Visual Inspection of Fuzzy Clustering Results

Frank Klawonn, Vera Chekhtman, and Edgar Janz

Department of Computer Science
University of Applied Sciences Braunschweig/Wolfenbuettel
Salzdahlumer Str. 46/48
D-38302 Wolfenbuettel, Germany
f.klawonn@fh-wolfenbuettel.de

1 Introduction

Clustering is an explorative data analysis method applied to data in order to discover structures or certain groupings in a data set. Therefore, clustering can be seen as an unsupervised classification technique. Fuzzy clustering accepts the fact that the clusters or classes in the data are usually not completely well separated and thus assigns a membership degree between 0 and 1 for each cluster to every datum.

The most common fuzzy clustering techniques aim at minimizing an objective function whose (main) parameters are the membership degrees and the parameters determining the localisation as well as the shape of the clusters. The algorithm will always compute a result that might represent an undesired local minimum of the objective function. Even if the global minimum is found, it might correspond to a bad result, when the cluster shapes or the number of the clusters are not chosen properly. Since the data are usually multi-dimensional, the visual inspection of the data is very limited. Methods like multi-dimensional scaling are available, but lead very often to unsatisfactory results. Nevertheless, it is important to evaluate the clustering result. Although cluster validity measures try to solve this problem, they tend to reduce the information of a large data set and a number of cluster parameters to a single value.

We propose in this paper to use the underlying principles of validity measures, but to refrain from the simplification to a single value and instead provide a graphical representation containing more information. This enables the user to identify inconsistencies in the clustering result or even in single clusters.

Section 2 briefly reviews the necessary background in objective function-based fuzzy clustering. The concept of validity measures is discussed in section 3. The techniques for visualisation are introduced in section 4 and in the

final conclusions we outline that the visualisation techniques tailored for fuzzy clustering are even useful in the case of crisp clustering.

2 Objective Function-Based Fuzzy Clustering

Fuzzy clustering is suited for finding structures in data. A data set is divided into a set of clusters and – in contrast to hard or deterministic clustering – a datum is not assigned to a unique cluster. In order to handle noisy and ambiguous data, membership degrees of the data to the clusters are computed. Most fuzzy clustering techniques are designed to optimise an object function with constraints. The most common approach is the so called probabilistic clustering with the objective function

$$f \;=\; \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} d_{ij} \tag{1}$$

under the constraints

$$\sum_{i=1}^{c} u_{ij} \;=\; 1 \qquad \text{for all } j = 1, \ldots, n. \tag{2}$$

It is assumed that the number of clusters c is fixed. The set of data to be clustered is $\{x_1, \ldots, x_n\} \subset \mathbb{R}^p$. u_{ij} is the membership degree of datum x_j to the ith cluster. d_{ij} is some distance measure specifying the distance between datum x_j and cluster i, for instance the (squared) Euclidean distance of x_j to the ith cluster centre. The parameter $m > 1$, called fuzzifier, controls how much clusters may overlap. The constraints (2) lead to the name probabilistic clustering, since in this case the membership degree u_{ij} can also be interpreted as the probability that x_j belongs to cluster i. The parameters to be optimised are the membership degrees u_{ij} and the cluster parameters that are not given explicitly here. They are hidden in the distances d_{ij}. Since this is a non-linear optimisation problem, the most common approach to minimize the objective function (1) is to alternatingly optimise either the membership degrees or the cluster parameters while considering the other parameter set as fixed. In this paper we are not interested in the great variety of specific cluster shapes (spheres, ellipsoids, lines, quadrics,...) that can be found by choosing suitable cluster parameters and an adequate distance function. (For an overview we refer to [2, 5].) Our considerations can be applied to almost all cluster shapes. However, for shell clustering there are better suited methods. Since most shell clustering algorithms are designed for image recognition, the data are usually two-dimensional so that special visualisation techniques are not required.

