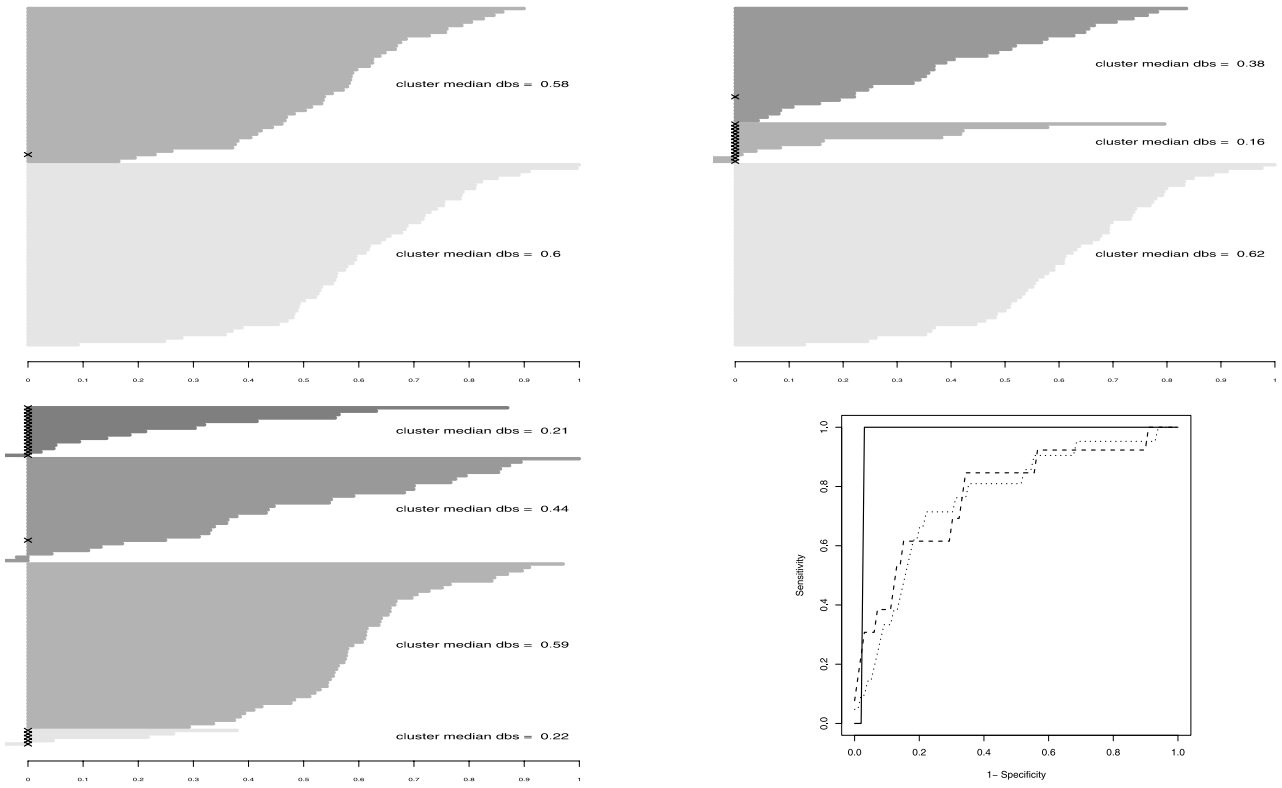

Fig. 8 Cf. Fig. 3. The first simulated data set has been clustered by MCLUST. When the procedure is forced to return a wrong number of clusters, the additional clusters have a small median *dbs* with respect to the actual groups

