Shape-Invariant Cluster Validity Indices

Greet Frederix[1] and Eric J. Pauwels[2]

[1] Hogeschool Limburg,
Dept IWT, Universitaire Campus,
Gebouw H, B-3590 Diepenbeek, Belgium
greet.frederix@hogelimb.be
[2] Centre for Mathematics and Computer Science,
CWI Kruislaan 413, 1098SJ Amsterdam, The Netherlands
eric.pauwels@cwi.nl

Abstract. This paper discusses two cluster validity indices that quantify the quality of a putative clustering in terms of label-homogeneity and connectivity. Because the indices are defined in terms of local data-density, they do not favour spherical or ellipsoidal clusters as other validity indices tend to do. A statistics-based decision framework is outlined that uses these indices to decide on the correct number of clusters.

1 Introduction and Notation

Clustering remains one of the mainstays in a large number of pattern recognition and machine learning applications. As a consequence there is no shortage of clustering algorithms and cluster validity indices (see e.g. [3, 4, 5] and references therein). Most of the latter measure various forms of *within versus between* variability and tend to favour clusters that are roughly spherical or ellipsoidal. In this paper we propose two simple *geometric cluster validity indices* that are completely free of such bias. Rather, they try to capture the basic geometric intuition that a cluster is a part of a point-set (in some metric space) which is relatively dense, as well as spatially isolated from the rest of the point-set. This point was also broached in [2, 7] although the solutions proposed in these papers are fundamentally different from the lines we pursue here. This work is an outgrowth of earlier work [6] in which we defined similar indices. However, in this paper we considerably improve on results obtained previously by introducing a methodology for estimating the statistical significance of the index-values.

At this point we should issue a disclaimer. When pursuing research on unsupervised learning, part of the difficulty is due to the fact that there is no clear-cut definition of what exactly a cluster is supposed to be. Our stance in this paper is pragmatic: the aim is to develop criteria that will yield the correct (or at least, an acceptable) clustering in those cases where the "correct" solution is obvious (see Fig. 7 for an example).

We end this section by introducing some notation for future reference. Consider a data set $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ of size n in p-dimensional space (i.e.

P. Permer (Eds.): ICDM 2004, LNAI 3275, pp. 96–105, 2004.

$\mathbf{x}_i \in \mathbb{R}^p$, which we assume equipped with the standard metric). An *L-clustering* is defined by a *labelling function* L which maps each point to its cluster label $g \in \mathcal{G} = \{1, 2, \ldots, K\}$: i.e. $L : \mathcal{D} \longrightarrow \mathcal{G} : \mathbf{x} \longmapsto g$. All points that are mapped to the same cluster label g will be called a *L-cluster*. To avoid confusion, we emphasize that in our notation, *(geometric) clusters* refer to the "real" clusters that are present in the dataset, while L-clusters are created through the (user-proposed) choice of an L-function. It goes without saying that the user's aim is to make sure that the L-clusters coincide with the geometric clusters, but attaining this goal is the essence of the problem.

Finally, we assume that there is (rough) estimate of the data density ϕ defined on $\mathcal{D}$; how this estimate is obtained is irrelevant for the discussion at hand. In fact, the methodology proposed in this paper is very robust with respect to estimate-induced variations of the density. For that reason, a quick-and-dirty density estimate based on nearest neighbours or a kernel estimate will do. We point out that such a rough estimate would not lend itself to clustering based on bump-hunting, as ϕ is likely to grossly under- or over-estimate the number of local extrema.