The visualisation methods we propose are also suited for noise clustering [3] where the principle of probabilistic clustering is maintained, but an additional noise cluster is introduced. All data have a fixed (large) distance to the noise

cluster. In this way, data that are near the border between two clusters, still have a high membership degree to both clusters as in probabilistic clustering. But data that are far away from all clusters will be assigned to the noise cluster and have no longer a high membership degree to other clusters.

We do not cover possibilistic clustering [6] where the probabilistic constraint is completely dropped and an additional term in the objective function is introduced to avoid the trivial solution $u_{ij} = 0$. However, the aim of possibilistic clustering is actually not to find the global optimum of the corresponding objective function, since this is obtained, when all clusters are identical [7].

3 Validity Measures

Cluster validity refers to the problem whether a given (fuzzy) partition fits to the data at all. We emphasize again that the clustering algorithm will always try to find the best fit for a fixed number of clusters and the parameterised cluster shapes. However, this does not mean that even the best fit is meaningful at all. Either the number of clusters might be wrong or the cluster shapes might not correspond to the groups in the data, if the data can be grouped in a meaningful way at all. The cluster validity problem is a similar one as in linear regression. One can always find the best fitting line for a given data set, even if the data come from an exponential function. This does not mean that there is a linear dependence in the data.

Cluster validity measures are used to validate a clustering result in general or also in order to determine the number of clusters. In order to fulfill the latter task, the clustering might be carried out with different numbers of clusters and the one yielding the best value of the validity measure is assumed to have the correct number of clusters.

Let us briefly review some cluster validity measures. It is beyond the scope of this paper to provide a complete overview on validity measures and we refer for a more detailed discussion to [2, 5]. We also restrict our considerations to global validity measures that evaluate a whole fuzzy partition and not single clusters.

The partition coefficient [1] is defined by

$$\frac{\sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^2}{n}.$$

The higher the value of the partition coefficient the better the clustering result. The highest value 1 is obtained, when the fuzzy partition is actually crisp, i.e. $u_{ij} \in \{0, 1\}$. The lowest value $1/c$ is reached, when all data are assigned to all clusters with the same membership degree $1/c$. This means that a fuzzy clustering result is considered better, when it is more crisp.

The partition entropy [1]

$$\frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\ln(u_{ij})}{n}$$

is inspired by the Shannon entropy. The smaller the value of the partition entropy, the better the clustering result. This means that similar to the partition coefficient crisper fuzzy partitions are considered better.

In [4] validity measures are proposed that take the volume of the clusters into account. Let

$$A_i \;=\; \frac{\sum_{j=1}^{n} u_{ij}^m (x_j - v_i)(x_j - v_i)^\top}{\sum_{j=1}^{n} u_{ij}^m}$$

denote the (fuzzy) covariance matrix of the ith cluster where

$$v_i \;=\; \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}$$

denotes the centre of the ith cluster and the x_j are the data vectors.

Then the validity measure called the fuzzy hypervolume is defined by

$$FHV \;=\; \sum_{i=1}^{c} \sqrt{\det(A_i)}.$$

A smaller value of FHV indicates compact and therefore better clusters.

The average partition density is given by

$$\frac{1}{c}\sum_{i=1}^{c} \frac{S_i}{\sqrt{\det(A_i)}}$$

where $S_i = \sum_{j \in Y_i} u_{ij}$ and

$$Y_j \;=\; \{j \in \{1,\ldots,n\} \mid (x_j - v_i)^\top A^{-1} (x_j - v_i) < 1\}$$

is the set of data near to the cluster centre v_i. S_i corresponds to the number of data assigned to cluster i that are near to the cluster centre v_i. Therefore $\frac{S_i}{\sqrt{\det(A_i)}}$ is proportional to the average density of the data in cluster i. A high density value indicates good clusters.

Finally, the partition density is defined by

$$\frac{\sum_{i=1}^{c} S_i}{FHV}$$

where a larger value refers again to better clusters.

It should be noted that validity measures like the partition coefficient or the partition entropy rely solely on the membership degrees whereas fuzzy hypervolume and (average) partition density also take the distance of the data to the clusters into account. We will use these ideas to develop graphical validity measures in the next section.