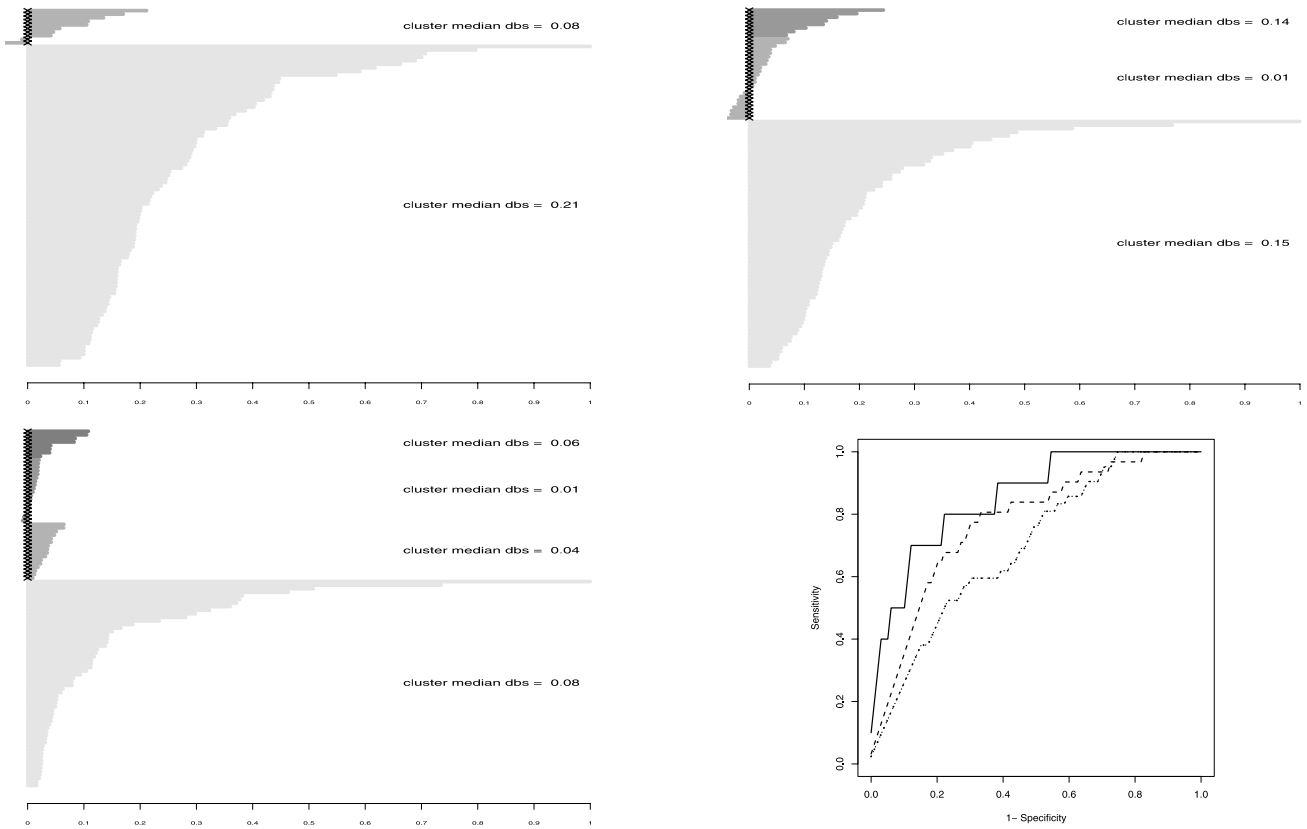

Fig. 9 Cf. Fig. 3. The second simulated data set has been clustered by AT. The smallest groups have a very low median *dbs*, being evidence of the absence of groups. Steep ROC curves show good diagnostic abilities of the *dbs* across the three partitions

analysis has been increasingly used as a tool to evaluate discriminate effects among different diagnostic methods. The ROC curve is a graphical plot of the true sensitivity vs. 1-specificity for a binary classifier as its discrimination threshold varies. It represents equally the plot of the fraction of true positives vs. the fraction of false positives when the classification threshold varies. The best possible diagnostic method would yield a point in the upper left corner or coordinate (0, 1) of the ROC space, representing 100% sensitivity (all true positives are found) and 100% specificity (no false positives are found). A completely random guess would give a point along a diagonal line from the bottom left to the top right corners. A ROC curve for evaluating the diagnostic ability of the *dbs* should be constructed by plotting, for each observed value m of the density-based Silhouette, the proportion of all the misclassified data with $dbs \leq m$ (the y-coordinate) versus the proportion of all the data with $dbs \leq m$ (the x coordinate). For computational convenience the results which follow refer to ROC curves built when m varies across a range of one hundred quantiles of the dbs, instead of across all the observed values.

