A New Index of Cluster Validity

Mu-Chun Su

Major Difficulties

- The presence of large variability in cluster geometric shapes, and
- The number of clusters cannot always be known a priori. Different distance measures lead to different types of clusters (e.g. compact hyperspheres, compact hyperellipsoids, lines, shells, etc.).

Cluster Validity (1)

- In fact, if cluster analysis is to make a significant contribution to engineering applications, much more attention must be paid to cluster validity issues that are concerned with determining the optimal number of clusters and checking the quality of clustering results.
- Many different indices of cluster validity have been proposed, such as the Bezdek's partition coefficient, the Dunn's separation index, the Xie-Beni's separation index, Davies-Bouldin's index, and the Gath-Geva's index, etc.
- Most of these validity indices usually assume tacitly that data points having constant density to the clusters. However, it is not sure of the real problems.

Indices of Cluster Validity (1)

- Cluster validation refers to procedures that evaluate the clustering results in a quantitative and objective function.
- Some kinds of validity indices are usually adopted to measure the adequacy of a structure recovered through cluster analysis.
- Determining the correct number of clusters in a data set has been, by far, the most common application of cluster validity.

Indices of Cluster Validity (2)

- In general, indices of cluster validity fall into one of three categories.
- Some validity indices measure partition validity by evaluating the properties of the crisp structure imposed on the data by the clustering algorithm.
- In the case of fuzzy clustering algorithms, some validity indices such as partition coefficient and classification entropy use only the information of fuzzy membership grades to evaluate clustering results.
- The third category consists of validity indices that make use of not only the fuzzy membership grades but also the structure of the data.

Dunn’s index

- The Dunn’s index is defined as

$$DI(c) = \min_{i \in c} \left\{ \min_{j \in c, j \neq i} \left\{ \frac{\delta(A_i, A_j)}{\max_{k \in c} \left\{ \Delta(A_k) \right\}} \right\} \right\}$$

where

$$\delta(A_i, A_j) = \min \left\{ d(\underline{x}_i, \underline{x}_j) \mid \underline{x}_i \in A_i, \underline{x}_j \in A_j \right\}$$

$$\Delta(A_k) = \max \left\{ d(\underline{x}_i, \underline{x}_j) \mid \underline{x}_i, \underline{x}_j \in A_i \right\}$$

d is a distance function and A_j is the set whose elements are the data points assigned to the *i*th cluster.

The main drawback with direct implementation of Dunn’s index is computational since calculating becomes computationally very expensive as *c* and *n* increase.

Davies-Bouldin’s Index (1)

- Its major difference from Dunn’s index is that it considers the average case by using the average error of each class.
- This index is a function of the ratio of the sum of within-cluster scatter to between-cluster separation, it uses both the clusters and their sample means.
- First, define the *within ith cluster scatter and the between ith and jth cluster as*

$$S_{i,q} = \left(\frac{1}{|A_i|} \sum_{\underline{x} \in A_i} \| \underline{x} - \underline{v}_i \|_2^q \right)^{1/q}$$

$$d_{ij,t} = \left\{ \sum_{s=1}^{p} | v_{si} - v_{sj} |^t \right\}^{1/t} = \| \underline{v}_i - \underline{v}_j \|_t$$

Davies-Bouldin Index (2)

- where $\underline{v}_i$ is the *i*th cluster center, $q,t \geq 1$, q is an integer and q,t can be selected independently of each other. $|A_i|$ is the number of elements in A_i
- Next, define

$$R_{i,qt} = \max_{j \in c, j \neq i} \left\{ \frac{S_{i,q} + S_{j,q}}{d_{ij,t}} \right\}$$

- Finally, the Davies-Bouldin index can be defined as

$$DB(c) = \frac{1}{c} \sum_{i=1}^{c} R_{i,qt}$$

Partition Coefficient (PC)

- Bezdek designed the partition coefficient (PC) to measure the amount of “overlap” between clusters.
- He defined the partition coefficient (PC) as follows.

$$PC(c) = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}(u_{ij})^2$$

- where u_{ij} $(i = 1, 2, \ldots, c\ ;\ j = 1, 2, \ldots, N)$ is the membership of data point *j* in cluster *i*.
- Disadvantages of the partition coefficient are its monotonic decreasing with *c* and the lack of direct connection to some property of the data themselves.

