Pattern Analysis and Machine Intelligence

Lecture Notes on Clustering (IV)
2011-2012

Davide Eynard

`eynard@elet.polimi.it`

Department of Electronics and Information

Politecnico di Milano

Course Schedule

Date	Topic
07/05/2012	Clustering I: Introduction, K-means
14/05/2012	Clustering II: K-M alternatives, Hierarchical, SOM
21/05/2012	Clustering III: Mixture of Gaussians, DBSCAN, J-P
28/05/2012	Clustering IV: Spectral clustering, evaluation measures

Lecture outline

- Cluster Evaluation
  - Internal measures
  - External measures
- Finding the correct number of clusters
- Framework for cluster validity

Cluster Evaluation

- Every algorithm has its pros and cons
  - (Not only about cluster quality: complexity, #clusters in advance, etc.)
- For what concerns cluster quality, we can *evaluate* (or, better, **validate**) clusters
- For supervised classification we have a variety of measures to evaluate how good our model is
  - Accuracy, precision, recall
- For cluster analysis, the analogous question is: *how can we evaluate the "goodness" of the resulting clusters?*
- But most of all... ***why*** *should we evaluate it?*

Cluster found in random data

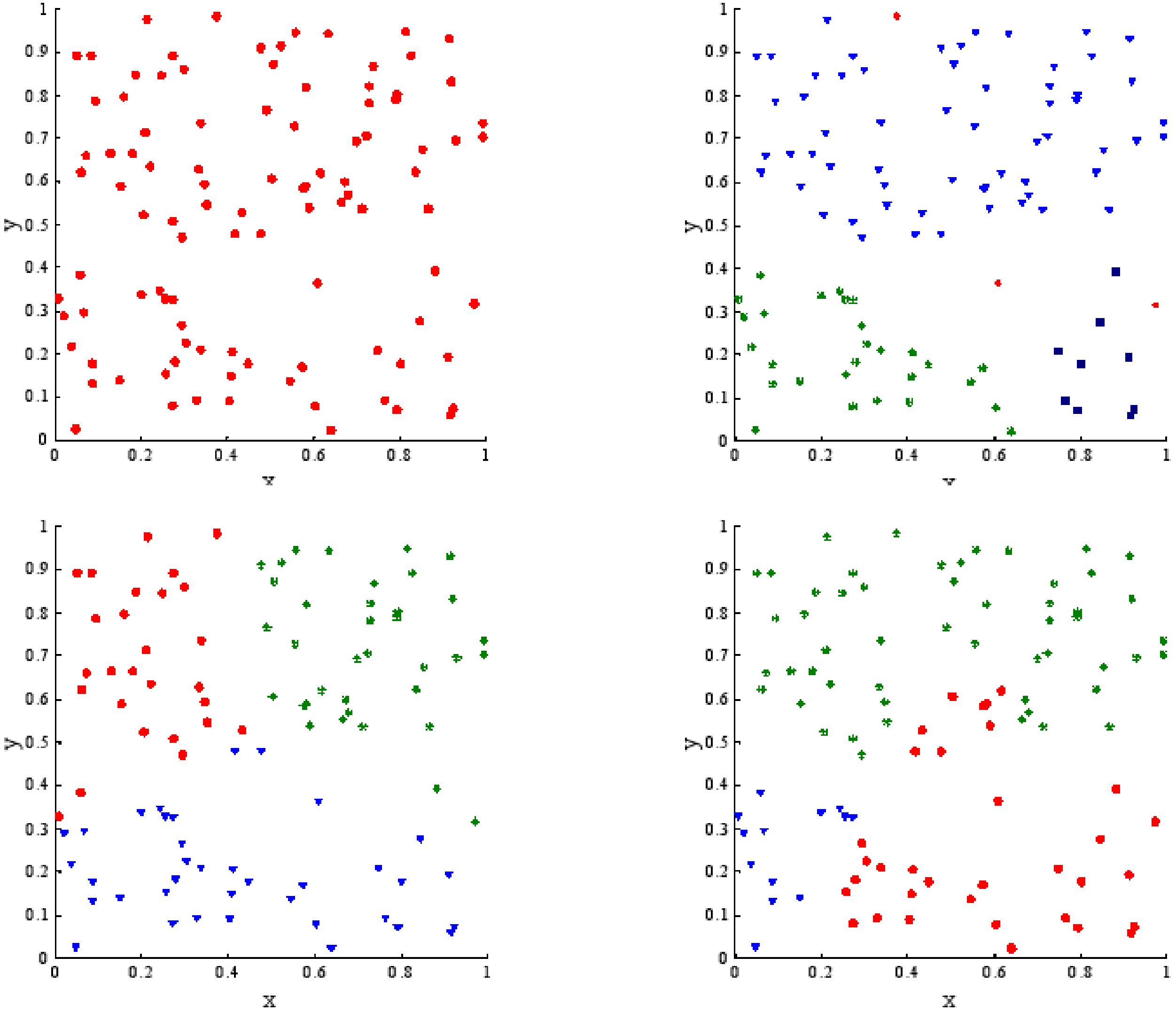
"Clusters are in the eye of the beholder"

Random Points

DBSCAN

K-means

Complete Link



Why evaluate?

