دانشگاه آزاد اسلامی واحد تبریز

نام درس: داده کاوی

بخش: خوشه بندی دوگانه و ارزیابی خوشه بندی

نام استاد: دکتر مسعود کارگر

Roadmap

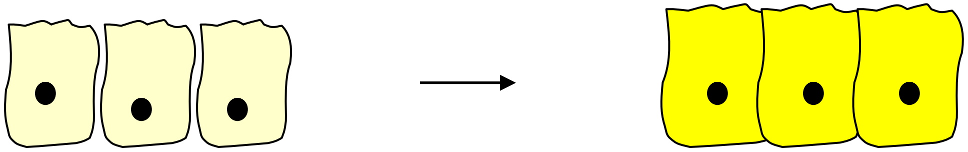
- Introduction to Bi-clustering
- Difference between Bi-clustering and two-way clustering
- Typical Procedure for Cluster discovery
- Why Cluster Validation/Evaluation?
- Different methods for clustering evaluation.

Clustering Analysis in Read-world: Microarray

If we observe gene activity (expression) over specified experimental conditions we can assign putative functions to genes.

Condition 1 Normal Cells

Condition 2 Cancerous Cells



Gene A

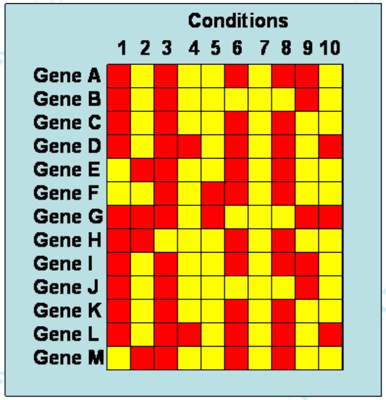
Normally expressed

Gene A ↑

Highly expressed

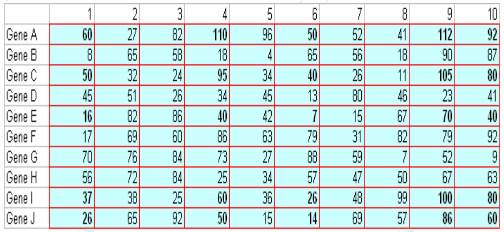
Conclusion: **Gene A** may be involved in onset of cancer

Microarray experiments have the capacity to analyze the expression of 10000’s of genes over many conditions.



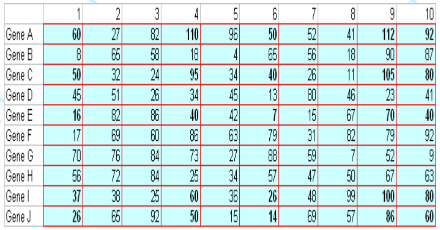
	Conditions									
	1	2	3	4	5	6	7	8	9	10
Gene A	🟥	🟨	🟥	🟨	🟨	🟥	🟨	🟥	🟥	🟨
Gene B	🟥	🟨	🟥	🟨	🟨	🟨	🟨	🟨	🟥	🟨
Gene C	🟥	🟨	🟥	🟨	🟨	🟥	🟨	🟥	🟨	🟨
Gene D	🟥	🟨	🟥	🟥	🟨	🟥	🟨	🟥	🟨	🟥
Gene E	🟨	🟥	🟥	🟨	🟨	🟥	🟨	🟥	🟨	🟨
Gene F	🟨	🟨	🟥	🟨	🟥	🟥	🟨	🟥	🟨	🟨
Gene G	🟥	🟥	🟥	🟨	🟥	🟨	🟨	🟨	🟥	🟥
Gene H	🟥	🟥	🟨	🟨	🟨	🟥	🟨	🟥	🟨	🟨
Gene I	🟥	🟨	🟥	🟨	🟨	🟥	🟨	🟥	🟥	🟨
Gene J	🟥	🟨	🟥	🟨	🟨	🟨	🟨	🟨	🟥	🟨
Gene K	🟥	🟨	🟥	🟨	🟨	🟥	🟨	🟥	🟨	🟨
Gene L	🟥	🟨	🟥	🟥	🟨	🟥	🟨	🟥	🟨	🟥
Gene M	🟨	🟥	🟥	🟨	🟨	🟥	🟨	🟥	🟨	🟨