Classification Entropy (CE)

- Classification entropy is defined as

$$CE(c) = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij} \log(u_{ij})$$

- Bezdek proved the relation $0 \leq 1 - PC(c) \leq CE(c)$ for all probabilistic cluster partitions *c*.
- It is basically a measure for the fuzziness of the cluster partition only, which is similar to the partition coefficient.

Separation index (S)

- The validity index S is based on the objective function *J* by determining the average number of data and the square of the minimum distances of the cluster centers. The separation index S is defined as

$$S(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij}{}^{2} d(\underline{x}_j - \underline{v}_i)^2}{N \min\limits_{\substack{m,n=1,\ldots,c \\ and\ \ m\neq n}} \{d(\underline{v}_m - \underline{v}_n)\}^2} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij}{}^{2} d(\underline{x}_j - \underline{v}_i)^2}{N * (d_{\min})^2}$$

- where $d_{\min}$ is the minimum Euclidean distance between cluster centers.
- The more separate the clusters, the larger $(d_{\min})^2$, and the smaller *S*. Thus, the smallest *S* indeed indicates a valid optimal partition.

Fuzzy Hypervolume (FHV)

- It can be argued that a good partition should yield a high value of fuzzy partition density and a low value of fuzzy hypervolume. Gath and Geva defined the volume of the clusters in a fuzzy partition as follows.

$$FHV(c) = \sum_{i=1}^{c} \sqrt{\det(F_i)}$$

- where F_i denotes the ith fuzzy covariance matrix.

CS Index

- The CS index is then defined as

$$CS(c)=\frac{\frac{1}{c}\sum_{i=1}^{c}\left\{\frac{1}{|A_i|}\sum_{\underline{x}_j\in A_i}\max_{\underline{x}_k\in A_i}\left\{d(\underline{x}_j,\underline{x}_k)\right\}\right\}}{\frac{1}{c}\sum_{i=1}^{c}\left\{\min_{j\in c,\, j\neq i}\left\{d(\underline{v}_i,\underline{v}_j)\right\}\right\}}=\frac{\sum_{i=1}^{c}\left\{\frac{1}{|A_i|}\sum_{\underline{x}_j\in A_i}\max_{\underline{x}_k\in A_i}\left\{d(\underline{x}_j,\underline{x}_k)\right\}\right\}}{\sum_{i=1}^{c}\left\{\min_{j\in c,\, j\neq i}\left\{d(\underline{v}_i,\underline{v}_j)\right\}\right\}}$$

Four Spherical Clusters (1)