4 Visualisation

Let us first use the membership degrees only for a graphical inspection, whether a fuzzy clustering result is acceptable. It should be noted that the underlying idea is always: The more crisp the fuzzy partition, the better the clustering result. Of course, decreasing the fuzzifier m always leads to crisper, but not necessarily better partitions. This must always be taken into account.

In order to illustrate our graphical validity criterions, we use the artificial data set shown in figure 1. This data set obviously contains three well separated clusters. We have clustered this data set with the well known fuzzy c-means algorithm with three clusters as well as four clusters. For three clusters the cluster centres are not shown in the figure, but are more or less exactly in the middle of the corresponding circles. When we use four clusters, the marked spots in figure 1 show the cluster centres. A data point has the highest membership degree to the nearest cluster centre.

As a first very simple visualisation of the membership degrees we can simply look at the distribution of the membership degrees. For the ideal case of a crisp result, i.e. each datum is assigned to exactly one cluster with member ship degree 1 and to all others with membership degree 0, we would expect the following: The relative frequency of the membership degree 1 should be $n/(cn) = 1/c$ and accordingly the relative frequency of the value 0 should be $(c-1)n/(cn) = (c-1)/c$. So, for the ideal case of crisp memberships, we would expect a distribution of the membership degrees whose chart diagram shows a value of $(c-1)/c$ on the left side and $1/c$ on the right side and zero values in between.

However, we belief that such a chart diagram is not very suitable, since its desired shape depends on the number of clusters. And for larger numbers of clusters the emphasis is mainly put only on the left side of the chart diagram. Therefore, we carried out a scaling of the values in such a way that in the ideal case the chart diagram would show a value of 1 on both the left and the right side. This means that, when counting the frequencies of the membership degrees, we introduce a weighting factor. The weighting for a membership degree of 0 is $c/(c-1)$ and for a membership degree of 1 it is c. For the computation of the chart diagram, the weighting of the membership degrees between 0 and 1 is simply linear, increasing from $c/(c-1)$ to c. A single chart in the diagram does not show the relative frequency of membership degrees in a range between a and b

$$\frac{1}{n}\,\mathrm{card}\{(i,j) \in \{1,\ldots,c\} \times \{1,\ldots,n\} \mid a \le u_{ij} < b\},$$

but the scaled frequency

$$\frac{1}{n} \sum_{(i,j):a \le u_{ij} < b} \left(\frac{c(c-2)}{c-1} u_{ij} + \frac{c}{c-1} \right).$$

