A method for visual cluster validation

Christian Hennig[1]

Fachbereich Mathematik - SPST,
Universität Hamburg, 20146 Hamburg, Germany

Abstract. Cluster validation is necessary because the clusters resulting from cluster analysis algorithms are not in general meaningful patterns. I propose a methodology to explore two aspects of a cluster found by any cluster analysis method: the cluster should be separated from the rest of the data, and the points of the cluster should not split up into further separated subclasses. Both aspects can be visually assessed by linear projections of the data onto the two-dimensional Euclidean space. Optimal separation of the cluster in such a projection can be attained by asymmetric weighted coordinates (Hennig (2002)). Heterogeneity can be explored by the use of projection pursuit indexes as defined in Cook, Buja and Cabrera (1993). The projection methods can be combined with splitting up the data set into clustering data and validation data. A data example is given.

1 Introduction

Cluster validation is the assessment of the quality and the meaningfulness of the outcome of a cluster analysis (CA). Most CA methods generate a clustering in all data sets, whether there is a meaningful structure or not. Furthermore, most CA methods partition the data set into subsets of a more or less similar shape, and this may be adequate only for parts of the data, but not for all. Often, different CA methods generate different clusterings on the same data and it has to be decided which one is the best, if any. Therefore, if an interpretation of a cluster as a meaningful pattern is desired, the cluster should be validated by information other than the output of the CA. A lot of more or less formal methods for cluster validation are proposed in the literature, many of which are discussed, e.g., in Gordon (1999, Section 7.2) and Halkidi et al. (2002). Six basic principles for cluster validation can be distinguished:

Use of external information External information is information that has not been used to generate the clustering. Such information can stem from additional data or from background knowledge. However, such information is often not available.

Significance tests for structure Significance tests against null models formalizing "no clustering structure at all" are often used to justify the interpretation of a clustering. While the rejection of homogeneity is a reasonable minimum requirement for a clustering, such tests cannot validate the concrete structure found by the CA algorithm.

Comparison of different clusterings on the same data Often, the agreement of clusterings based on different methods is taken as a confirmation of clusters. This is only meaningful if sufficiently different CA methods have been chosen, and in the case of disagreement it could be argued that not all of them are adequate for the data at hand.

Validation indexes In some sense, the use of validation indexes is similar to that of different clusterings, because many CA methods optimize indexes that could otherwise be used for validation.

Stability assessment The stability of clusters can be assessed by techniques such as bootstrap, cross-validation, point deletion, and addition of contamination.

Visual inspection Recently (see, e.g., Ng and Huang (2002)), it has been recognized that all formal approaches of cluster validation have limitations due to the complexity of the CA problem and the intuitive nature of what is called a "cluster". Such a task calls for a more subjective and visual approach. To my knowledge, the approach of Ng and Huang (2002) is the first visual technique which is specifically developed for the validation of a clustering.

Note that these principles address different aspects of the validation problem. A clustering that is well interpretable in the light of external information will not necessarily be reproduced by a different clustering method. Structural aspects such as homogeneity of the single clusters and heterogeneity between different clusters as indicated by validation indexes or visual inspection are not necessarily properties of clusters which are stable under resampling. However, these aspects are not "orthogonal". A well chosen clustering method should tend to reproduce well separated homogeneous clusters even if the data set is modified.

In the present paper, a new method for visual cluster validation is proposed. As opposed to the approach of Ng and Huang (2002), the aim of the present method is to assess every cluster individually. The underlying idea is that a valid cluster should have two properties:

- separation from the rest of the data, so that it should not be joined with other parts of the data,
- homogeneity, so that the points of the cluster can be said to "belong together".

In Section 2, asymmetric weighted coordinates (AWCs) are introduced. AWCs provide a linear projection of the data in order to separate the cluster under study optimally from the rest of the data. In Section 3, I propose the application of some projection pursuit indexes to the points of the cluster to explore its heterogeneity. Additionally, if there is enough data to split the data set into a "training sample" and a "validation sample", the projections obtained from clustering and visual validation on the training sample can be applied also to the points of the validation sample to see if the found patterns can

be reproduced. Throughout the paper, the data is assumed to come from the p-dimensional Euclidean space. The methods can also be applied to distance data after carrying out an appropriate multidimensional scaling method. Euclidean data is only needed for the validation; the clustering can be done on the original distances. In Section 4, the method is applied to a real data set.