The bottom right panel of Figs. 3 to 10 compares the ROC curves of the *dbs* across a range of partitions with two, three and four clusters. The analysis confirms the previous considerations because the ROC curves of the proposed method lie above the bisector of the ROC space. A remarkable situation occurs in Figs. 4, 8 and 10 where the ROC curves referring to the actual number of clusters jump almost immediately to one, meaning that minimum values of the *dbs* correspond to incorrect labels and no misclassified observation has a large *dbs* (no false negatives are found). Moreover, the *dbs* looks consistent across partitions with a distinct number of groups, the ROC curves being steep in almost all the considered examples. However, quite low ROC curves are returned when MCLUST is forced to return fewer clusters than the actual structure of groups (see the real data examples), suggesting a possible bias of the density-based Silhouette with respect to the number of clusters. The issue of possible forms of bias affecting the clustering validation techniques has been discussed by Handl et al. (2005) which show that the correct partitioning may not score well under the Silhouette information when distinct partitions are considered.

Some further investigations have been conducted with the twofold aim of catching possible forms of bias and considering the use of a summarizing measure of the global quality. In particular, the behaviour of the overall mean and median *dbs* has been evaluated when the number of clusters varies. The analysis has shown that there do not seem to be rea-

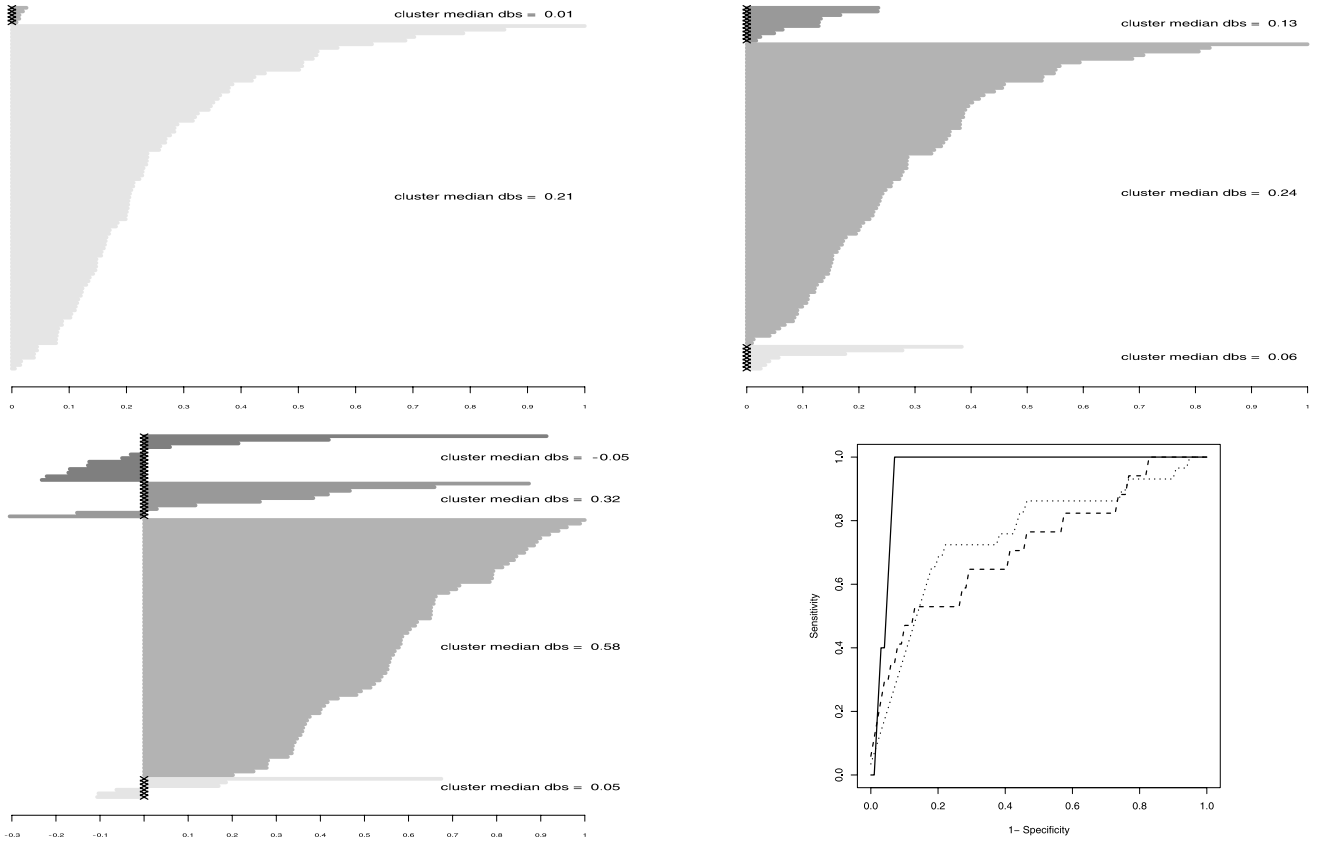

Fig. 10 Cf. Fig. 3. The second simulated data set has been clustered by MCLUST. The minor groups have a very small median *dbs*, being evidence of the absence of groups. Steep ROC curves mean good diagnostic abilities of the *dbs* across the three partitions, especially the first