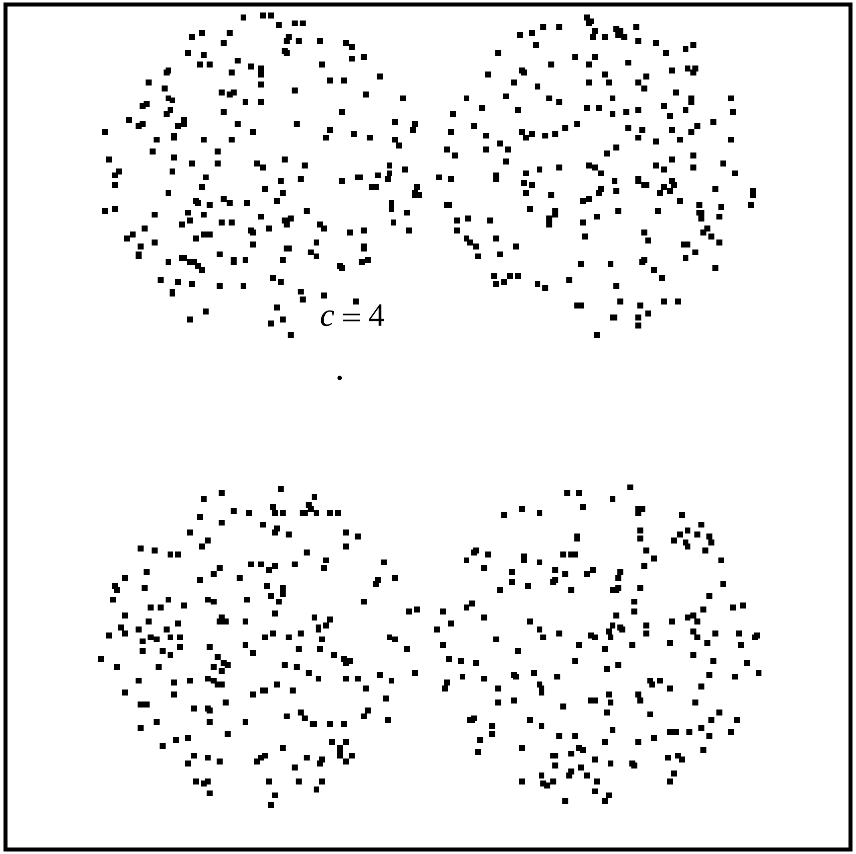

Fig. 3(a). The data set in example 1: It contains of a mixture of compact spherical and ellipsoidal clusters

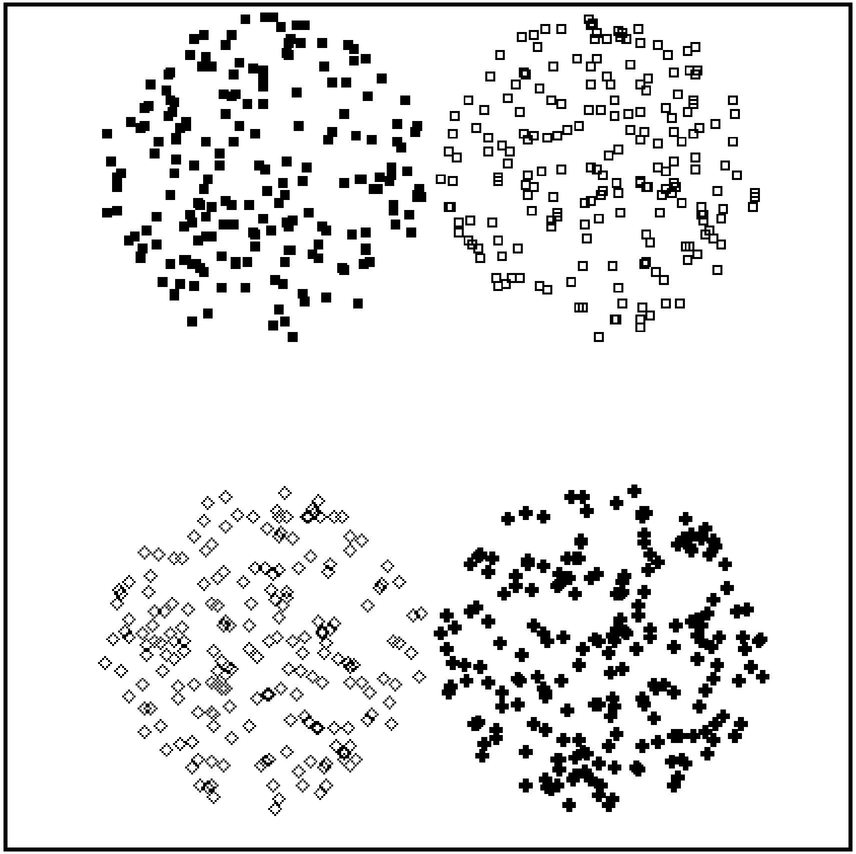

Fig. 3(b). The final clustering result achieved by the FCM algorithm at

Four Spherical Clusters (2)

c	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
DI	**0.236**	0.038	0.078	0.008	0.013	0.029	0.024	0.012	0.020
DB	0.800	**0.618**	0.624	0.767	0.955	0.972	0.800	0.869	0.916
PC	**0.779**	0.703	0.720	0.651	0.598	0.560	0.541	0.522	0.511
CE	**0.365**	0.546	0.569	0.719	0.847	0.935	1.037	1.079	1.112
S	0.162	0.165	**0.080**	0.176	0.351	0.255	0.250	0.187	0.184
FHV	1.124	1.047	**0.780**	0.851	0.927	0.929	1.075	0.975	0.901
CS	1.058	0.874	**0.771**	0.967	1.132	1.316	1.031	1.20	1.246