2 Optimal projection for separation

The most widespread linear projection technique to separate classes goes back to Rao (1952) and is often called "discriminant coordinates" (DCs). The first DC is defined by maximizing the ratio

$$F(\mathbf{c}_1) = \frac{\mathbf{c}_1'\mathbf{B}\mathbf{c}_1}{\mathbf{c}_1'\mathbf{W}\mathbf{c}_1}, \text{ where}$$

$$\mathbf{W} = \frac{1}{n-s}\sum_{i=1}^{s}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\mathbf{m}_i)(\mathbf{x}_{ij}-\mathbf{m}_i)',$$

$$\mathbf{B} = \frac{1}{n(s-1)}\sum_{i=1}^{s} n_i(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})'.$$

n denotes the number of points, n_i is the number of points of class i, s denotes the number of classes, $\mathbf{x}_{i1},\ldots,\mathbf{x}_{in_i}$ are the p-dimensional points of class i, $\mathbf{m}_i$ is the mean vector of class i and $\mathbf{m}$ is the overall mean. The further DCs maximize F under the constraint of orthogonality to the previous DCs w.r.t. $\mathbf{W}$. $\mathbf{B}$ is a covariance matrix for the class means and $\mathbf{W}$ is a pooled within-class covariance matrix. Thus, F gets large for projections that separate the means of the classes as far as possible from each other while keeping the projected within-class variation small. Some disadvantages limit the use of DCs for cluster validation. Firstly, separation is formalized only in terms of the class means, and points of different classes far from their class means need not to be well separated (note that the method of Ng and Huang (2002) also aims at separating the cluster centroids). Secondly, $s-1$ dimensions are needed to display all information about the separation of s classes, and therefore there is no guarantee that the best separation of a particular cluster shows up in the first two dimensions in case of $s > 3$. This could in principle be handled by declaring the particular cluster to be validated as class 1 and the union of all other clusters as class 2 (this will be called the "asymmetry principle" below). But thirdly, DCs assume that the covariance matrices of the classes are equal, because otherwise $\mathbf{W}$ would not be an adequate covariance matrix estimator for a single class. If the asymmetry principle is applied to a clustering with $s > 2$, the covariance matrices of these classes cannot be expected to be equal, not even if they would be equal for the s single clusters.

A better linear projection technique for cluster validation is the application of asymmetric linear dimension reduction (Hennig (2002)) to the two

classes obtained by the asymmetry principle. Asymmetry means that the two classes to be projected are not treated equally. Asymmetric discriminant coordinates maximize the separation between class 1 and class 2 while keeping the projected variation of class 1 small. Class 2, i.e., the union of all other data points, may appear as heterogeneously as necessary. Four asymmetric projection methods are proposed in Hennig (2002), of which asymmetric weighted coordinates (AWCs) are the most suitable for cluster validation. The first AWC is defined by maximizing

$$F^*(\mathbf{c}_1) = \frac{\mathbf{c}_1'\mathbf{B}^*\mathbf{c}_1}{\mathbf{c}_1'\mathbf{S}_1\mathbf{c}_1}, \text{ where}$$

$$\mathbf{S}_1 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(\mathbf{x}_{1j} - \mathbf{m}_1),$$

$$\mathbf{B}^* = \sum_{i,j} w_j(\mathbf{x}_{1i} - \mathbf{x}_{2j})(\mathbf{x}_{1i} - \mathbf{x}_{2j})',$$

$$w_j = \min\left(1, \frac{d}{(\mathbf{x}_{2j} - \mathbf{m}_1)'\mathbf{S}_1^{-1}(\mathbf{x}_{2j} - \mathbf{m}_1)}\right),$$