- To determine the **clustering tendency** of the dataset, that is distinguish whether non-random structure actually exists in the data
- To determine the **correct number of clusters**
- To evaluate how well the results of a cluster analysis fit the data *without* reference to external information
- To compare the results of a cluster analysis to externally known results, such as externally provided class labels
- To compare two sets of clusters to determine which is better

Note:

- the first three are *unsupervised techniques*, while the last two require external info
- the last three can be applied to the entire clustering or just to individual clusters

Open challenges

Cluster evaluation has a number of challenges:

- a measure of cluster validity may be quite limited in the scope of its applicability
  - ie. dimensions of the problem: most work has been done only on 2- or 3-dimensional data
- we need a framework to interpret any measure
  - How good is "10"?
- if a measure is too complicated to apply or to understand, nobody will use it

Measures of Cluster Validity

Numerical measures that are applied to judge various aspects of cluster validity are classified into the following three types:

- **Internal (unsupervised) Indices:** Used to measure the goodness of a clustering structure without respect to external information
  - cluster *cohesion* vs cluster *separation*
  - e.g. Sum of Squared Error (SSE)
- **External (supervised) Indices:** Used to measure the extent to which cluster labels match externally supplied class labels
  - e.g. entropy, purity, precision, ...
- **Relative Indices:** Used to compare two different clusterings or clusters
  - External or internal indices can be used, e.g. SSE or entropy

External Measures

- Entropy
  - The degree to which each cluster consists of objects of a single class
  - For cluster i we compute p_{ij}, the probability that a member of **cluster** i belongs to **class** j, as $p_{ij} = m_{ij}/m_i$, where m_i is the number of objects in cluster i and m_{ij} is the number of objects of class j in cluster i
  - The **entropy** of each cluster i is $e_i = -\sum_{j=1}^{L} p_{ij} log_2 p_{ij}$, where L is the number of classes
  - The **total entropy** is $e = \sum_{i=1}^{K} \frac{m_i}{m} e_i$, where K is the number of clusters and m is the total number of data points

External Measures

- Purity
  - Another measure of the extent to which a cluster contains objects of a single class
  - Using the previous terminology, the **purity** of cluster i is $p_i = max(p_{ij})$ for all the j
  - The **overall purity** is $purity = \sum_{i=1}^{K} \frac{m_i}{m} p_i$

External Measures

- Precision
  - The fraction of a cluster that consists of objects of a specified class
  - The precision of cluster i with respect to class j is $precision(i, j) = p_{ij}$
- Recall
  - The extent to which a cluster contains all objects of a specified class
  - The recall of cluster i with respect to class j is $recall(i, j) = m_{ij}/m_j$, where m_j is the number of objects in class j

External Measures

- F-measure
  - A combination of both precision and recall that measures the extent to which a cluster contains *only* objects of a particular class and *all* objects of that class
  - The F-measure of cluster i with respect to class j is $F(i,j) = \frac{2 \times precision(i,j) \times recall(i,j)}{precision(i,j) + recall(i,j)}$

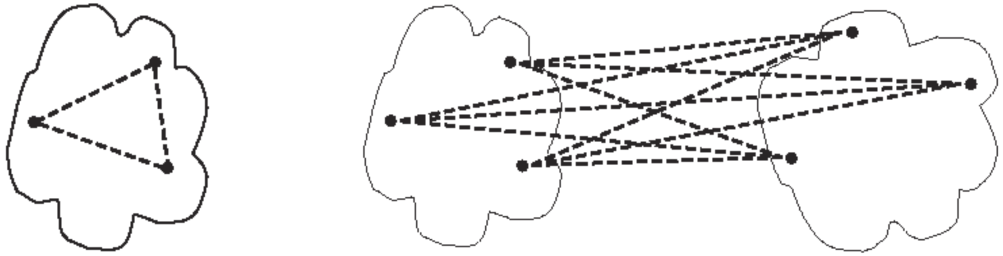
External Measures: example

Table 8.9. K-means clustering results for the *LA Times* document data set.

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

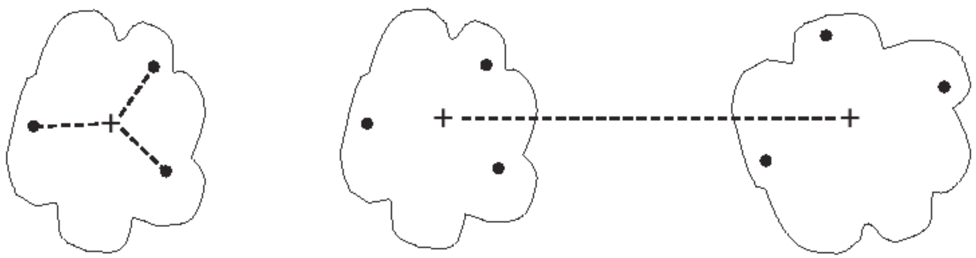
Internal measures: Cohesion and Separation

- Graph-based view



(a) Cohesion. (b) Separation.

- Prototype-based view



(a) Cohesion. (b) Separation.

Internal measures: Cohesion and Separation

- **Cluster Cohesion:** Measures how closely related objects in a cluster are

$$cohesion(C_i) = \sum_{x \in C_i, y \in C_i} proximity(x, y)$$

$$cohesion(C_i) = \sum_{x \in C_i} proximity(x, c_i)$$

- **Cluster Separation:** Measure how distinct or well-separated a cluster is from other clusters

$$separation(C_i, C_j) = \sum_{x \in C_i, y \in C_j} proximity(x, y)$$

$$separation(C_i, C_j) = proximity(c_i, c_j)$$

$$separation(C_i) = proximity(c_i, c)$$

Cohesion and separation example

- Cohesion is measured by the within cluster sum of squares (SSE)

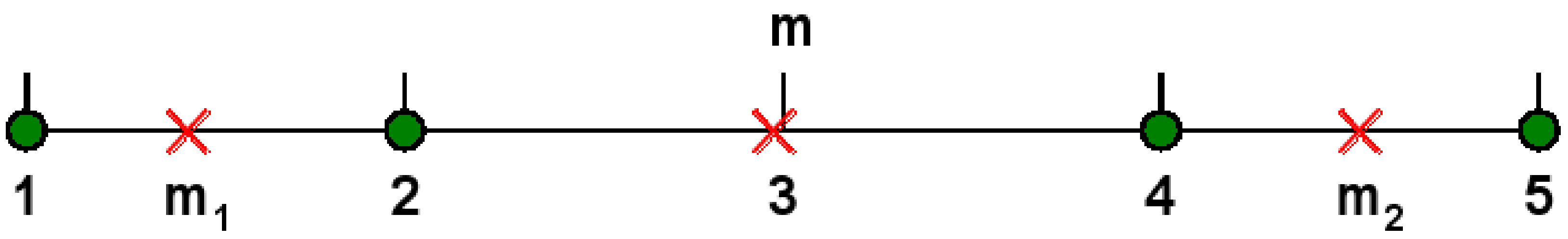
$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