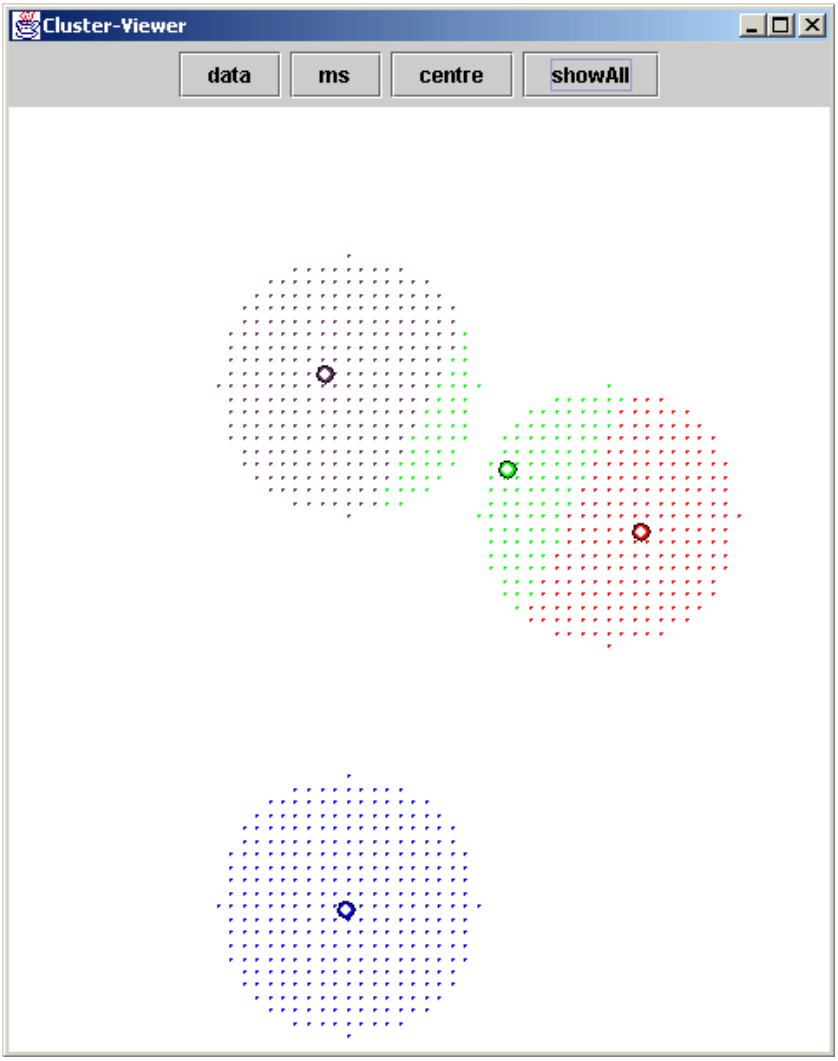

Fig. 1. Clustering result with four clusters

Figure 2 shows these scaled chart diagrams for our artificial data set, when clustered with three and four clusters. The height of the chart diagrams in the graphics is normalised, the value on the left side is 0.92 for the left and 0.90 for the right chart diagram. What is more significant is that we can see immediately that the chart diagrams differ in the middle and the right side. The right chart diagram shows less values near 1 and more ambiguous membership degrees. This is an indicator for a non-optimal clustering result.

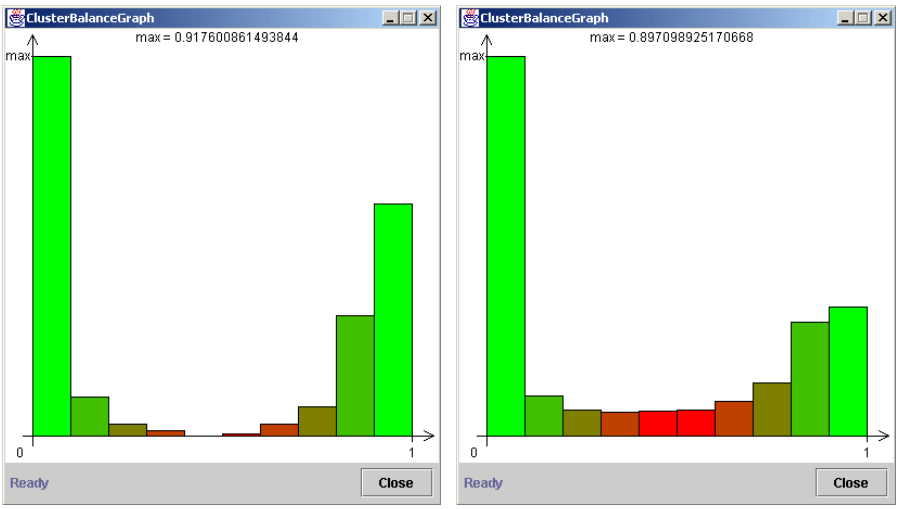

Fig. 2. Scaled membership distributions for three (left) and four (right) clusters

Although these chart diagrams provide already interesting information on the clustering result, we propose to have a look at another diagram as well. For each datum x_j we determine the cluster for which x_j has the highest membership degree, say clusters i_1, and the cluster yielding the second highest membership degree, say i_2. Then for each x_j we plot a point at the coordinates $(u_{i_1 j}, u_{i_2 j})$ leading to a diagram as shown in figure 3.