A Mixture of Spherical and Ellipsoidal Clusters (1)

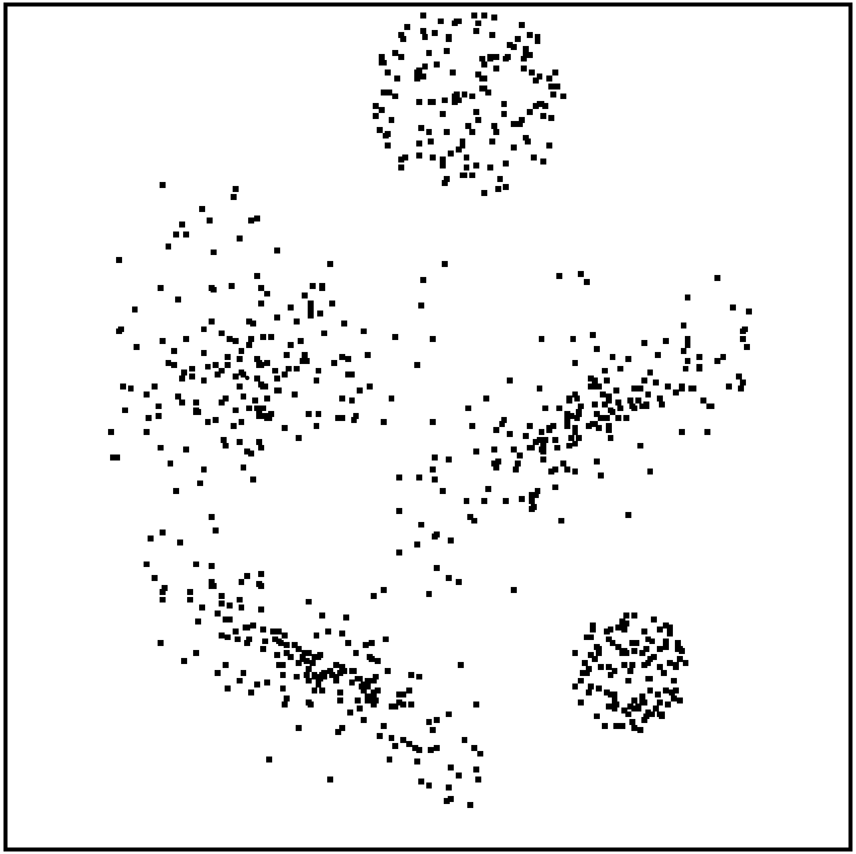

Fig. 4.(a) The data set in example 2: It contains five compact clusters.