- Separation is measured by the between cluster sum of squares

$$BSS = \sum_i |C_i|(m - m_i)^2$$

where $|C_i|$ is the size of cluster i

Cohesion and separation example



- K=1 cluster:

$$WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$
$$BSS = 4 \times (3-3)^2 = 0$$
$$Total = 10 + 0 = 10$$

- K=2 clusters:

$$WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$
$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$
$$Total = 1 + 9 = 10$$

Evaluating individual clusters and Objects

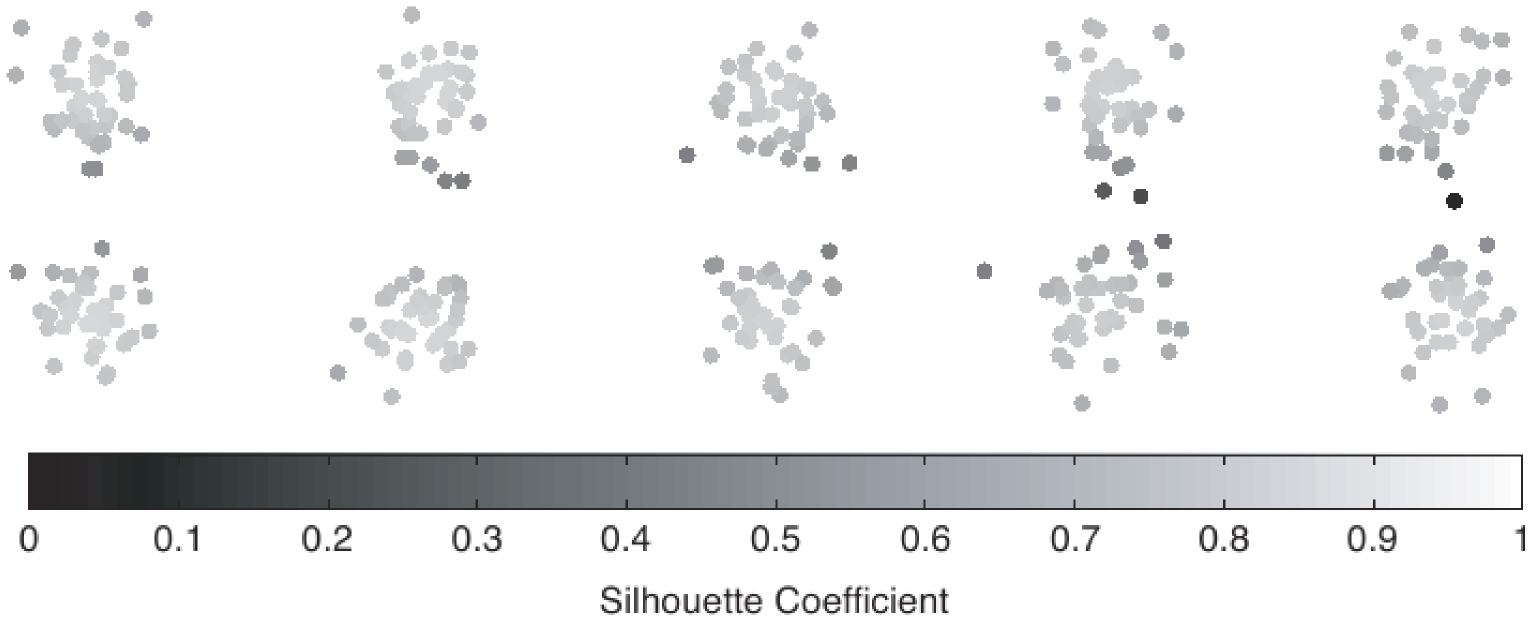
- So far, we have focused on evaluation of a group of clusters
- Many of these measures, however, also can be used to evaluate individual clusters and objects
  - For example, a cluster with a high cohesion may be considered better than a cluster with a lower one
- This information often can be used to improve the quality of the clustering
  - Split not very cohesive clusters
  - Merge not very separated ones
- We can also evaluate the objects within a cluster in terms of their contribution to the overall cohesion or separation of the cluster

The Silhouette Coefficient

- Silhouette Coefficient combine ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings
- For an individual point, i
  - Calculate a_i = average distance of i to the points in its cluster
  - Calculate b_i = min (average distance of i to points in another cluster)
  - The silhouette coefficient for a point is then given by $s_i = (b_i - a_i)/max(a_i, b_i)$

The Silhouette Coefficient

- Silhouette Coefficient combine ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings



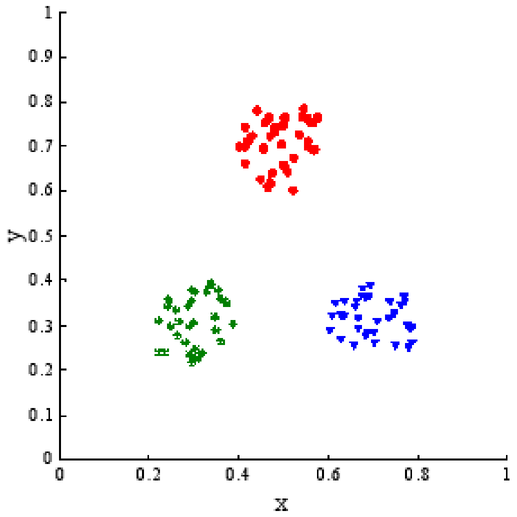
Measuring Cluster Validity via Correlation

If we are given the similarity matrix for a data set and the cluster labels from a cluster analysis of the data set, then we can evaluate the "goodness" of the clustering by looking at the **correlation** between the similarity matrix and an ideal version of the similarity matrix based on the cluster labels