It is clear that all points must lie within the triangle defined by the points (0,0), (0.5,0.5) and (1,0), since the first coordinate must always be larger than the second one and according to the probabilistic constraint (2) we have $u_{i_1 j} + u_{i_2 j} \leq 1$.

Having again the ideal case of (almost) crisp membership degrees in mind, all points would be plotted near the point (1,0). Points near (0.5,0.5) indicate ambiguous data that are shared by two clusters. Points near (0,0) usually originate from noise data that have a low membership degree to all clusters.

The upper left graph in figure 3 shows this diagram at the beginning of the clustering algorithm, when the cluster centres are initialised randomly. Of course, with random cluster centres very few data are near to a cluster centre and we find only few points near (1,0).

The lower part of figure 3 shows this diagram of maximum membership degrees after we have carried out the clustering completely with three and four clusters. Since for the partition with three clusters more points are concentrated near (1,0) than for four clusters, we see from the diagram that three clusters should be preferred.

In the case of very large data sets, we recommend not to plot a single point for each datum. A colour plot where the intensity of the colour represents the

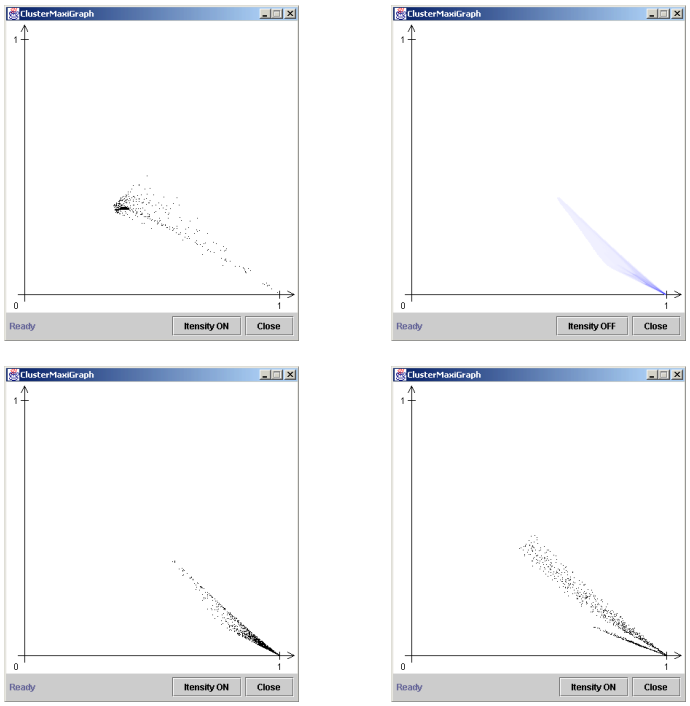

Fig. 3. Maximum membership degrees for the random initialisation (upper left), for three (lower left) and four (lower right) clusters after convergence and an intensity plot for a large data set (upper right)

density of points should be chosen instead. The upper right graph in figure 3 shows such an intensity plot for a similar data set as in figure 1, except that the density of the data is increased, so that we have 192231 instead of 951 data.

So far we have considered only the membership degrees for our visualisation. We have already seen in our short review of validity measures that this is a suitable approach, but we can use more information from the clustering output. Validity measures like the fuzzy hypervolume or the (average) partition density also take the distances of the data to the clusters into account. A lot of insight can be gained from a plot of the membership degrees over the distances for each cluster. For each cluster i we plot for every datum x_j a point at (d_{ij}, u_{ij}). This leads to diagrams as they can be seen in figure 4.

How should an ideal graph look like? We would expect high membership degrees for small distances and low membership degrees for large distances. Let us briefly discuss what kind of effects can occur, when cluster centres are not chosen appropriately, although there are valid clusters in the data. Typical problems that occur in this case are the following.