Microarray

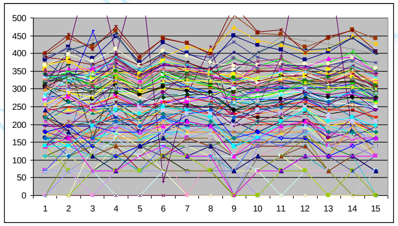


	1	2	3	4	5	6	7	8	9	10
Gene A	**60**	27	82	**110**	96	**50**	52	41	**112**	**92**
Gene B	8	65	58	18	4	65	56	18	90	87
Gene C	**50**	32	24	**95**	34	**40**	26	11	**105**	**80**
Gene D	45	51	26	34	45	13	80	46	23	41
Gene E	**16**	82	86	**40**	42	**7**	15	67	**70**	**40**
Gene F	17	69	60	86	63	79	31	82	79	92
Gene G	70	76	84	73	27	88	59	7	52	9
Gene H	56	72	84	25	34	57	47	50	67	63
Gene I	**37**	38	25	**60**	36	**26**	48	99	**100**	**80**
Gene J	**26**	65	92	**50**	15	**14**	69	57	**86**	**60**

Clustering Analysis of Gene Expression Datasets

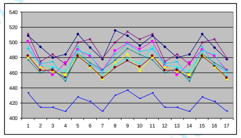


	1	2	3	4	5	6	7	8	9	10
Gene A	**60**	27	82	**110**	96	**50**	52	41	**112**	**92**
Gene B	8	65	58	18	4	65	56	18	90	87
Gene C	**50**	32	24	**95**	34	**40**	26	11	**105**	**80**
Gene D	45	51	26	34	45	13	80	46	23	41
Gene E	**16**	82	86	**40**	42	**7**	15	67	**70**	**40**
Gene F	17	69	60	86	63	79	31	82	79	92
Gene G	70	76	84	73	27	88	59	7	52	9
Gene H	56	72	84	25	34	57	47	50	67	63
Gene I	**37**	38	25	**60**	36	**26**	48	99	**100**	**80**
Gene J	**26**	65	92	**50**	15	**14**	69	57	**86**	**60**

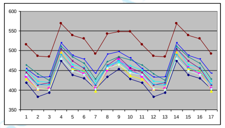


Graph of Gene Expression V's Conditions

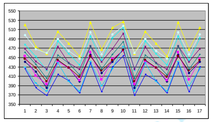
Clustering…



Cluster A



Cluster B



Cluster C

Similar rows are grouped together into unique clusters. The premise is that each Cluster may represent a group of functionally related genes (**Biological Module**).

Possible Drawbacks with Clustering:

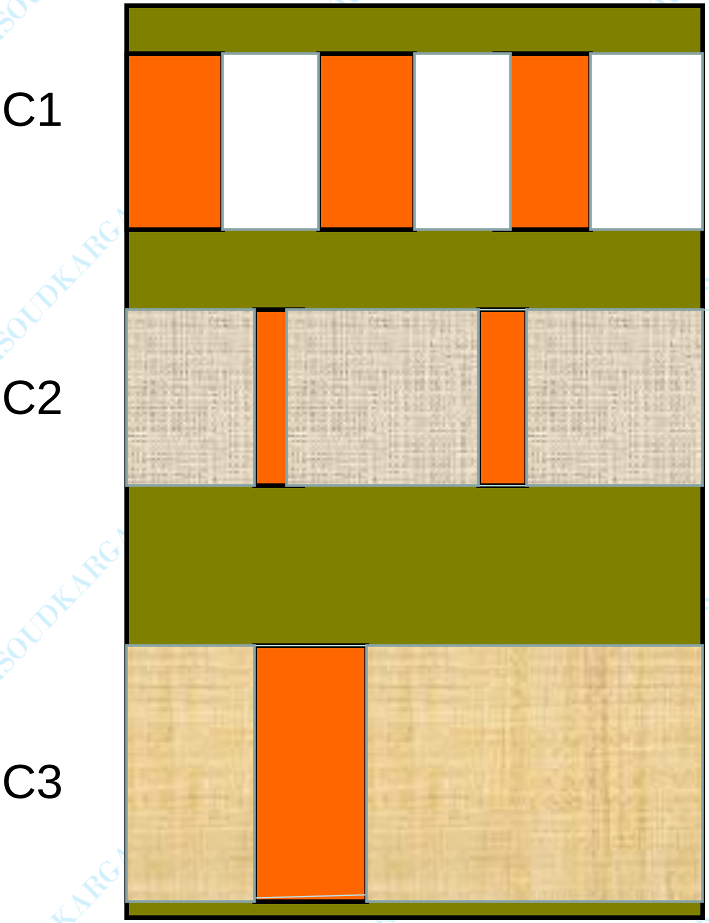
1. In high dimensional expression datasets similarity may not exist over all conditions.