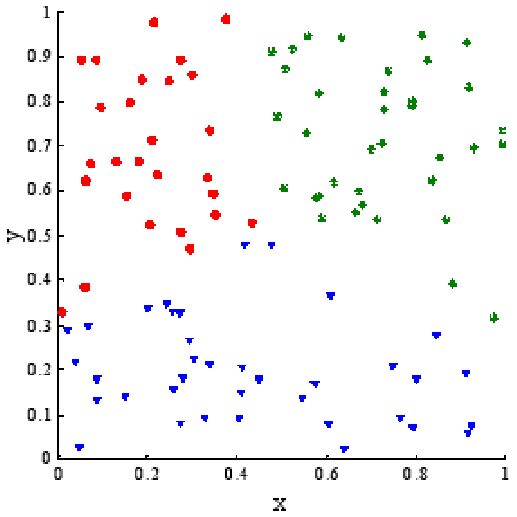
- Similarity/Proximity Matrix
- Ideal Matrix
  - One row and one column for each data point
  - An entry is 1 if the associated pair of points belongs to the same cluster
  - An entry is 0 if the associated pair of points belongs to different clusters

Measuring Cluster Validity via Correlation

- Compute the correlation between the two matrices
  - Since the matrices are symmetric, only the correlation between $n(n-1)/2$ entries needs to be calculated
- High correlation indicates that points that belong to the same cluster are close to each other



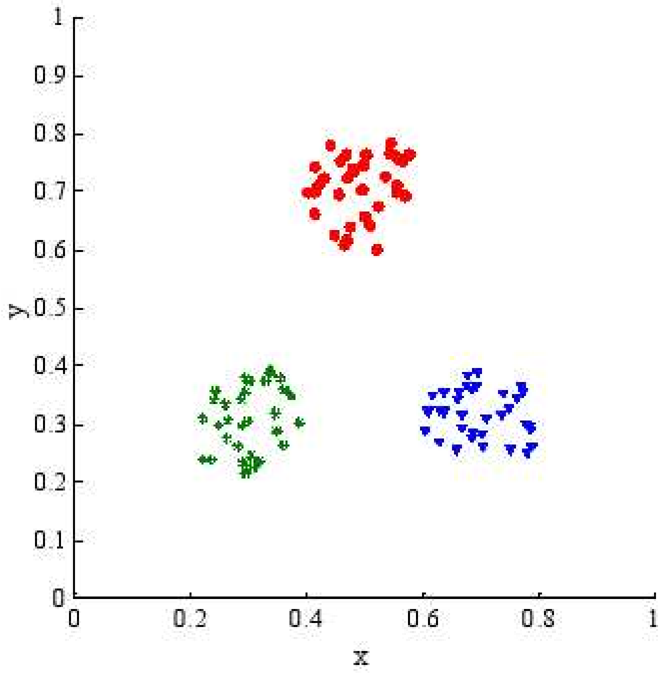
Corr = -0.9235

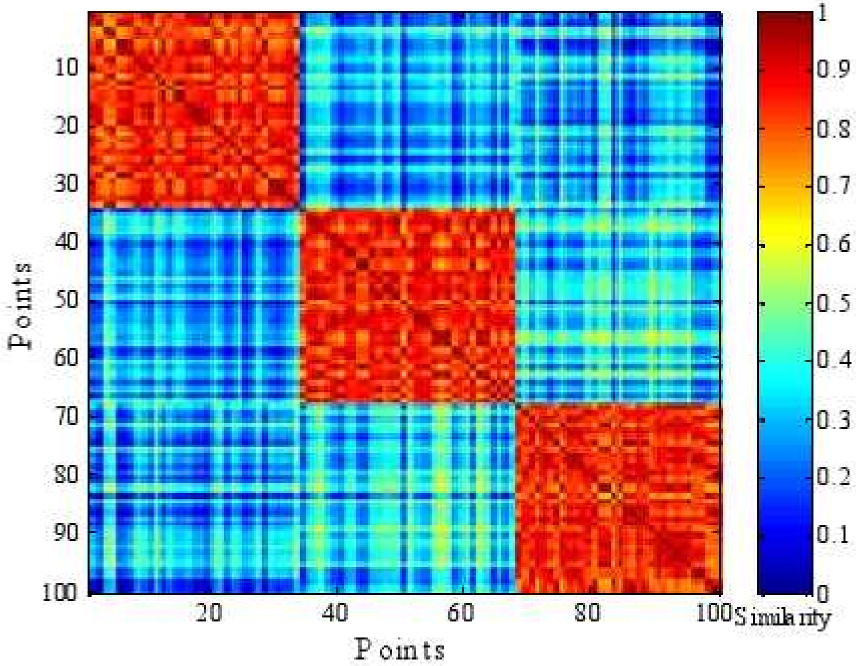


Corr = -0.5810

Using Similarity Matrix for Cluster Validation

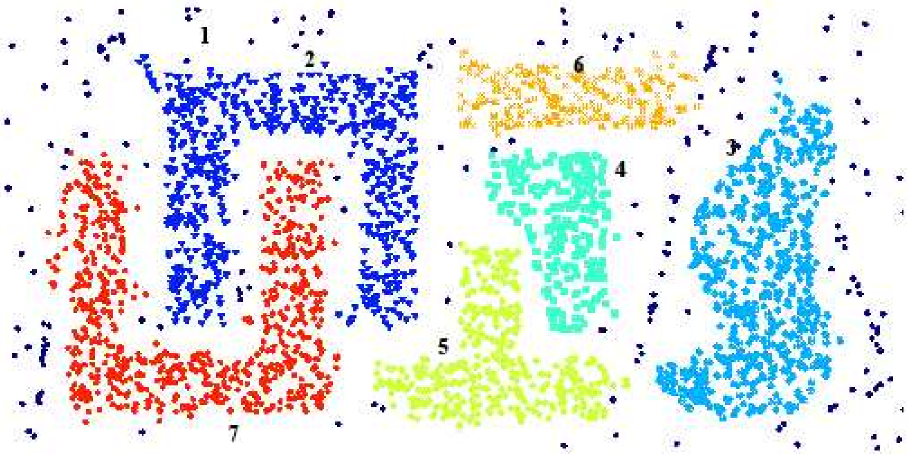
- Order the similarity matrix with respect to cluster labels and inspect visually