2 NN-Based Cluster Tension

2.1 Definition

The first cluster validity index captures the idea that clusters are *locally homogeneous* in that neighbouring points tend to have the same label. Put differently, lots of label-variation in the neighbourhood of a large number of points is indicative of poor clustering. To recast this straightforward intuition into a quantitative measure we investigate nearest neighbours. Denote by $N_k(\mathbf{x})$ the set of the k nearest neighbours of $\mathbf{x}$. By convention, the centre-point $\mathbf{x}$ does not belong to $N_k(\mathbf{x})$. We then define the *local diversity* (at $\mathbf{x}$) associated with labelling L by counting the number of neighbours that have a label different from the label at the centre point: $\delta_k(\mathbf{x}; L) = \#\{\mathbf{y} \in N_k(\mathbf{x}) \,|\, L(\mathbf{y}) \neq L(\mathbf{x})\}/k$. The (global) *NN-tension* (induced by the cluster labelling L) is then obtained by computing the weighted average over all data points:

$$T_k(L) = \frac{1}{n} \sum_{i=1}^{n} \delta_k(\mathbf{x}_i; L)\phi(\mathbf{x}_i). \tag{1}$$

The rationale for including the density as a weight-factor is that label diversity in high density regions is more significant as a contra-indication for good clustering. Notice that the number k of nearest neighbours is in fact a sort of scale-parameter, determining the size of the smallest clusters that can be picked up. In our experiments we took k to be equal to 5% of the dataset size.

The way this validity index can be used needs little explanation: Consider the case where a labelling L erroneously splits a single geometric cluster into two L-clusters (say A and B, see Fig. 1, top row). This will then give rise to a relatively high local NN-tension along the "faultline" separating A and B,

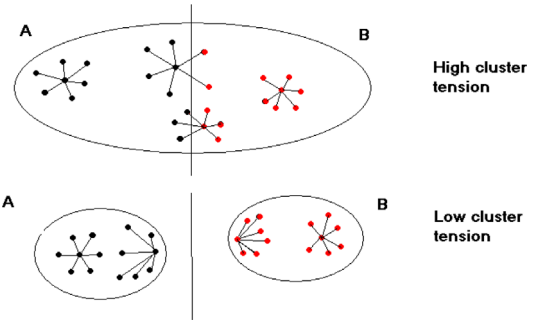

Fig. 1. *Top:* A single geometric cluster is erroneously split into two L-clusters A and B. This gives rise to relatively high NN-tension along the "faultline" separating A and B. *Bottom:* When the L-clusters A and B correspond to geometric clusters, the NN-tension is low or even zero

which is translated in a relatively high global NN-tension. Contrast this to the case where the L-clusters do in fact correspond to genuine geometric clusters (Fig. 1 bottom). Now NN-tension will be low throughout the dataset, keeping the global NN-tension low.

These simple considerations reinforce our original intuition: "high" tension is indicative of erroneous mergers, while "low" tension bolsters our confidence in the validity of the proposed labelling. However, in order to turn this qualitative appreciation into an operational decision criterion, we need some value for the *typical value* and *expected variability* of the NN-tension. Have another look at Fig. 1 where in both cases the putative clustering is indicated by a vertical line. Now, imagine that the proposed clustering was in fact generated by a *horizontal* line cutting both data-sets into two (approximately equal sized) groups. It is clear that for the top data-set the resulting tension would be comparable to the original one. For the second data-set (bottom) however, such a split would generate a tension which is significantly higher than the original one (which was probably close, or equal, to zero). This leads quite naturally to the following construction: To decide on the acceptability of the NN-tension $T^{(0)} \equiv T(L_0)$ associated with a putative cluster labelling L_0, we generate R random cluster labellings L_r $(r = 1, \ldots, R)$ by separating the data using R random hyperplanes (i.e. hyperplane through a random data-point, and orthogonal to a random direction) and compute the corresponding tensions $T^{(r)} \equiv T(L_r)$. Next, estimate the p-value of $T^{(0)}$ with respect to the set $\{T^{(1)}, \ldots, T^{(R)}\}$, e.g. by computing the fraction of this set that is smaller than $T^{(0)}$, i.e. $p(T^{(0)}) = \#\{T^{(r)} \,|\, T^{(r)} \leq T^{(0)}/R$. We will now conclude that the proposed clustering is acceptable if this p-value is exceptionally small (e.g. $p < 0.05$, or even $p < 0.01$). We refer to Fig. 4 for a illustrative example.