2. Related genes may naturally **correlate over some conditions** and not others.

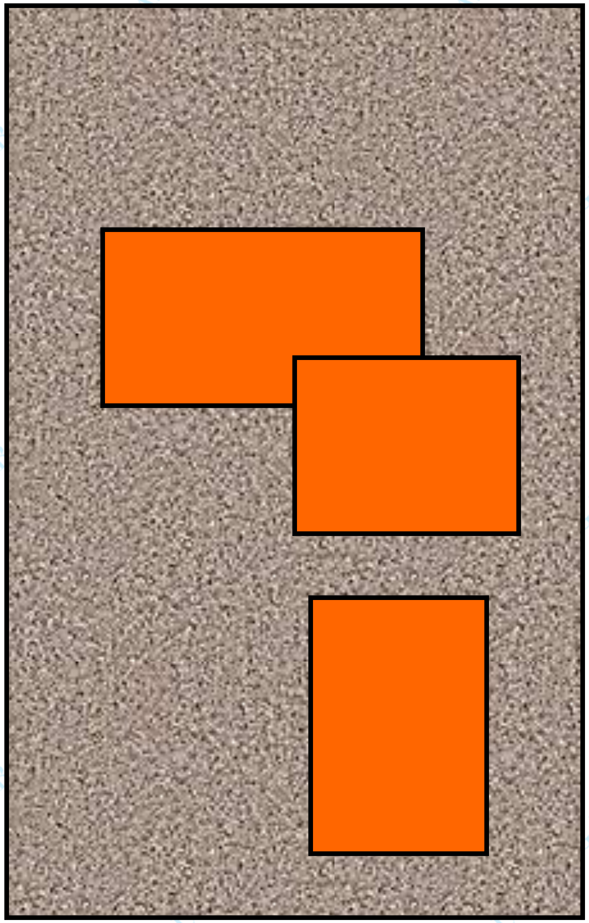
3. Genes may have more than one function (**clusters may not be disjoint but overlap**).

With such datasets it is more beneficial to cluster the data over both rows and columns or to employ **Biclustering**