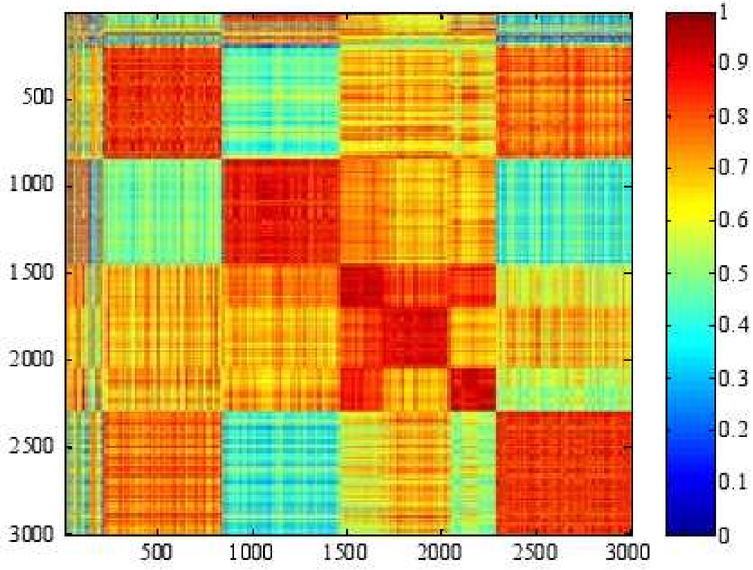




Using Similarity Matrix for Cluster Validation

- Order the similarity matrix with respect to cluster labels and inspect visually