sons for preferring the median to the mean as a measure of a global accuracy (unlike the evaluation of the quality of a single cluster, there is no need to use a robust location index after the *dbs* normalization). In fact, given that the mean is affected by negative and small values of the *dbs* it slightly overperforms the *dbs* median. A slight tendency of the *dbs* in favouring model-based partitions with two clusters has emerged, but in the majority of the considered situations the partition with the largest mean and median *dbs* is the one which most overlaps with the actual clustering structure.[2]

It should be stressed that, unlike supervised problems of classification, in a clustering framework the observations are not associated with a true label class and the solution provided by the application of a clustering method represents just one possible partition of the data in groups of similar observations. While keeping in mind these considerations, the conducted empirical analysis may give us some suggestions about how to use the *dbs* information in practice:

– among distinct partitions, choose the one with the highest median (or mean) *dbs*;
– clusters having a median *dbs* close to zero are likely to be spurious groups;
– negative values of *dbs* are interpreted as misclassified observations, i.e. observations being more similar to the objects belonging to alternative groups;
– in general, observations getting a *dbs* value close to zero lie on the margins of the groups, corresponding to the valley of the density estimate. We are not able to decide about the correct or incorrect allocations of these observations but their *dbs* value denotes a poor homogeneity with respect to the other observations belonging to the same group.

Our analysis has included also the computation of the original Silhouette information on the clusters returned by the distance-based Ward clustering method. The evaluation has been conducted by cutting the dendrogram produced by the Ward method to output a given number of clusters, set equal to the actual number of groups. Pointedly, an exception has been made for the standard normal data where two clusters have been considered instead of one, in accordance with the density-based clustering methods. Figure 11 shows the results. From the considered examples we see that the Silhouette information is quite sensitive at recognizing misclas-

[2]Further details may be found on the supplementary material.