2.2 Adjusting for Variations in Shape

Although definition (1) of global tension seems natural and straightforward to use, closer examination reveals a problem which is highlighted by the follow-

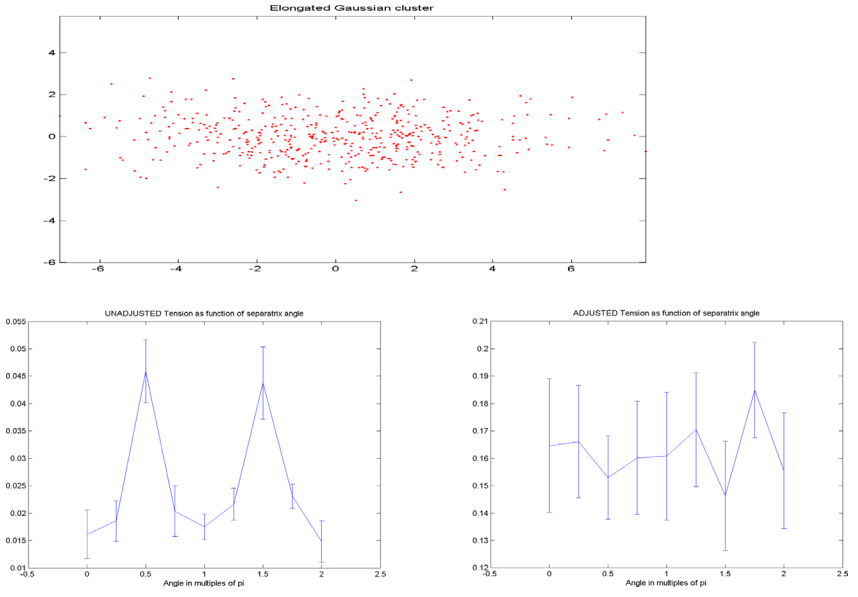

Fig. 2. *Top:* Elongated Gaussian cluster illustrating the dependency of the total tension on the split direction. *Bottom, left:* The unadjusted cluster tension (1) clearly shows a dependency on the angle of the separatrix, with low tension at angles where the separatrix is horizontal (i.e. angle $= 0, \pi, 2\pi$), and high tension when the separatrix is vertical (i.e. angle $= \pi/2$ and $3\pi/2$). For each angle, the curve shows the mean and standard deviation for 10 independent resamplings of the Gaussian cluster. This is also the reason why the results for 0, π and 2π are not identical! *Bottom, right:* The adjusted cluster tension (2) no longer exhibits such a systematic cyclic trend

ing simple example. Consider an elongated Gaussian 2-dim cluster, centered at the origin and positioned such that its long axis coincides with the x-axis (see Fig. 2). Now, suppose that this cluster has been split into two by using the k-means algorithm (for $k = 2$). A moment's thought will convince the reader that k-means will generate a separatrix (line separating the two groups) which is approximately vertical and passes through the origin (i.e. the cluster's centre of gravity). This entails that the "faultline" is relatively short, especially when compared to the separatrix that would result from a horizontal split (i.e. when this separatrix would coincide with the x-axis). As a consequence, a horizontal split will result in more points straddling the faultline and since eq.(1) averages over *all* points, this will result in higher average tension. This is detrimental to the computation of the p-value as proposed in the previous section, as it means that the original tension $T^{(0)}$ is systematically smaller than the ones that are obtained from random cuts that are not vertical. Phrased in statistical terminology, the original tension is *biased towards small values,* resulting in an artificially low p-value, even if the underlying cluster is compact (but elongated).

To remove this bias, we adjust definition (1) by restricting the averaging to all points that have *non-vanishing local tension* (we add one to the denominator to avoid potential *division by zero* problems):

$$T_k(L) = \frac{1}{N_p + 1} \sum_{i=1}^{n} \delta_k(\mathbf{x}_i; L)\phi(\mathbf{x}_i), \quad \text{where} \quad N_p = \#\{\mathbf{x}_i \,|\, \delta_k(\mathbf{x}_i; L) > 0\} \quad (2)$$

Since the denominator is now proportional to the number of points that effectively contribute to the total tension, this compensates for the shape bias, as illustrated in Fig.2 (bottom, right).