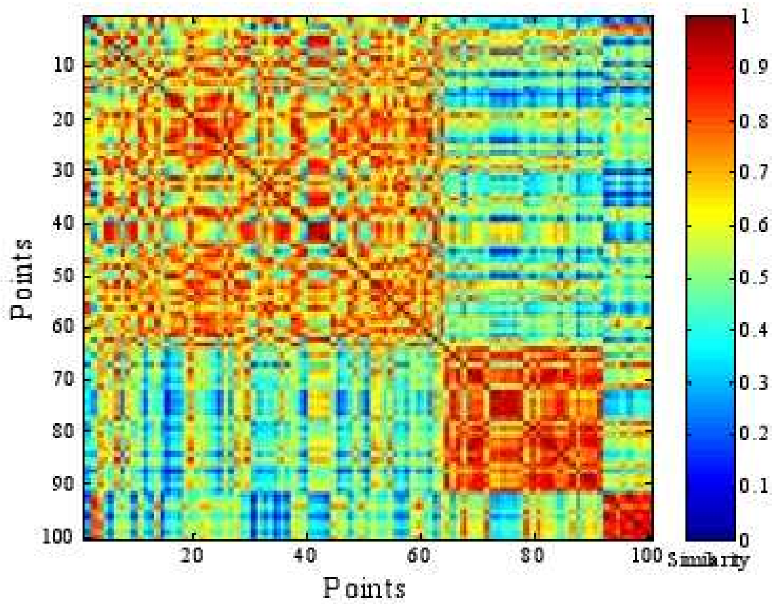


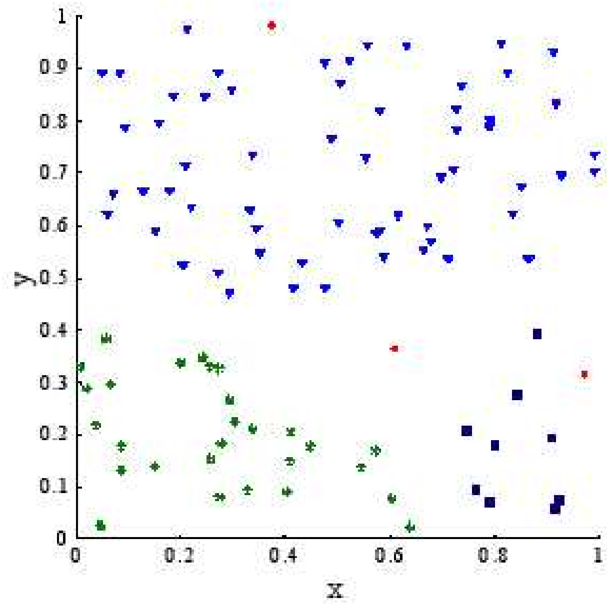


DBSCAN

Using Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp

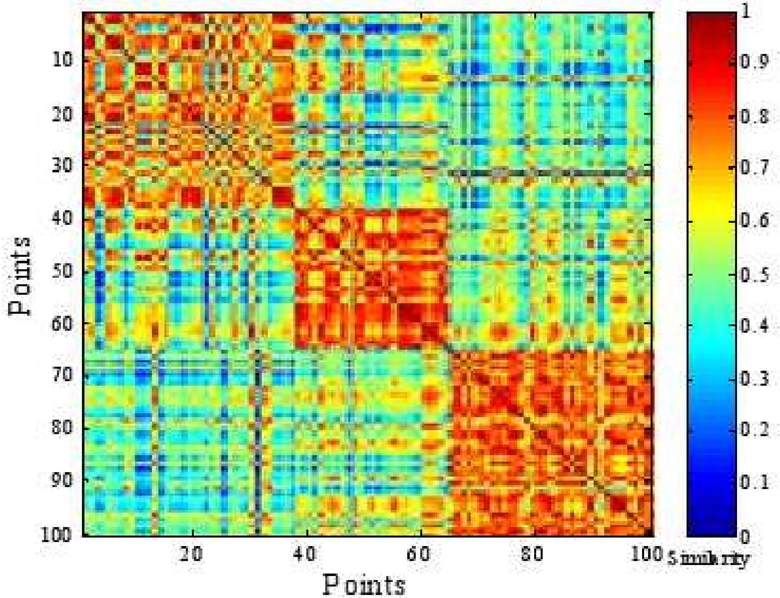


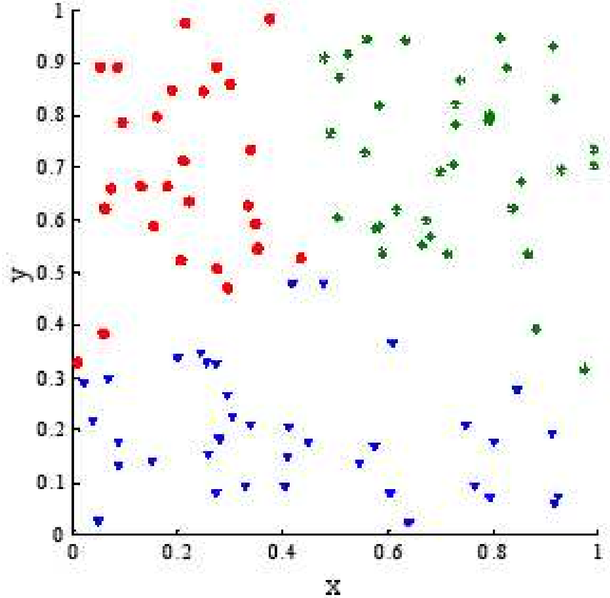


DBSCAN

Using Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp

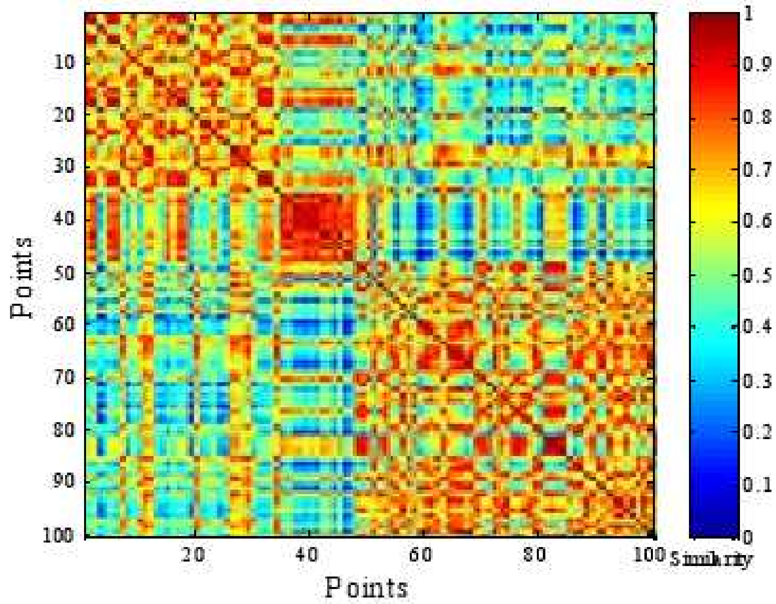


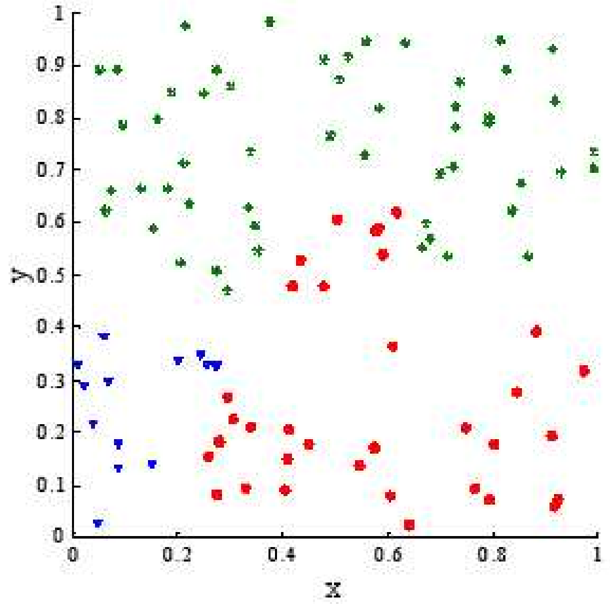


K-means

Using Similarity Matrix for Cluster Validation

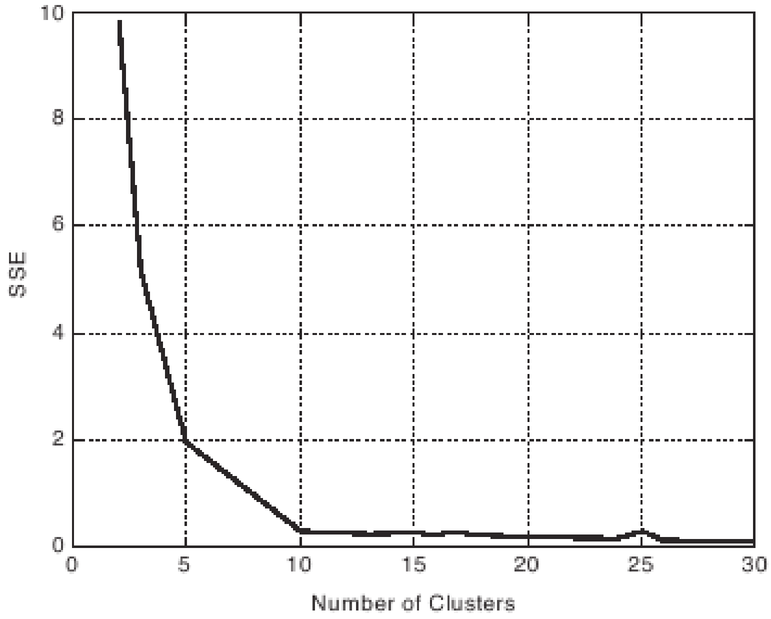
- Clusters in random data are not so crisp