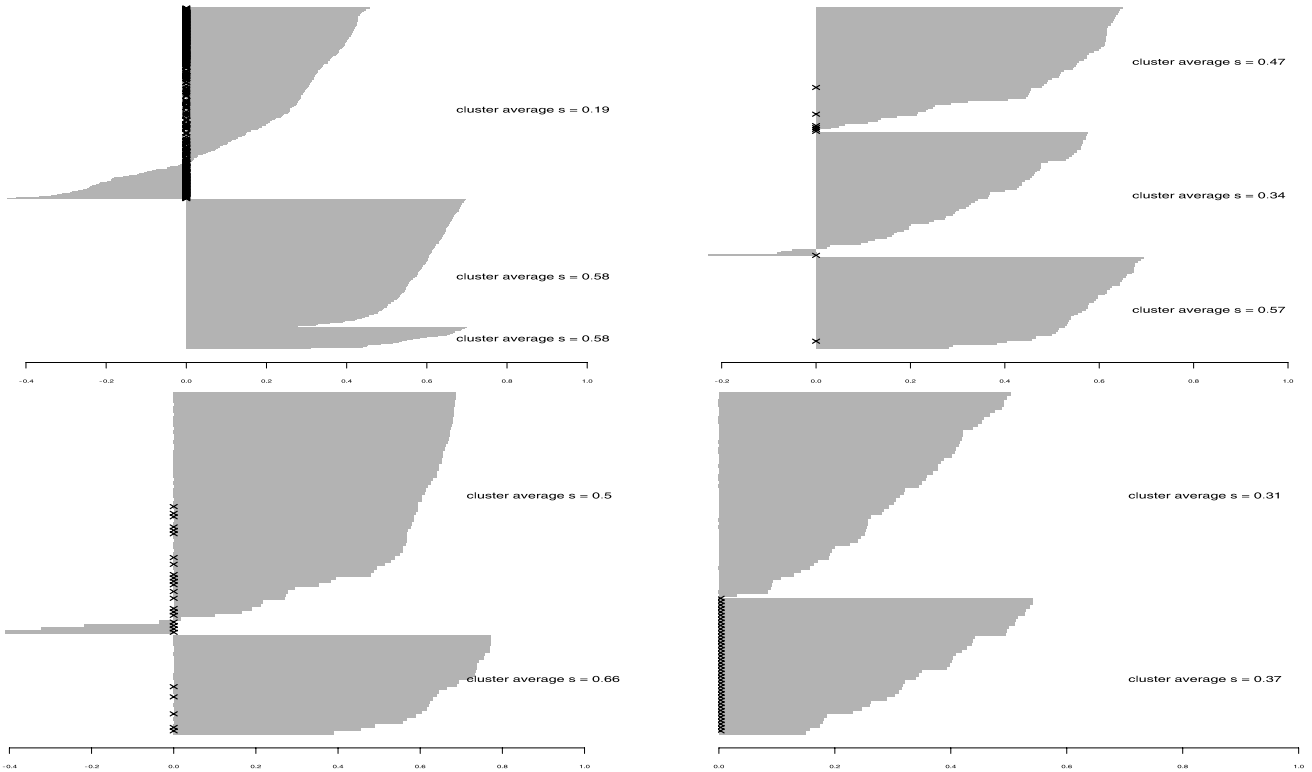

Fig. 11 Silhouette plot of the Ward method on the olive oil (*top left panel*), wine data (*top right panel*), simulated two groups (*bottom left panel*) and normal data (*bottom right panel*). A poor performance emerges from the use of s on the olive oil data and normal data partitions

sified observations. However, it is generally outperformed by its density-based version. The comparison between the ROC curves of s and *dbs* (Fig. 12) shows that the *dbs* produce uniformly steeper curves than the Silhouette information in three of the four considered data sets. Moreover, the ROC curves show even more clearly that, when the data are not partitionable, the density-based Silhouette shows good performance while the original Silhouette does not recognize the absence of groups (both the density-based and the distance-based Silhouette have been computed on the two group clustering of the standard normal data).

The ROC analysis suggests some further reflections on the *dbs* behaviour when it is used in a model-based framework compared with a nonparametric framework. Unlike the parametric clustering, the latter approach basically associates the clusters with the bumps in the density estimate, thus resulting in groups which are apparently more separated than groups correspondent to the components of a mixture model. It might follow that an optimistic evaluation of *dbs* is due to the large log ratios between the conditional probabilities involved in the (3). However, the normalization factor largely reduces this risk. In fact, the presented analysis does not provide elements to think that the *dbs* is biased toward favouring parametric or nonparametric methods.