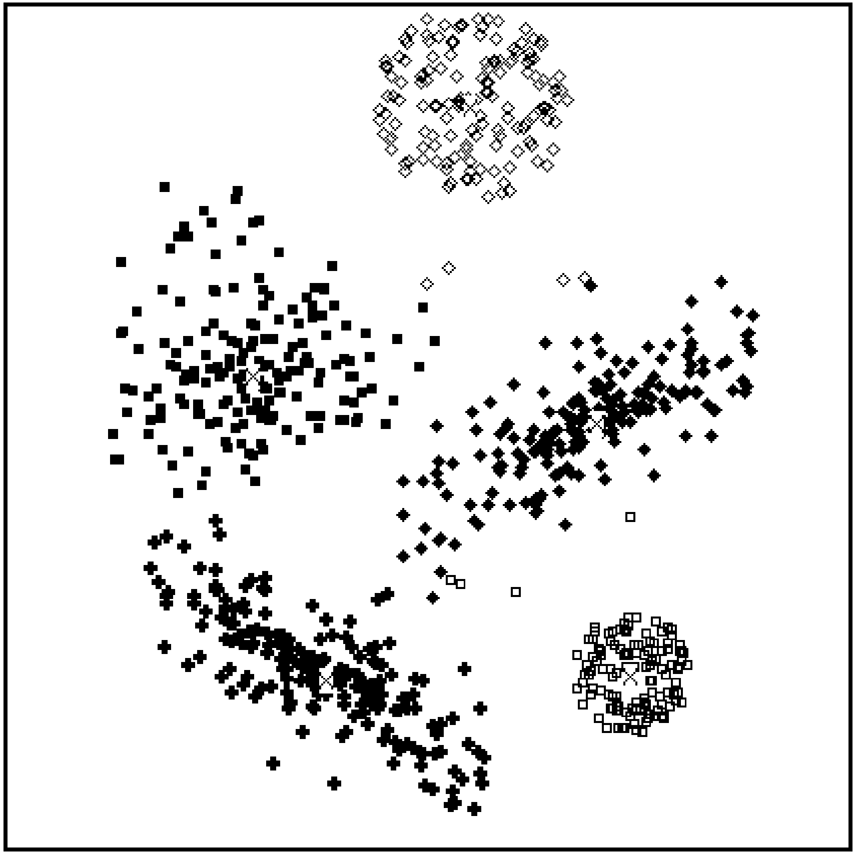

Fig. 4(b). The final clustering result achieved by the Gustafson-Kessel algorithm at

A Mixture of Spherical and Ellipsoidal Clusters (2)

c	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
DI	**0.034**	0.016	0.015	0.022	0.012	0.004	0.006	0.004	0.008
DB	1.265	0.958	0.715	**0.501**	0.744	0.924	0.821	0.960	1.086
PC	**0.835**	0.783	0.755	0.780	0.731	0.689	0.654	0.625	0.591
CE	**0.287**	0.408	0.498	0.483	0.592	0.688	0.771	0.834	0.911
S	0.398	0.269	0.153	**0.082**	0.383	0.497	0.296	0.315	0.796
FHV	1.858	1.570	1.253	**0.921**	1.044	1.061	1.073	1.055	1.083
CS	1.758	1.517	1.035	**0.866**	1.099	1.369	1.156	1.476	1.790

Five Clusters (1)