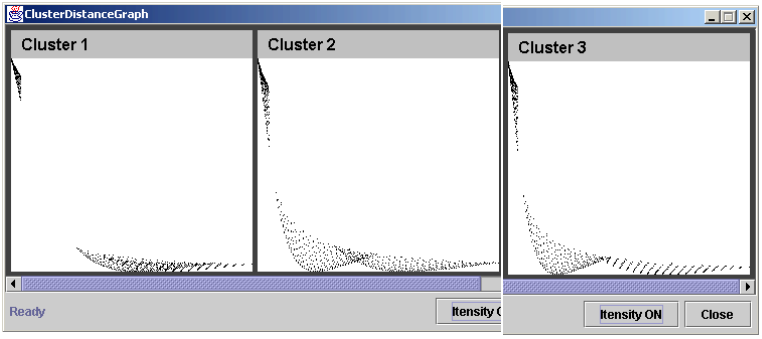

Fig. 4. Membership degrees over distances for three clusters

- One cluster has to cover two or more data clusters. This occurs especially, when the number of clusters is chosen too small. In this case, we would see almost no data with small distances in the upper left part of the diagram for this cluster.
- Two or more clusters compete and share the same data cluster. This usually occurs, when the number of clusters is chosen to high. In this case, there will occur small membership degrees even for small distances. This means, we find points in the lower left part of the diagrams of the corresponding clusters.

Figure 4 shows the corresponding diagrams for our data set, when we use three clusters. For all three clusters the diagrams look quite well. It can even be seen that cluster 1 (covering the lower data cluster in figure 1) is best separated from the other clusters.

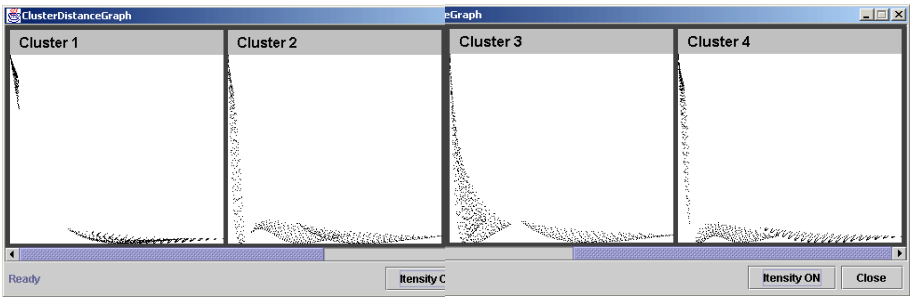

Fig. 5. Membership degrees over distances for four clusters

In figure 5 we have the diagrams for (the inappropriate choice of) four clusters. Cluster 1 corresponds to the lower data cluster in figure 1 and the diagram coincides more or less with the first cluster in figure 4. The diagram

for cluster 4 (the upper left cluster centre in figure 1) is still more or less acceptable. Indeed, we can see from figure 1 that it covers the upper left data cluster almost correctly. However, compared to the diagrams in figure 4 we see a more continuous slide from high to low membership degrees, indicating that the cluster is not very well separated from the others. Cluster 3, represented by the middle cluster centre in figure 1, is the worst one. We have low membership degrees even for small distances, since it shares data with cluster 2 (the right one). There is also a gap in medium values for the membership degrees arising from the fact that cluster 3 actually covers data from two different data clusters. Cluster 2 has also low membership degrees for small distances, because of the competing cluster 3.