6 Concluding remarks

In this work a diagnostic tool aimed at evaluating the quality of the partition generated by a clustering procedure has been presented. This method is similar to the Silhouette information but, unlike the Rousseeuw index, it is developed for appraising the quality of a density-based clustering method. It is based on the estimation of the posterior probabilities that the observations belong to the detected groups.

The idea of using the Bayes formula for evaluating the allocation of data points to the clusters is, indeed, not completely new. Fraley and Raftery (2002), for instance, assess the confidence of each clustered data by estimating the posterior probability that each observation does not belong to the group where it has been allocated. This measure highly agrees (in a reverse sense) with the *dbs* both when two clusters only are detected and when groups are well separated. Indeed, while the uncertainty index of Fraley and Raftery basically evaluates the cluster compactness or homogeneity, the density-based Silhouette compares a measure of compactness with a measure of separation.

The density-based Silhouette claims to be a completely general technique, not linked to any specific density estimator nor to a clustering method. An application to real and simulated data has shown the ability of the proposed method in recognizing well clustered data, (likely) misclassified or

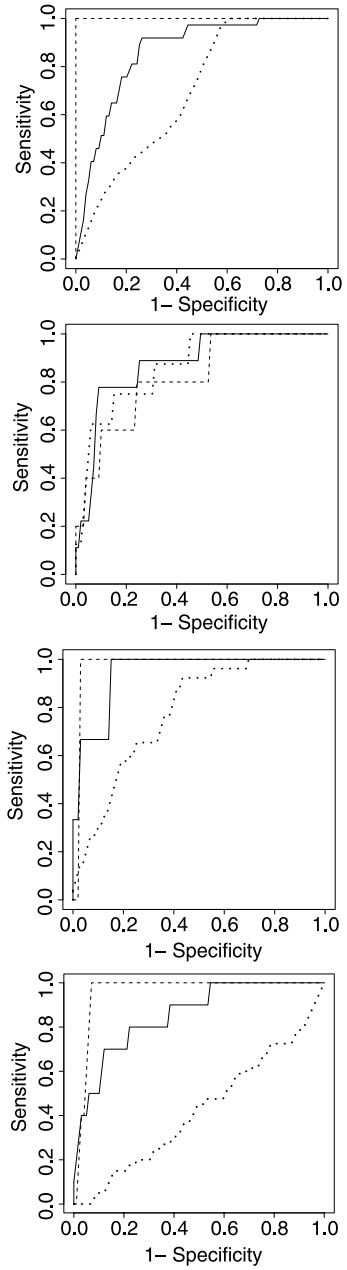

Fig. 12 From the *top*: ROC curves for olive oil, wine, simulated two groups and normal data, respectively. The curves correspond to the best partitions detected by each clustering method (displayed in Figs. 3 to 10 and 11). The *solid lines* refer to the *dbs* returned by applying AT, the *dashed lines* correspond to MCLUST, the *dotted line* is the ROC curve applied on the Silhouette information of the Ward partition

less confident observations, and in determining the best partition among several groupings. A ROC analysis has further highlighted good levels of sensitivity and specificity and a generally better diagnostic ability than the original Silhouette.

Acknowledgements The author wishes to gratefully acknowledge the Associate Editor and the reviewers for their useful remarks. A special thanks to professor Nicola Torelli for his comments and to professor Adelchi Azzalini for his suggestions that greatly improved the presentation of this paper.

Supplementary information: Enlarged figures and some supplementary material are available at http://www2.units.it/~nirdses/sito_inglese/working papers/files for wp/wp125.pdf.