At first sight, this modification might seem to emasculate the criterion. Indeed, consider the extreme case where you have two 2-dimensional standard normal Gaussian cluster (labeled 1 and 2) which are well separated (e.g. by a distance of 10 say). Furthermore, assume there is a single point right in the middle between these two clusters. Clearly, because the Gaussians are well separated all points in these two clusters will only have neighbours with the same label (hence no tension). Whatever the label of the point in the middle is, it will have approximately half of its neighbours in the first Gaussian, and half of them in the second, resulting in a local diversity of approximately 0.5. Since eq. (2) now averages over a single point, one could get the impression that tension will be relatively high. However, remember that local tension is obtained by multiplying diversity and data-density. Clearly, the density at the isolated single point will be very low, resulting in a low value for the adjusted tension, as intended.

2.3 Experiments

The following experiments are meant to illustrate the potential of NN-based tension. In a first experiment, we generate two standard normal 5-dimensional Gaussian clusters (of 700 points each), such that the centra are a distance of 6 apart. To create two L-clusters we then generate a hyperplane orthogonal to the line connecting the two centra: points on either side of this hyperplane get different labels (say 1 and 2). The position of the hyperplane is then systematically shifted along the line: distance 0 means that it cuts through the centre of the first Gaussian, while distance 6 indicates that it cuts through the centre of the second Gaussian. Clearly, the optimal position corresponds to distance 3 when it cuts the connecting line exactly at the midpoint. For each configuration the associated tension is generate.

The results (shown in Fig. 3) confirm our intuition. The tension attains its highest values when the separatrix hyperplane slices straight through the cluster centres, whereas the minimum is obtained around the midpoint (as expected). It's also interesting to observe that these results are very robust with respect to the estimated density (ϕ in eq.(2)): In one case the density was estimated using the distance to the nearest neighbours, while the other result is based on the exact Gaussian density (from which the samples were drawn). Both graphs are almost identical, buttressing the point we made earlier.

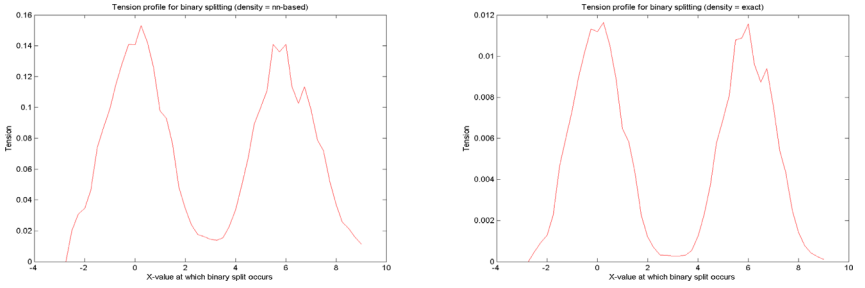

Fig. 3. Evolution of the NN-based tension T as a function of the position splitting hyperplane (creating the L-clustering: see main text for detailed description). *Left:* Tension based on the NN-based estimate for density ϕ. *Right:* Tension based on the exact version of density ϕ

In a second experiment, we tested the discriminatory power of the tension measure by comparing the tension (and the corresponding p-values) for cluster configurations for which the answer should be obvious. More precisely, we generated two (2-dimensional standard-normal) Gaussian clusters at a distance d apart. For small values of d (e.g. $d = 1$) the two Gaussians merge into a single geometric cluster (i.e. an independent observer who has no knowledge about the underlying generative process would judge it to be one cluster, see Fig. 4, top). Increasing the distance will progressively pry the Gaussians apart, up to a point (e.g. for $d = 5$) where two constituent clusters are clearly discernable (see Fig. 4, bottom). Typical results for the p-values of the original labelling relative to simulated labellings are shown in the right column of Fig. 4.