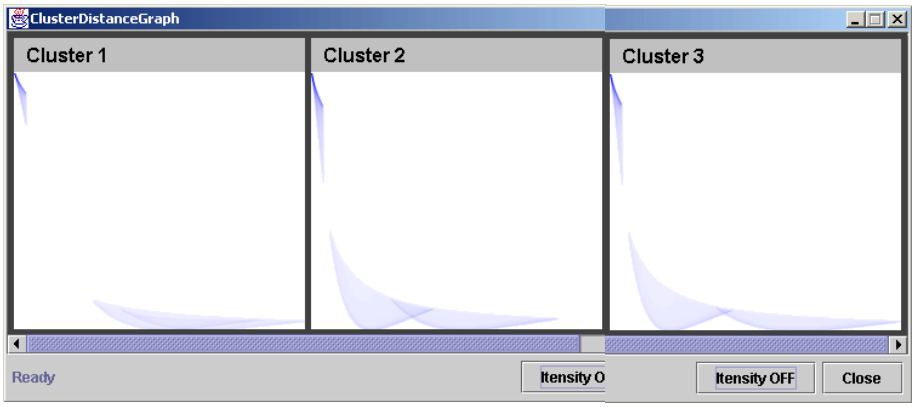

Fig. 6. Intensity plot for membership degrees over distances for a large data set

Analogously to the diagrams showing the maximum membership degrees, we recommend to replace the point plots by intensity plots for larger data sets as shown in figure 6, where we have used again the previously described data set with 192231 data points.

As another example we have used the artificial data set with some noise added (see figure 7).

Figure 8 shows the chart diagram for the membership degree distribution and the maximum membership degree diagram for the noisy data when clustered with three clusters. These diagrams should be compared to the diagrams in figure 2 and 3, respectively, for the data set without noise. The diagrams for the membership degrees over the distances for the noisy data are shown in figure 9 to be compared to figures 4 and 5. The similarities and differences are obvious and need no further explanation.

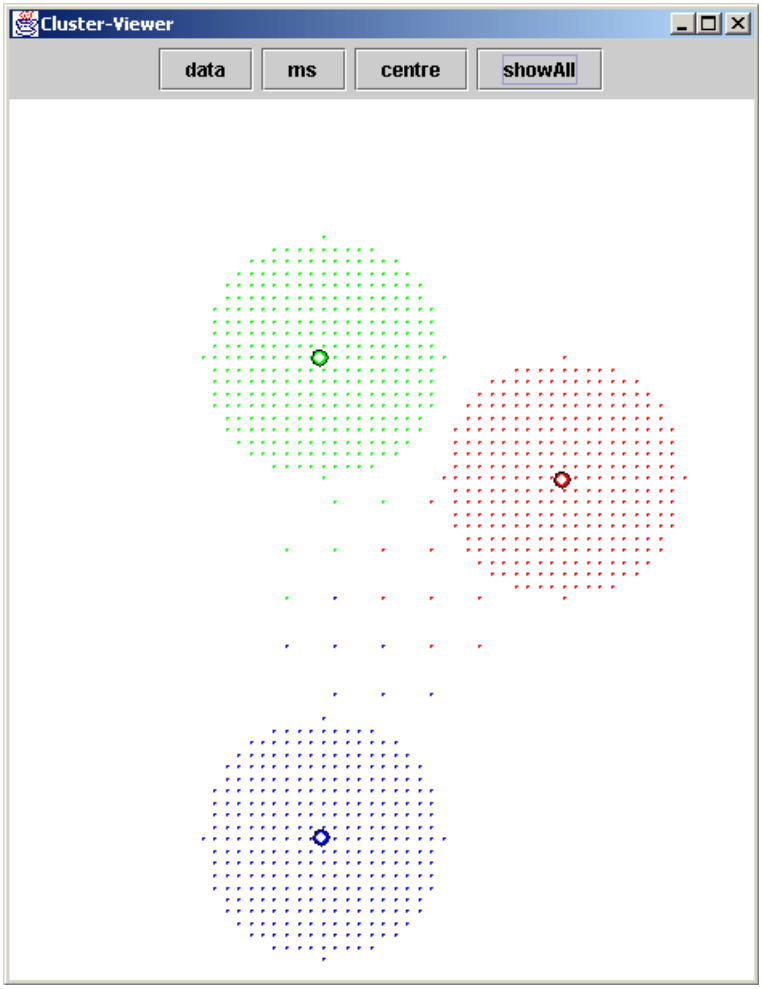

Fig. 7. Clustering result with three clusters for noisy data

5 Conclusions

We have a proposed a number of diagrams that support the visual validation of a fuzzy clustering result. Classical validity measures can be used to carry out clustering completely unsupervised. But this means that we rely on a very strict information compression performed by these validity measures. With our visualisation techniques much more information is available and even single good or bad clusters can be identified by inspecting the diagrams.