References

Ankerst, M., Breuning, M.M., Kriegel, H., Sander, J., Optics: Ordering points to identify the clustering structure. In: Proc. ACM SIGMOD Int. Conf. on Manag. Data (SIGMOD-96), pp. 49–60 (1999)

Azzalini, A., Torelli, N.: Clustering via nonparametric density estimation. Stat. Comput. **17**, 71–80 (2007)

Banfield, J.D., Raftery, A.E.: Model-based Gaussian and non-Gaussian clustering. Biometrics **49**, 803–821 (1993)

Bensmail, H., Celeux, G., Raftery, A.E., Robert, C.P.: Inference in model-based cluster analysis. Stat. Comput. **7**, 1–10 (1997)

Bezdek, J.C., Pal, N.R.: On cluster validity for the fuzzy c-means model. IEEE Trans. Fuzzy Syst. **3**, 190–193 (1995)

Binder, D.A.: Bayesian cluster analysis. Biometrika **65**, 31–38 (1978)

Binder, D.A.: Approximations to bayesian clustering rules. Biometrika **68**, 275–285 (1981)

Chang, W.C.: On using principal components before separating a mixture of two multivariate normal distributions. Appl. Stat. **32**, 267–275 (1983)

Cuevas, A., Febrero, M., Fraiman, R.: Cluster analysis: a further approach based on density estimation. Comput. Stat. Data Anal. **36**, 441–459 (2001)

Cutler, A., Windham, M.P.: Information-based validity functionals for mixture analysis. In: Proc. 1st US/Japan Conf. Front. Stat. Model., Bozdogan. Kluwer Academic, Norwell (1994)

Davies, D., Bouldin, D.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **1**, 224–227 (1979)

Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley, New York (1973)

Dunn, J.: Well separated clusters and optimal fuzzy partitions. J. Cybern. **57**, 3–32 (1974)

Dy, J.G., Brodley, C.: Feature selection for unsupervised learning. J. Mach. Learn. Res. **5**, 845–889 (2004)

Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density based algorithm for discovering clusters in large spatial databases withe noise. In: Proc. 2nd Int. Conf. Knowl. Discov. Data Min. (KDD-96). AAAI Press, Menlo Park (1996)

Fawcett, T.: An introduction to ROC analysis. Pattern Recognit. Lett. **27**, 861–874 (2006)

Forina, M., Armanino, C., Lanteri, S., Tiscornia, E.: Classification of olive oils from their fatty acid composition. In: Martens, M., Russwurm, H.J. (eds.) Food Research and Data Analysis, pp. 189–214. Appl. Sci., London (1983)

Forina, M., Armanino, C., Castino, M., Ubigli, M.: Multivariate data analysis as a discriminating method of the origin of wines. Vitis **25**, 189–201 (1986)

Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis and density estimation. J. Am. Stat. Assoc. **97**, 611–631 (2002)

Fraley, C., Raftery, A.E.: MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Tech. Rep. 504, Univ. of Washington, Dep. of Stat. (2006)

Handl, J., Knowles, J., Kell, D.B.: Computational cluster validation in post-genomic data analysis. Bioinformatics **21**(15), 3201–3212 (2005)

Hartigan, J.A.: Clustering Algorithms. Wiley, New York (1975)

Hubert, L.J., Schultz, J.W.: Quadratic assignment as a general data analysis strategy. Br. J. Math. Stat. Psychol **29**, 190–241 (1976)

Kaufman, L., Rousseeuw, P.J.: Finding Groups in data: an introduction to cluster analysis. Wiley, New York (1990)

Maulik, U., Bandyopadhyay, S.: Performance evaluation of some clustering algorithms and validity indices. IEEE Trans. Pattern Anal. Mach. Intell. **24**, 1650–1654 (2002)

McLachlan, G.J., Peel, D.: Robust Cluster Analysis via Mixtures of Multivariate t-Distributions, pp. 658–666. Springer, Berlin (1998),

Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)

Sander, J., Ester, M., Kriegel, H.P., Xu, X.: Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. Data Min. Knowl. Discov. **2**, 169–194 (1998)

Stuetzle, W.: Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. J. Classif. **20**, 25–47 (2003)

Xie, X.L., Beni, G.: A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. **13**, 841–847 (1991)