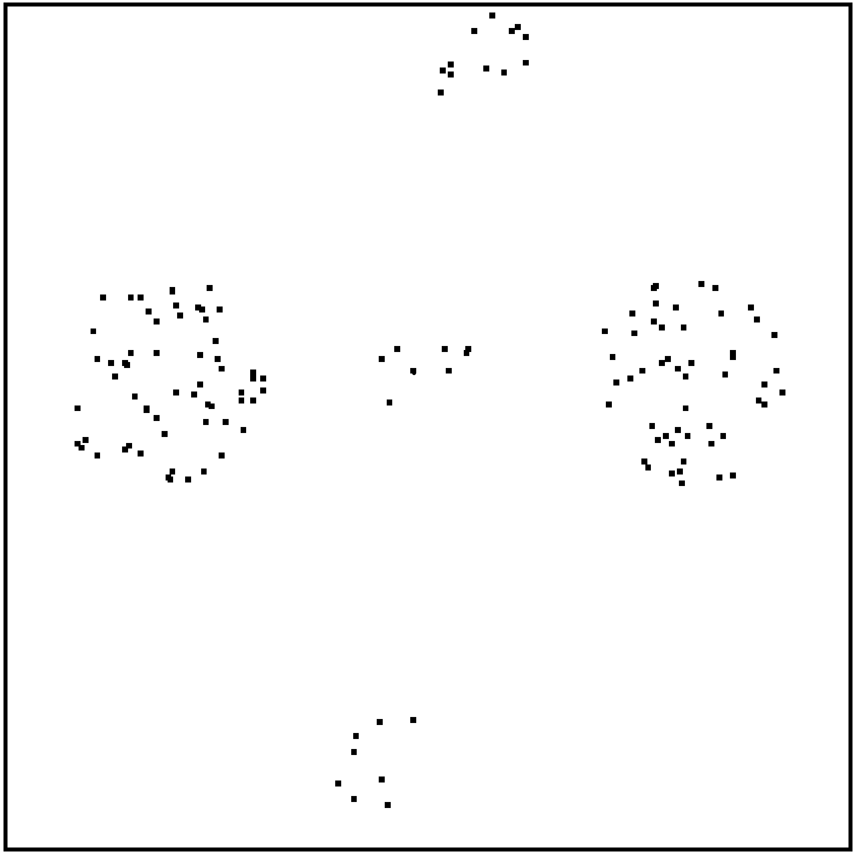

Fig. 5(a). The data set in example 3: It contains distributed on five clusters

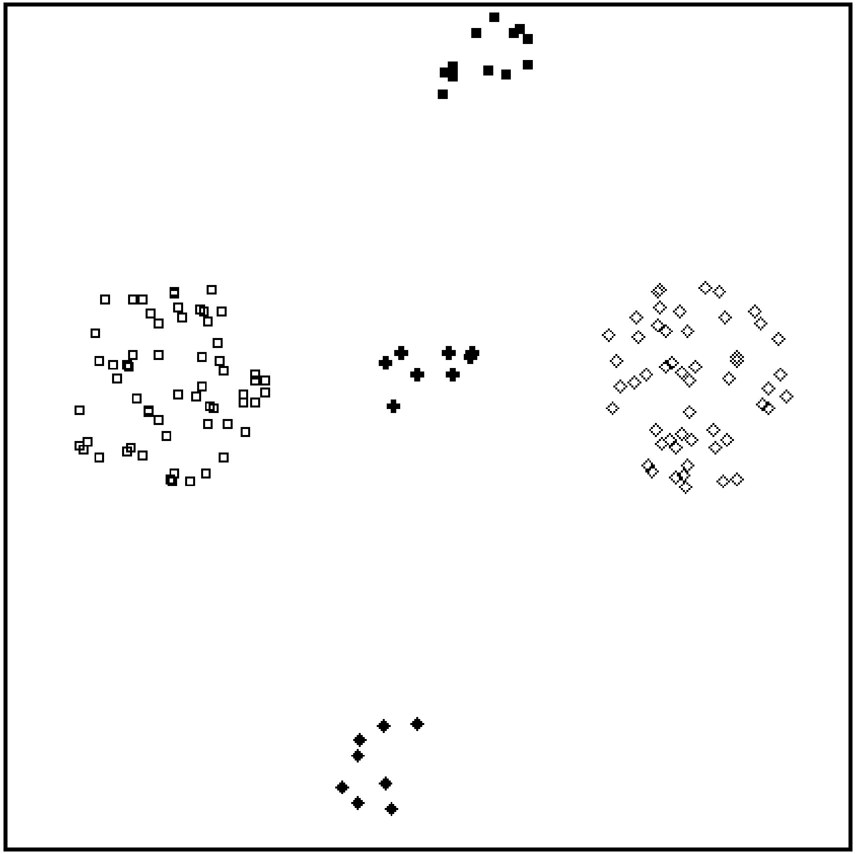

Fig. 5(b). The final clustering result achieved by the FCM algorithm at

Five Clusters (2)

c	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
DI	0.062	0.069	0.105	**0.588**	0.047	0.040	0.040	0.092	0.078
DB	0.591	0.435	**0.316**	0.354	0.845	0.731	0.673	0.595	0.693
PC	**0.866**	0.827	0.838	0.786	0.676	0.684	0.671	0.669	0.650
CE	**0.222**	0.340	0.362	0.480	0.631	0.621	0.700	0.722	0.773
S	0.090	0.074	**0.039**	0.063	0.310	0.204	0.169	0.124	0.873
FHV	1.957	1.725	1.072	0.925	0.933	0.754	0.725	**0.653**	0.751
CS	1.022	0.589	0.428	**0.396**	0.782	0.594	0.566	0.552	0.682