3 Connectivity Index

The second geometric cluster validity measure we discuss is the so-called *connectivity index*. This index captures the intuition that any two points in a cluster can be connected by a *high density path*, i.e. a path which at no point needs to traverse a "void". To fix ideas we will start with a simple setup (illustrated in Fig. 5). On the left, we have two geometric clusters (A and B) which we assume are recognized as such by the labelling. If we now pick random pairs of points in either cluster (yielding a and b, a' and b' respectively) and evaluate the cluster density ϕ at the midpoints m and m', we'll get a relatively high value as these points tend to be situated at high density locations. Contrast this to the situation on the right in Fig. 5 where we assume that the labelling L has erroneously lumped the two geometric clusters A and B in one big L-cluster. As a consequence, when we are drawing pairs of random points a and b (from the same L-cluster) there will be a significant fraction of pairs for which these points are part of different geometric clusters (illustrated by the points a' and b'; if A and B have comparable size this will occur with an approximate probability of

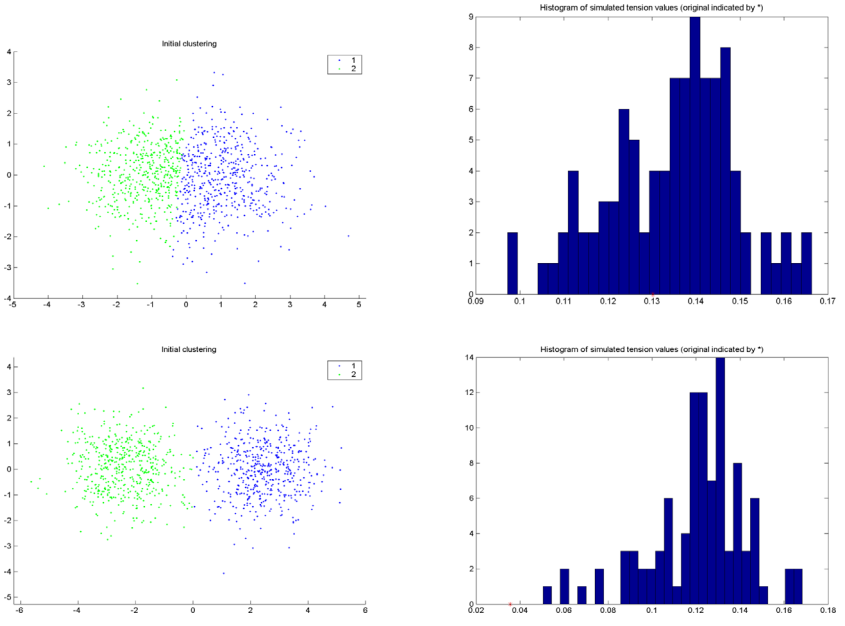

Fig. 4. *Top left:* Single geometric cluster erroneously split by vertical midline, resulting in an NN-tension $T_0 = 0.13$. *Top right:* Histogram for NN-tensions generated by 100 random clusterings. The p-value for T_0 equals 0.38 (large!), indicating that the putative split is erroneous. *Bottom left:* Two genuine geometric clusters for which $T_0 = 0.036$. *Bottom right:* Histogram for NN-tensions generated by 100 random clusterings. The p-value for T_0 now is less than 0.01 (small!), confirming the appropriateness of the proposed split

0.5). In such an event, the midpoint m' will likely fall in the void between the two clusters, and therefore register a low density $\phi(m')$.