Necessity is the Mother of Invention



Traditional Clustering

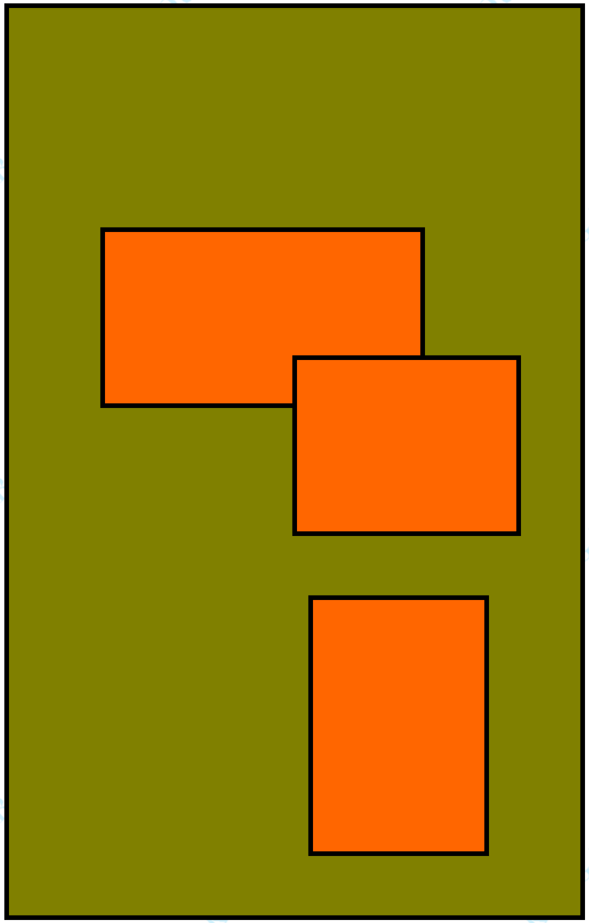


Biclustering

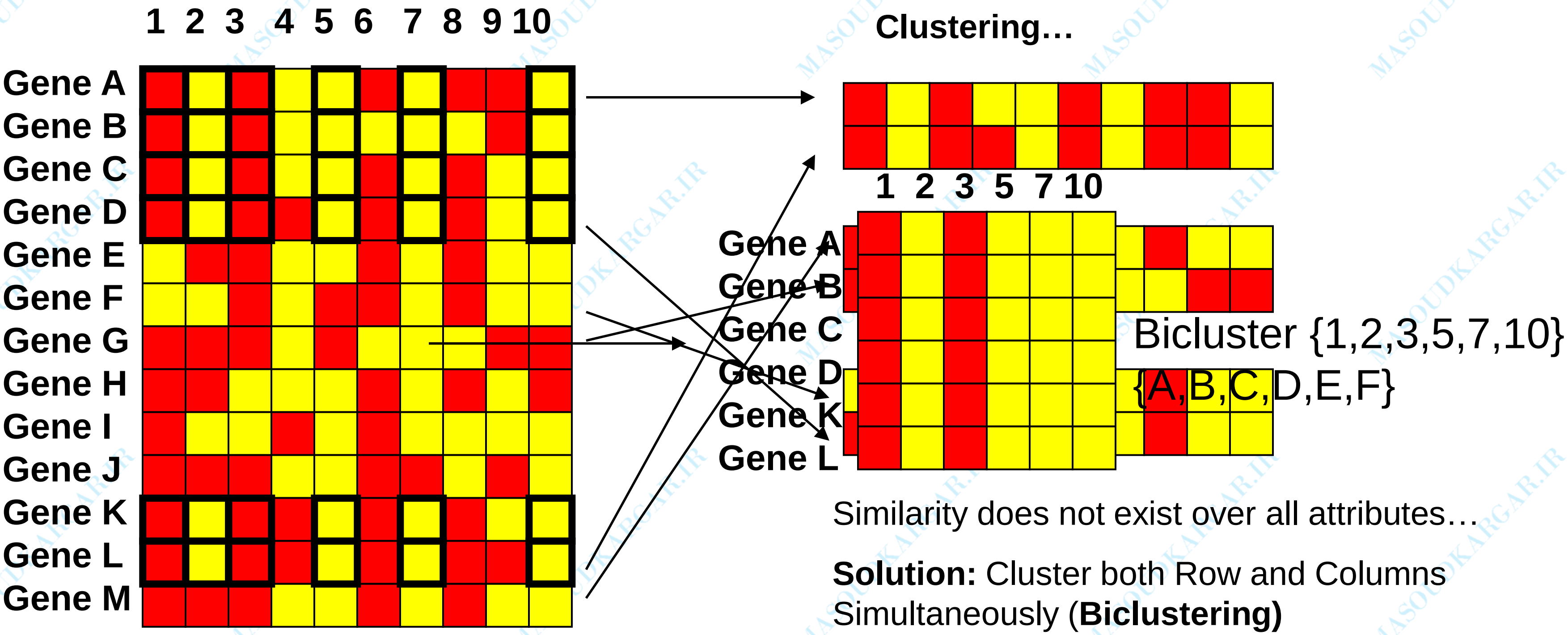
Biclustering concept was proposed in 1960s. But rarely used or studied until 2000 !