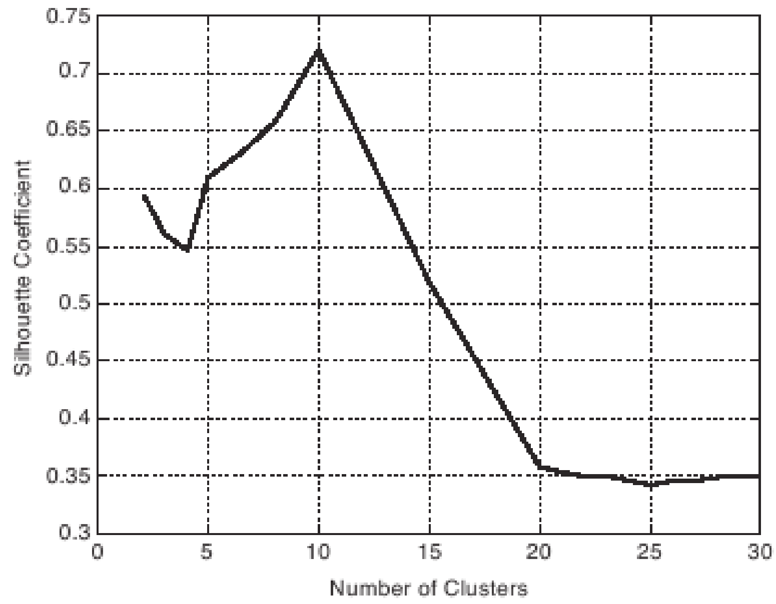




Complete Link

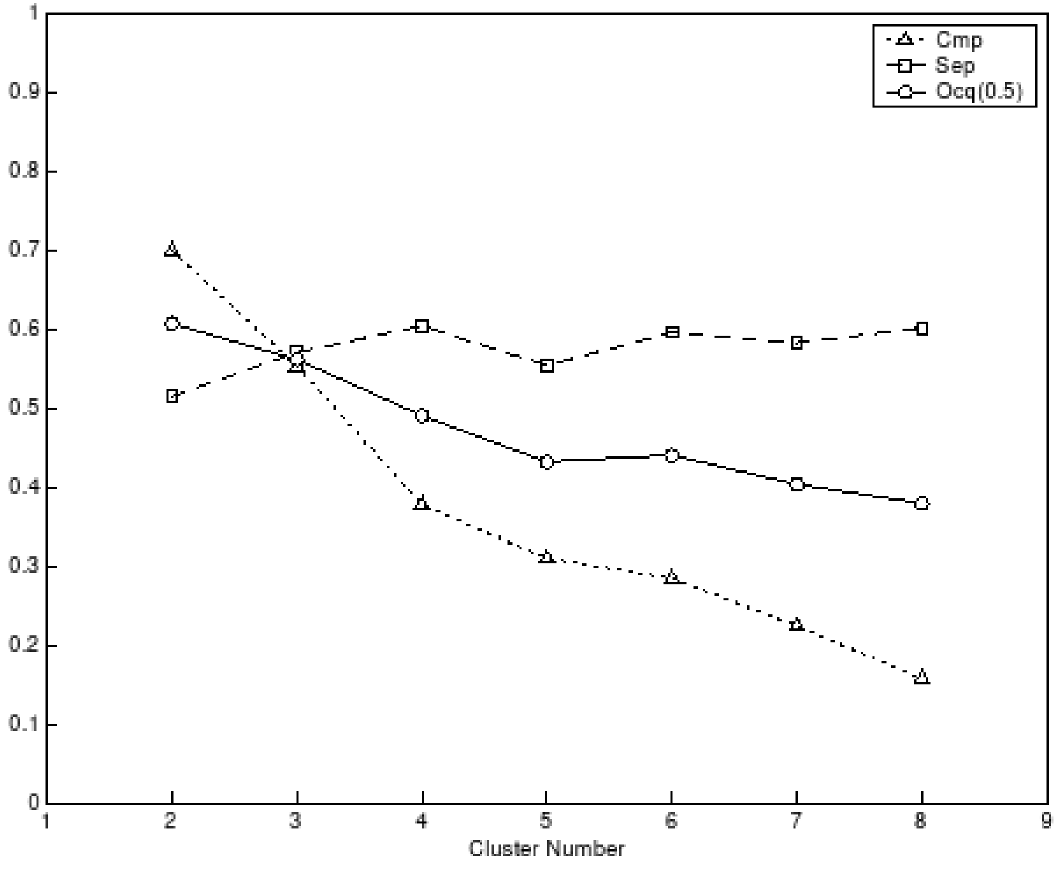
Finding the Correct Number of Clusters





- Look for the number of clusters for which there is a knee, peak, or dip in the plot of the evaluation measure when it is plotted against the number of clusters

Finding the Correct Number of Clusters



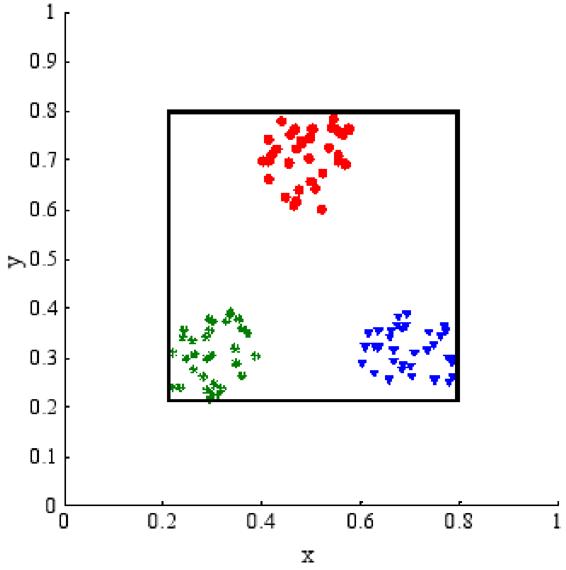
- Of course, this isn't always easy...