$d > 0$ being some constant, for example the 0.99-quantile of the χ^2_p-distribution. The second AWC $\mathbf{c}_2$ maximizes F^* subject to $\mathbf{c}_1'\mathbf{S}_1^{-1}\mathbf{c}_2 = 0$ and so on. $\mathbf{c}_1'\mathbf{B}^*\mathbf{c}_1$ gets large if the projected differences between points of class 1 and class 2 are large. The weights w_j downweight differences from points of class 2 that are very far away (in Mahalanobis distance) from class 1. Otherwise, $\mathbf{c}_1'\mathbf{B}^*\mathbf{c}_1$ would be governed mainly by such points, and class 1 would appear separated mainly from the furthest points in class 2, while it might be mixed up more than necessary with closer points of class 2. The weights result in a projection that separates class 1 also from the closest points as well as possible. More motivation and background is given in Hennig (2002). As for DCs, the computation of AWCs is very easily be done by an Eigenvector decomposition of $\mathbf{S}_1^{-1}\mathbf{B}^*$. Note that AWCs can only be applied if $n_1 > p$, because otherwise class 1 could be projected onto a single point, thus $\mathbf{c}_1'\mathbf{S}_1^{-1}\mathbf{c}_1 = 0$. If n_1 is not much larger than p, $\mathbf{c}_1'\mathbf{S}_1^{-1}\mathbf{c}_1$ can be very small, and some experience (e.g., with simulated data sets from unstructured data) is necessary to judge if a seemingly strong separation is really meaningful.

3 Optimal projection for heterogeneity

Unfortunately, AWCs cannot be used to assess the homogeneity of a cluster. The reason is that along projection directions that do not carry any information regarding the cluster, the cluster usually does not look separated, but often more or less homogeneous. Thus, to assess separation, the projected separation has to be maximized, which is done by AWCs. But to assess homogeneity, it is advantageous to maximize the projected *heterogeneity* of the cluster.

Projection pursuit is the generic term for linear projection methods that aim for finding "interesting", i.e., heterogeneous projections of the data (Huber (1985)). The idea is to project only the points of the cluster to be validated in order to find a most heterogeneous visualization. There are lots of projection pursuit indexes. Some of them are implemented in the data visualization software XGOBI (Buja et al. (1996)). A main problem with projection pursuit is that the indexes can only be optimized locally. XGOBI visualizes the optimization process dynamically, and after a local optimum has been found, the data can be rotated toward new configurations to start another optimization run.

Two very simple and useful indexes have been introduced by Cook et al. (1993) and are implemented in XGOBI. The first one is the so-called "holes index", which is defined by minimizing

$$F^{**}(\mathbf{C}) = \sum_{i=1}^{n_1} \varphi_2(\mathbf{C}'\mathbf{x}_{1i}),$$

over orthogonal $p \times 2$-projection matrices $\mathbf{C}$, where φ_2 denotes the density of the two-dimensional Normal distribution and the points $\mathbf{x}_{1i}$ are assumed to be centered and scaled. F^{**} becomes minimal if as few points as possible are in the center of the projection, in other words, if there is a "hole". Often, such a projection shows a possible division of the cluster points into subgroups.

It is also useful to maximize F^{**}, which is called "central mass index" in XGOBI. This index attempts to project as many points as possible into the center, which can be used to find outliers in the cluster. But it can also be useful to try out further indexes, as discussed in Cook et al. (1993).

4 Example

As an example, two CA methods have been applied to the "quakes" data set, which is part of the base package of the free statistical software R (to obtain from `www.R-project.org`). The data consist of 1000 seismic events on Fiji, for which five variables have been recorded, namely geographical longitude and latitude, depth, Richter magnitude and number of stations reporting. Because of the favorable relation of n to p, I divided the data set into 500 points that have been used for clustering and 500 points for validation.

The first clustering has been generated by MCLUST (Fraley and Raftery (2003)), a software for the estimation of a Normal mixture model including noise, i.e., points that do not belong to any cluster. The Bayesian information criterion has been used to decide about the number of clusters and the complexity of the model for the cluster's covariance matrices. It resulted in four clusters with unrestricted covariance matrices plus noise. As a comparison, I have also performed a 5-means clustering on sphered data.