Biclustering: The Background

- Usual clustering algorithms are based on global similarities of rows or columns of an expression data matrix.
- But the similarity of the expression profiles of a group of genes may be restricted to **certain experimental conditions**.
- Goal of biclustering: identify “homogeneous” **submatrices**.
- Difficulties: computational complexity, assessing the statistical significance of results



Biclustering V.S. Clustering



Similarity does not exist over all attributes…

Solution: Cluster both Row and Columns Simultaneously (**Biclustering)**

Cheng and Church (2000) introduced the concept of **Biclustering** to the area of gene expression analysis.

They developed a function called the **Mean Squared Residue Score** to score sub-matrices and locate those with good row and column correlation (**Biclusters**)

The exhaustive search for and scoring of all sub-matrices is **NP-hard** and they employed a **Greedy Search Heuristic** in their approach.

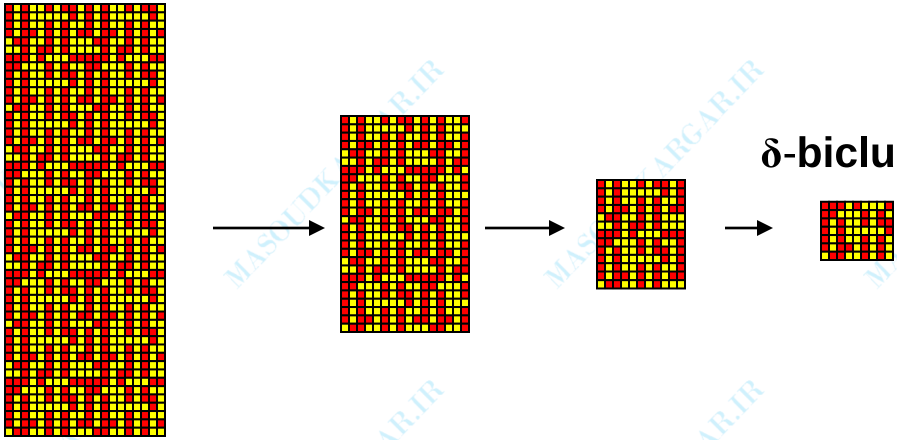
Cheng and Church Node Deletion Approach

Cheng and Church’s greedy search approach involved deleting rows and columns from the parent matrix which most improved the Mean Squared Residue Score. The search is stopped upon reaching a predefined score (δ), this solution is referred to as the **δ-bicluster.**

Input:

Data matrix

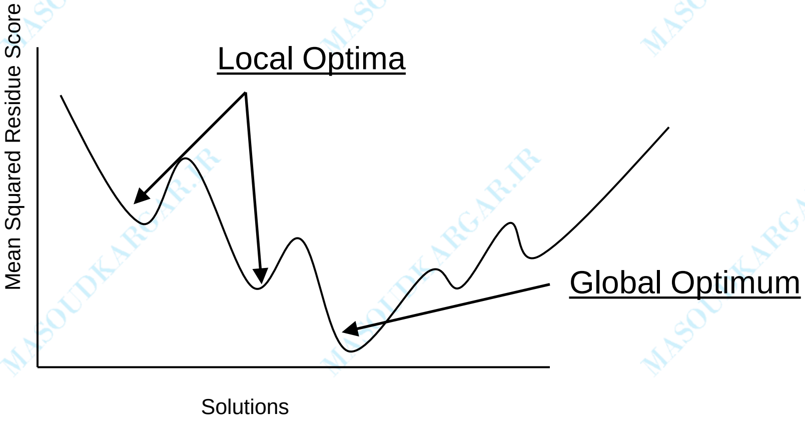
δ = 300



δ-bicluster

Score: 1,052 **Score: 543** **Score:423** **Score: 300**

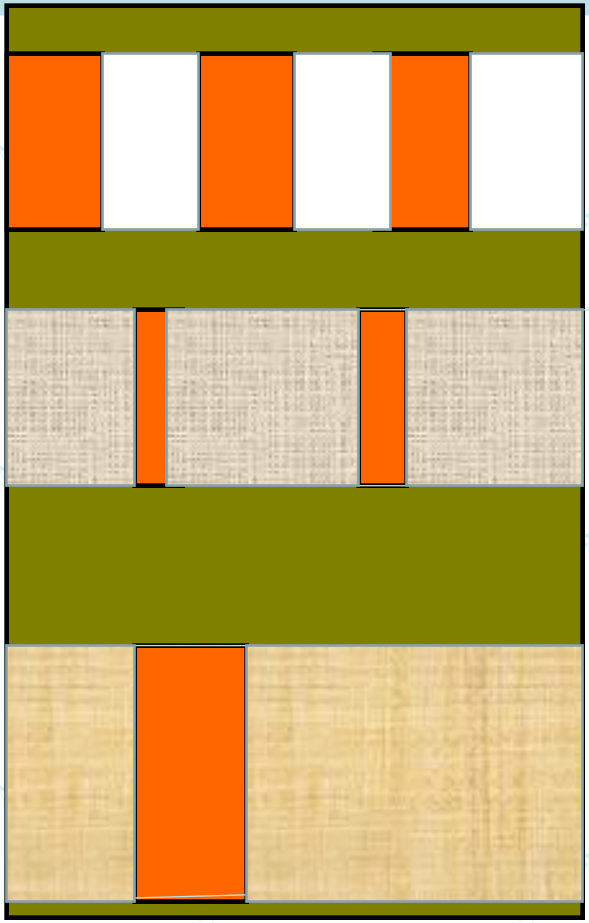
Cheng, Y. and Church, G.M. (2000) Biclustering of expresssion data. ISMB 2000



Overview of the Biclustering Methods

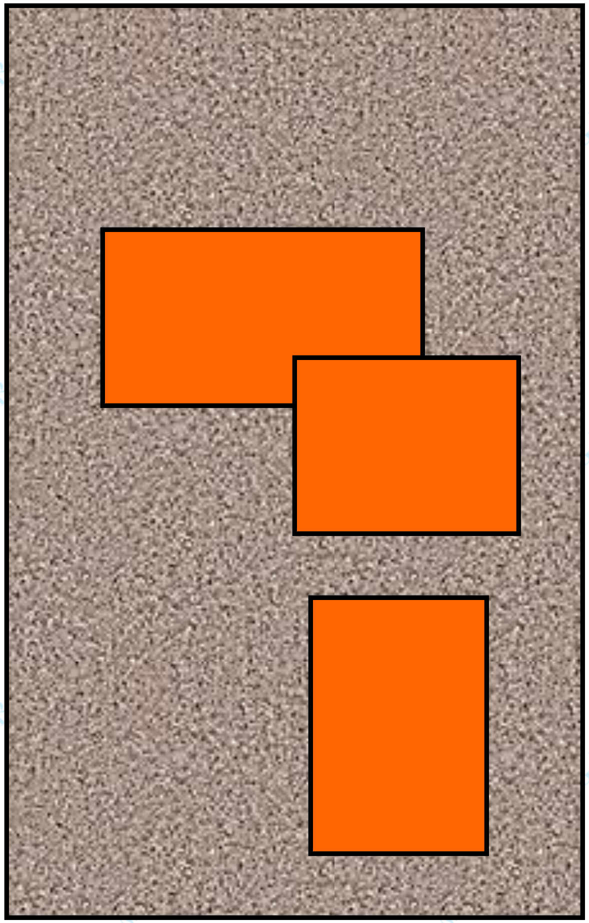
Method	Publish	Cluster Model	Goal
Cheng & Church	ISMB 2000	Background + row effect + column effect	Minimize mean squared residue of biclusters
Getz et al. (CTWC)	PNAS 2000	Depending on plugin clustering algorithm	Depending on plugin clustering algorithm
Lazzeroni & Owen (Plaid Models)	Bioinformatics 2000	Background + row effect + column effect	Minimize modeling error
Ben-Dor et al. (OPSM)	RECOMB 2002	All genes have the same order of expression values	Minimize the p-values of biclusters
Tanay et al. (SAMBA)	Bioinformatics 2002	Maximum bounded bipartite subgraph	Minimize the p-values of biclusters
Yang et al. (FLOC)	BIBE 2003	Background + row effect + column effect	Minimize mean squared residue of biclusters
Kluger et al. (Spectral)	Genome Res. 2003	Background × row effect × column effect	Finding checkerboard structures