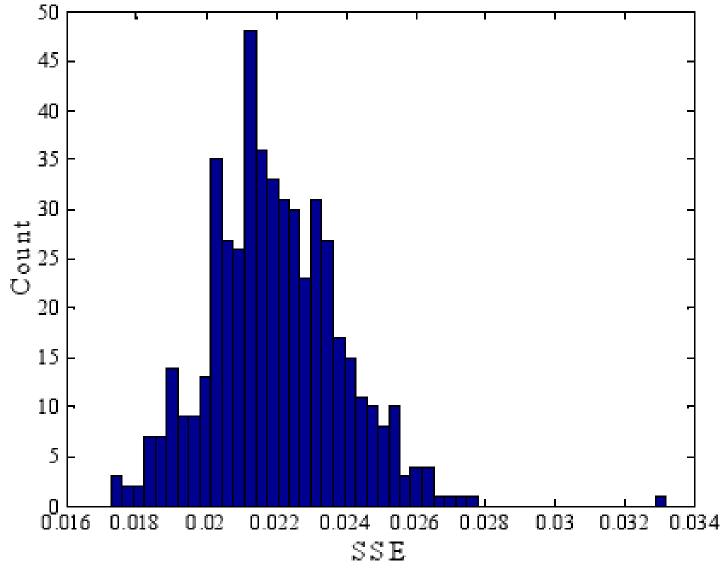
Framework for Cluster Validity

- Need a framework to interpret any measure.
  - For example, if our measure of evaluation has the value "10", is that good, fair, or poor?
- Statistics provide a framework for cluster validity
  - The more atypical a clustering result is, the more likely it represents valid structure in the data
  - Can compare the values of an index that result from random data or clusterings to those of a clustering result: if the value of the index is unlikely, then the cluster results are valid
  - These approaches are more complicated and harder to understand
- For comparing the results of two different sets of cluster analyses, a framework is less necessary
  - However, there is the question of whether the difference between two index values is significant

Statistical Framework for SSE

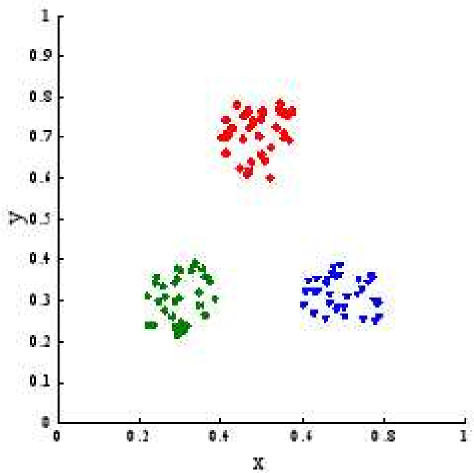
- Example
  - Compare SSE of 0.005 against three clusters in random data
  - Histogram shows SSE of three clusters in 500 sets of random data points of size 100 distributed over the range 0.2   0.8 for x and y values

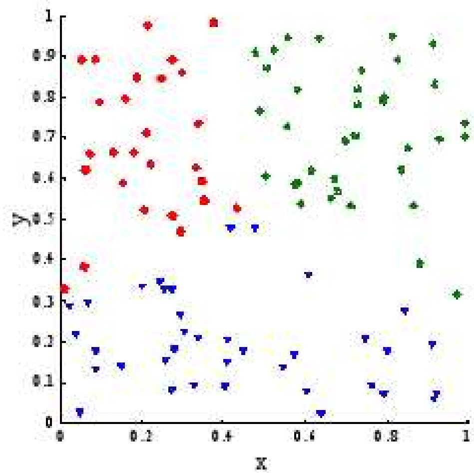


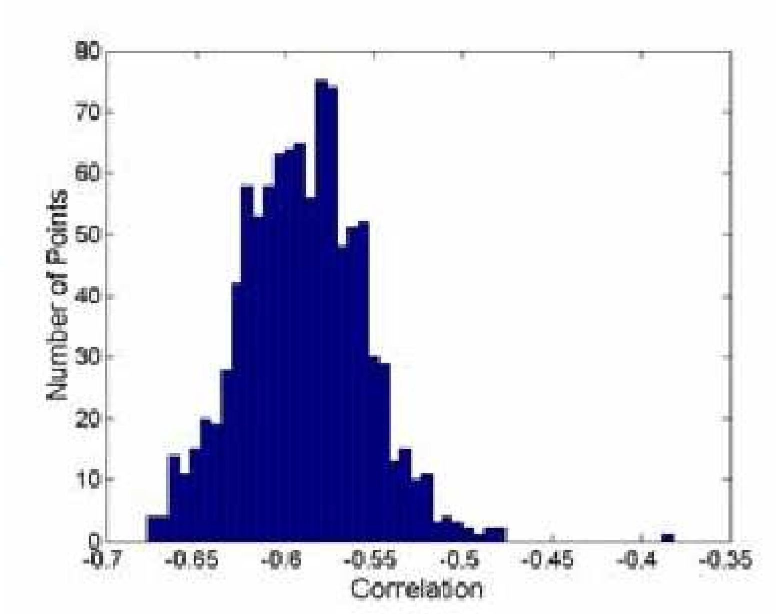


Statistical Framework for Correlation

- Correlation of incidence and proximity matrices for the K-means clusterings of the following two data sets







Corr = -0.9235

Corr = -0.5810

Final Comment on Cluster Validity

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

Algorithms for Clustering Data, Jain and Dubes

Bibliography

- Slides about clustering for the Data Mining course
  prof. Salvatore Orlando (link)
- Tan, Steinbach, Kumar: "Introduction to Data Mining", Ch. 8
  http://www-users.cs.umn.edu/ kumar/dmbook/index.php

- The end (really!)