We are now in a position to provide a formal definition for the connectivity index C associated with a labelling L:

1. Draw from the dataset K random pairs of points (a_k, b_k), making sure that the points in each pair belong to the same L-cluster, i.e. $\forall i : L(a_k) = L(b_k)$;
2. Construct for each pair the corresponding midpoint m_k and evaluate the data-density ϕ at that point. The connectivity index C is then defined to be the average over all random pairs:

$$C(L) = \frac{1}{K} \sum_k \phi(m_k). \tag{3}$$

To render this index really useful, we need to make it slightly more robust. The right panel in Fig. 5 shows what needs to be done: When confronted with a curved cluster, the chances are that the midpoint will fall in a "convexity void" and therefore return an underestimate for the density. If we allow m to hill-climb

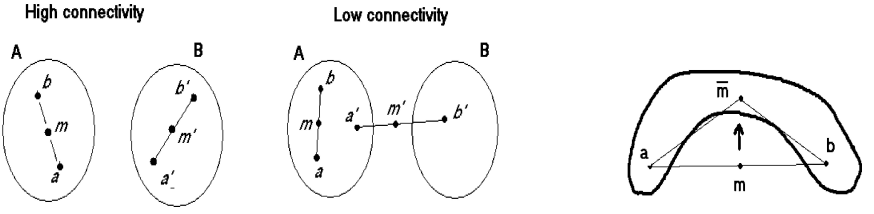

Fig. 5. Illustrating the definition of the connectivity index. *Left and middle:* Basic definition of connectivity index; see text for more details. *Right:* Better value for connectivity index is obtained by measuring the density along a high density ridge obtained by hill-climbing from m to $\bar{m}$ (constrained by the condition $d(a, \bar{m}) \approx d(b, \bar{m})$))

towards the high density point $\bar{m}$, (keeping $\bar{m}$ approximately in the "middle" of the anchor points a and b by insisting that $d(a, \bar{m}) \approx d(b, \bar{m})$), we get a more representative density estimate. Sensitivity of this index can be further increased by repeating the procedure for the midpoints between a and $\bar{m}$, and $\bar{m}$ and b (which then contribute to the mean in eq.(5)). Basically, this means that we are estimating the density along a "snake" connecting a and b, which is attracted by the high density ridge.

To illustrate the validity of this concept, have another look at the data-sets in Fig. 4, but this time ignore the different labels, i.e. assume that each data-set is considered to be one L-cluster. In that case, the top dataset should have high connectivity, while the lower dataset should score a significantly lower value due to the gap between the two geometric clusters. This is borne out by the tests reported in Fig. 6 where the histogram shows the result for 200 computed connectivity indices, 100 for each dataset. They nicely cluster in two groups of 100 indices each: the left group corresponds to the low connectivity indices obtained for the lower dataset (consisting of two disconnected geometric clusters), while the right group comprises the higher connectivity indices obtained for the upper dataset in Fig. 4.

We conclude this section by indicating how the connectivity index is used in practice. To fix ideas, consider once again the case depicted in Fig. 5(top). Since there are indeed two geometric clusters, we expect the connectivity index to be *significantly lower* when we assume that both clusters are lumped together in a single L-cluster. To quantify what should be considered "significant", we proceed as follows. Recall that C as defined in eq. 3 is an average over a random sample of K paired anchor points. We can therefore easily resample and compute another instance of this parameter. Let $C_i^{(1)}$ denote the i^{th} realisation of the C-index assuming that both clusters are lumped together in a single L-cluster, while $C_j^{(2)}$ denotes similar results under the assumption that the clusters are different. Allowing both i and j to run over r repeats, we subsequently compute the corresponding means ($M^{(1)}$ and $M^{(2)}$) and standard deviations ($S^{(1)}$ and $S^{(2)}$). To test whether $M^{(1)}$ is indeed much lower than $M^{(2)}$, we apply a standard T-test and evaluate whether the Student t-statistic