Our methods can also be applied in the context of crisp clustering. Of course, all our methods are based on membership degrees. But even if we have

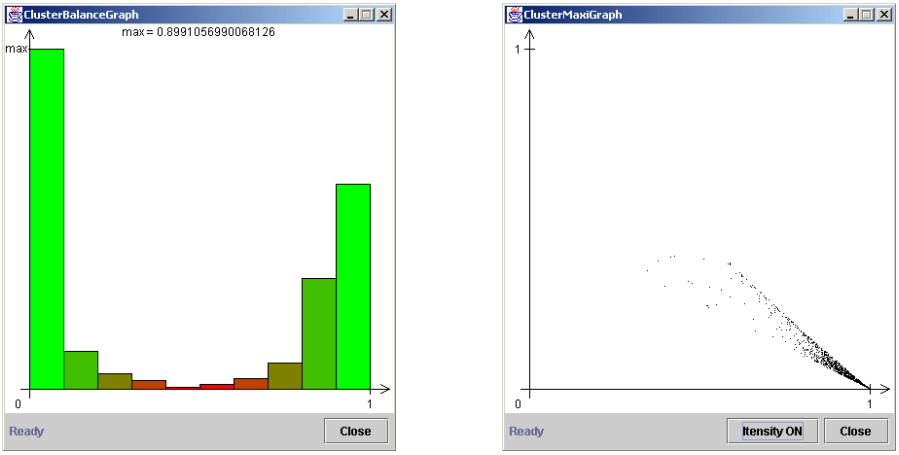

Fig. 8. Scaled chart diagram (left) and maximum membership degrees diagram (right) for noisy data

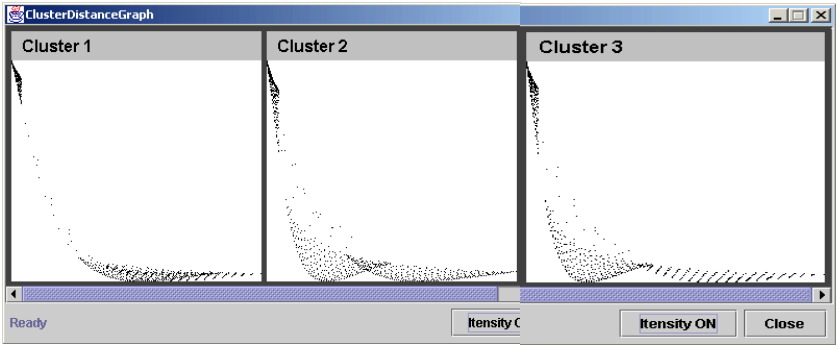

Fig. 9. Membership degrees over distances for noisy data

carried out a crisp clustering, we can afterwards compute membership degrees by the formulae known from fuzzy clustering and then apply our visualisation methods.

References

1. Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum Press
2. Bezdek JC, Keller J, Krishnapuram R, Pal NR (1999) Fuzzy models and algorithms for pattern recognition and image processing. Kluwer, Boston
3. Davé, RN (1991) Characterization and detection of noise in clustering. Pattern Recognition Letters 12: 657–664

4. Gath I, Geva AB (1989) Unsupervised optimal fuzzy clustering. IEEE Trans. Pattern Analysis and Machine Intelligence 11: 773–781
5. Höppner F, Klawonn F, Kruse R, Runkler T (1999) Fuzzy cluster analysis. Wiley, Chichester
6. Krishnapuram R, Keller J (1993) A possibilistic approach to clustering. IEEE Trans. on Fuzzy Systems 1: 98-110
7. Timm H, Borgelt C, Kruse R (2002) A modification to improve possibilistic cluster analysis. IEEE Intern. Conf. on Fuzzy Systems, Honululu (2002)

Index