Generally, the validity of the clusters of the MCLUST-solution can be confirmed. In Figure 1, the AWC plot is shown for the second cluster (points

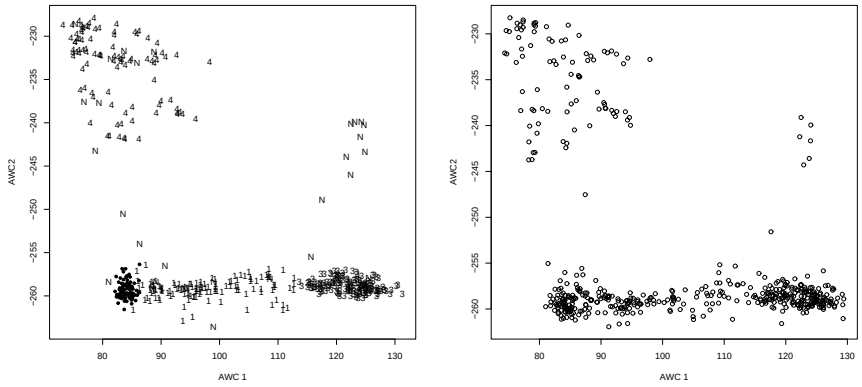

Fig. 1. Left: AWCs of cluster 2 (black points) of the MCLUST solution. Right: validation data set projected onto the AWCs.

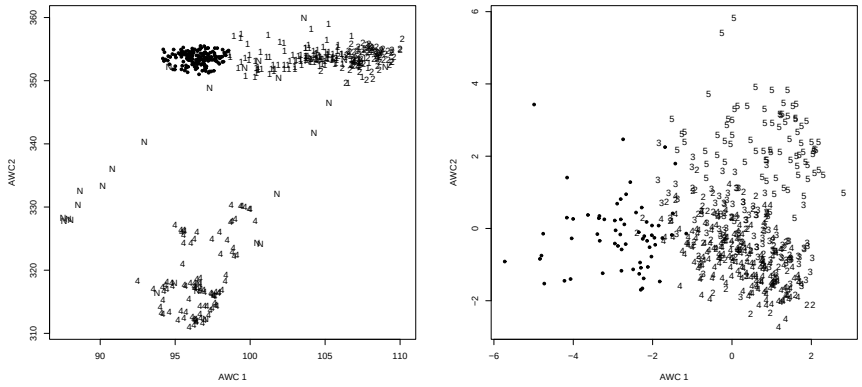

Fig. 2. Left: AWCs of cluster 3 (black points) of the MCLUST solution. Right: AWCs of cluster 1 (black points) of the 5-means solution.

of other clusters are always indicated with the cluster numbers). These points do neither appear separated in any scatterplot of two variables nor in the principal components (not shown), but they are fairly well separated in the AWC plot, and the projection of the validation points on the AWCs (right side) confirms that there is a meaningful pattern. Other clusters are even better separated, e.g., cluster 3 on the left side of Figure 2. Some of the clusters of the 5-means solution have a lower quality. For example, the AWC-plot of cluster 1 (right side of Figure 2) shows the separation as dominated by the variation within this cluster.

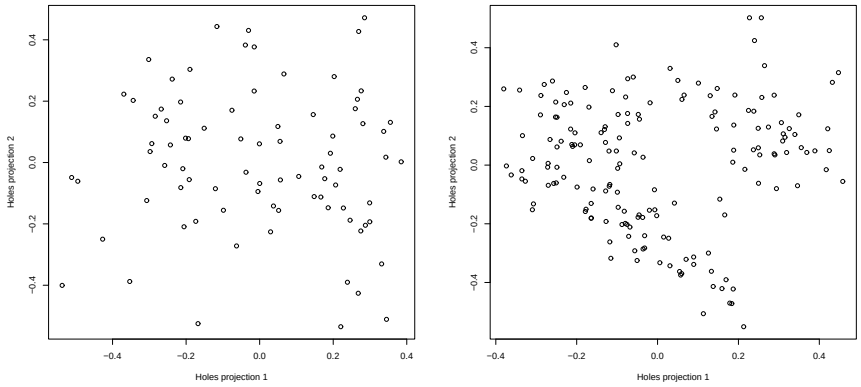

Fig. 3. Left: "holes" projection of cluster 2 of the MCLUST solution. Right: "holes" projection of cluster 3 of the MCLUST solution.