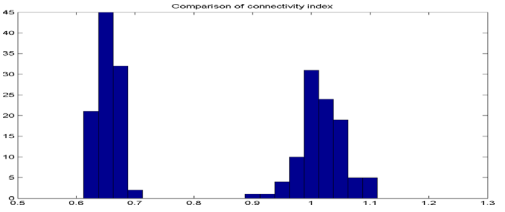

Fig. 6. Histograms for 200 connectivity indices; see main text for more details

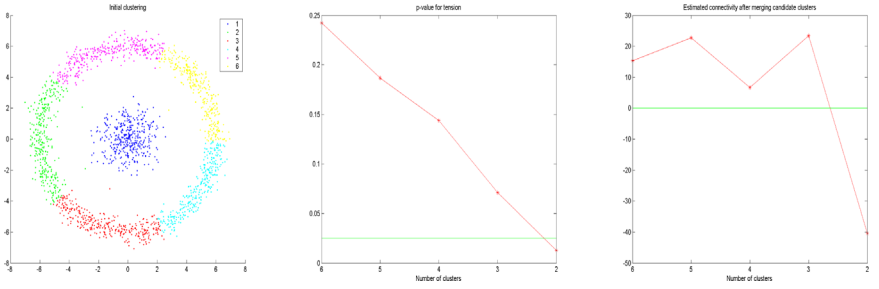

Fig. 7. Illustration of cluster selection based on both geometric cluster validity indices. *Left:* Complicated dataset initially divided up in 6 groups using k-means clustering. These groups are then systematically merged, such that each merger maximizes the reduction in NN-tension. The number of groups is thus stepwise reduced until finally two clusters remain: the ring and the central core. *Middle:* This graph charts the evolution of the p-value for the NN-tension during the evolution from 6 to 2 clusters. The final p-value (for 2 clusters) is exceptionally low ($p < 0.025$), indicating that the NN-tension for 2 clusters is exceptionally low. *Right:* This graph shows the evolution of the V-parameter (defined in section 3) for the connectivity index. Again, we see that this measure drops below the 0-threshold when two clusters remain. Both indices therefore agree on *two* being the correct number of clusters

$$t = \frac{M^{(1)} - M^{(2)}}{\sqrt{((S^{(1)})^2 + (S^{(2)})^2)/r}}$$

is less than -2 (which corresponds to an approximate p-value of 0.025). For convenience we introduce a new parameter $V = t + 2$, such that $V < 0$ flags that further mergers are contra-indicated.

A final illustration is provided in Fig. 7 were we applied **both** validity criteria on a challenging data-set in which points are distributed over a central cluster and a surrounding ring. An initial k-means clustering (with k=6) returns 6 groups, five of which are situated in the ring. Reducing the number of clusters in a stepwise fashion by merging the neighbouring clusters whose unification produces the largest reduction in NN-tension, finally yields two clusters (the ring and the central core). At this point, both validity indices indicate that

further reduction of the number of clusters is contra-indicated, thus confirming that two is the correct number of clusters.

Acknowledgments. This work was partially supported by EC Project FOUNDIT (IST-2000-28427) and EC Network of Excellence MUSCLE (FP6-507752).

References

1. J.C. Bezdek, J. Keller, R. Krisnapuram, N.R. Pal: Fuzzy Models and Algorithms for Pattern Recognition in Image Processing, Kluwer Academic Publishers, 1999.
2. J.C. Dunn Well Separated Cluserds and Optimal Fuzzy Partitions, Journal of Cybernetics, **Vol. 4**, (1974), 95–104.
3. A.K. Jain, M.N. Murty, P.J. Flynn Data Clustering: A Review, ACM Computing Surveys, **Vol. 31**, No. 3, (Sept 1999), 264-323.
4. K. Jajuga, A. Sokolowski and H. Bock, Classification, Clustering and Data Analysis, IFCS Conference Poland, Springer, 2002.
5. L. Kaufman and P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, J. Wiley and Sons, 1990.
6. E.J. Pauwels and G. Frederix, Finding Salient Regions in Images, Computer Vision and Image Understanding, **Vol. 75**, No. 1/2, (July/August 1999), 73-85.
7. X.L. Xie and G. Beni: A Validity Measure for Fuzzy Clustering, IEEE Trans on Pattern Analysis and Machine Intelligence, **Vol. 13**, No. 8, (August 1991), 841-847