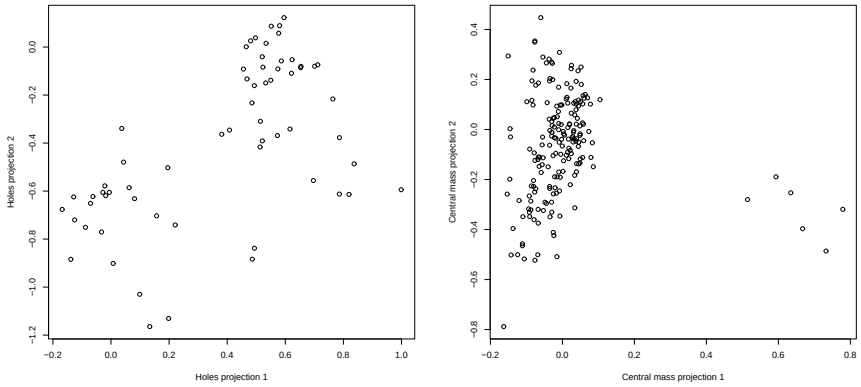

Fig. 4. Left: "holes" projection of cluster 1 of the 5-means solution. Right: "central mass" projection of cluster 4 of the 5-means solution.

Optimization of the holes index did not reveal any heterogeneity in MCLUST-cluster 2, see the left side of Figure 3, while in cluster 3 (right side) two subpopulations could roughly be recognized. Sometimes, when applying MCLUST to other 500-point subsamples of the data, the corresponding pattern is indeed divided into two clusters (it must be noted that there is a non-negligible variation in the resulting clustering structures from MCLUST, including the estimated number of clusters, on different subsamples). Some of the 5-means clusters show a much clearer heterogeneity. The holes index reveals some subclasses of cluster 1 (right side of Figure 4), while the central mass index highlights six outliers in cluster 4 (right side).

5 Conclusion

A combination of two plots for visual cluster validation of every single cluster has been proposed. AWCs optimize the separation of the cluster from the rest of the data while the cluster is kept homogeneous. Projection pursuit is suggested to explore the heterogeneity of a cluster.

Note that for large p compared to n, the variety of possible projections is large. Plots in which the cluster looks more or less separated or heterogeneous are found easily. Thus, it is advisable to compare the resulting plots with the corresponding plots from analogous cluster analyses applied to data with the same n and p generated from "null models" such as a normal or uniform distribution to assess if the cluster to be validated yields a stronger pattern. This may generally be useful to judge the validity of visual displays.

The proposed plots are static. This has the advantage that they are reproducible (there may be a non-uniqueness problem with projection pursuit) and they are optimal with respect to the discussed criteria. However, a further dynamical visual inspection of the data by, e.g., the grand tour as implemented in XGOBI (Buja et al. (1996)), can also be useful to assess the stability of separation and heterogeneity as revealed by the static plots.

AWCs are implemented in the add-on package FPC for the statistical software package R, available under `www.R-project.org`.

References

BUJA, A., COOK, D. and SWAYNE, D. (1996): Interactive High-Dimensional Data Visualization. *Journal of Computational and Graphical Statistics, 5, 78–99.*

COOK, D., BUJA, A. and CABRERA, J. (1993): Projection Pursuit Indexes Based on Orthonormal Function Expansions. *Journal of Computational and Graphical Statistics, 2, 225–250.*

FRALEY, C. and RAFTERY, A. E. (2003): Enhanced Model-Based Clustering, Density Estimation and Discriminant Analysis Software: MCLUST. *Journal of Classification 20, 263–293.*

GORDON, A.D. (1999): *Classification* (2nd Ed.). Chapman & Hall/CRC, Boca Raton.

HALKIDI, M., BATISTAKIS, Y., VAZIRGIANNIS, M. (2002): Cluster Validity Methods: Part I. *SIGMOD Record 31, 40–45.*

HENNIG, C. (2002): Symmetric, asymmetric and robust linear dimension reduction for classification. To appear in *Journal of Computational and Graphical Statistics*, `ftp://ftp.stat.math.ethz.ch/Research-Reports/108.html`.

HUBER, P. J. (1985): Projection pursuit (with discussion). *Annals of Statistics, 13, 435–475.*

NG, M. and HUANG, J. (2002): M-FastMap: A Modified FastMap Algorithm for Visual Cluster Validation in Data Mining. In: M.-S. Chen, P. S. Yu and B. Liu (Eds.): *Advances in Knowledge Discovery and Data Mining. Proceedings of PAKDD 2002, Taipei, Taiwan.* Springer, Heidelberg, 224–236.

RAO, C. R. (1952): *Advanced Statistical Methods in Biometric Research*, Wiley, New York.