Difference Between Biclustering and Two-way clustering



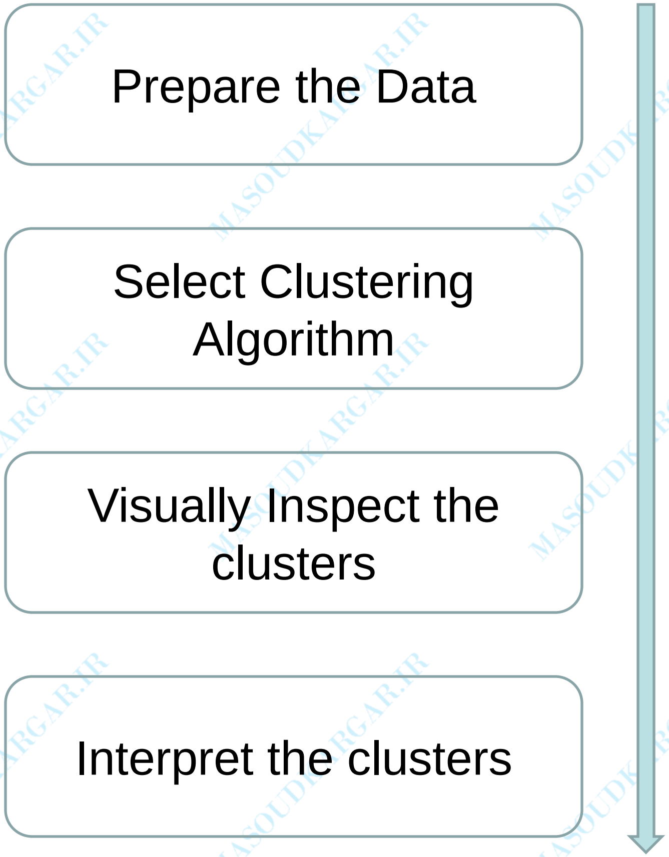
Clustering by Row(Genes)

Clustering by Column



Biclustering

Clustering Analysis in Real-world

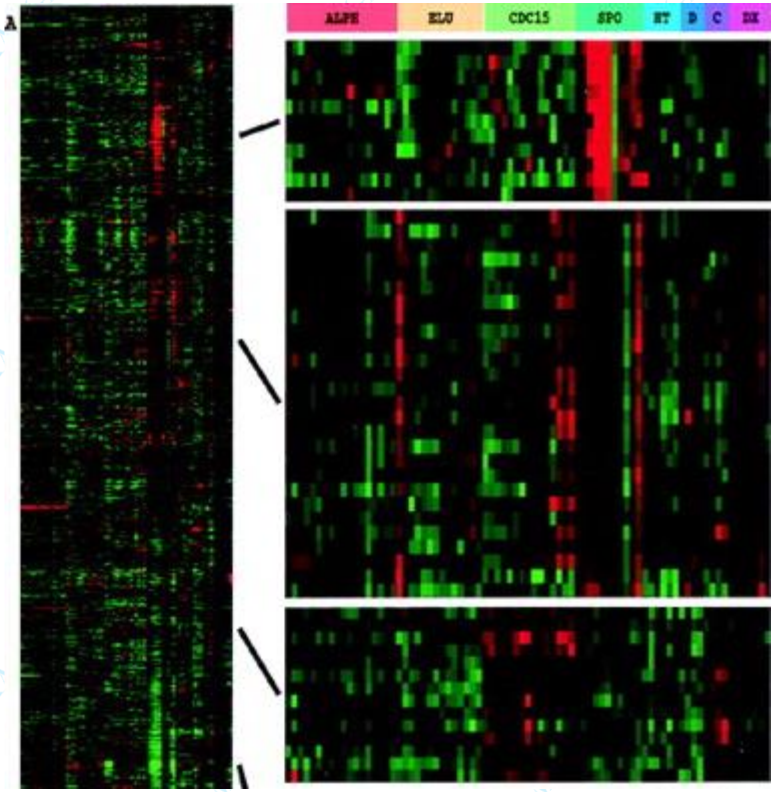


Prepare the Data

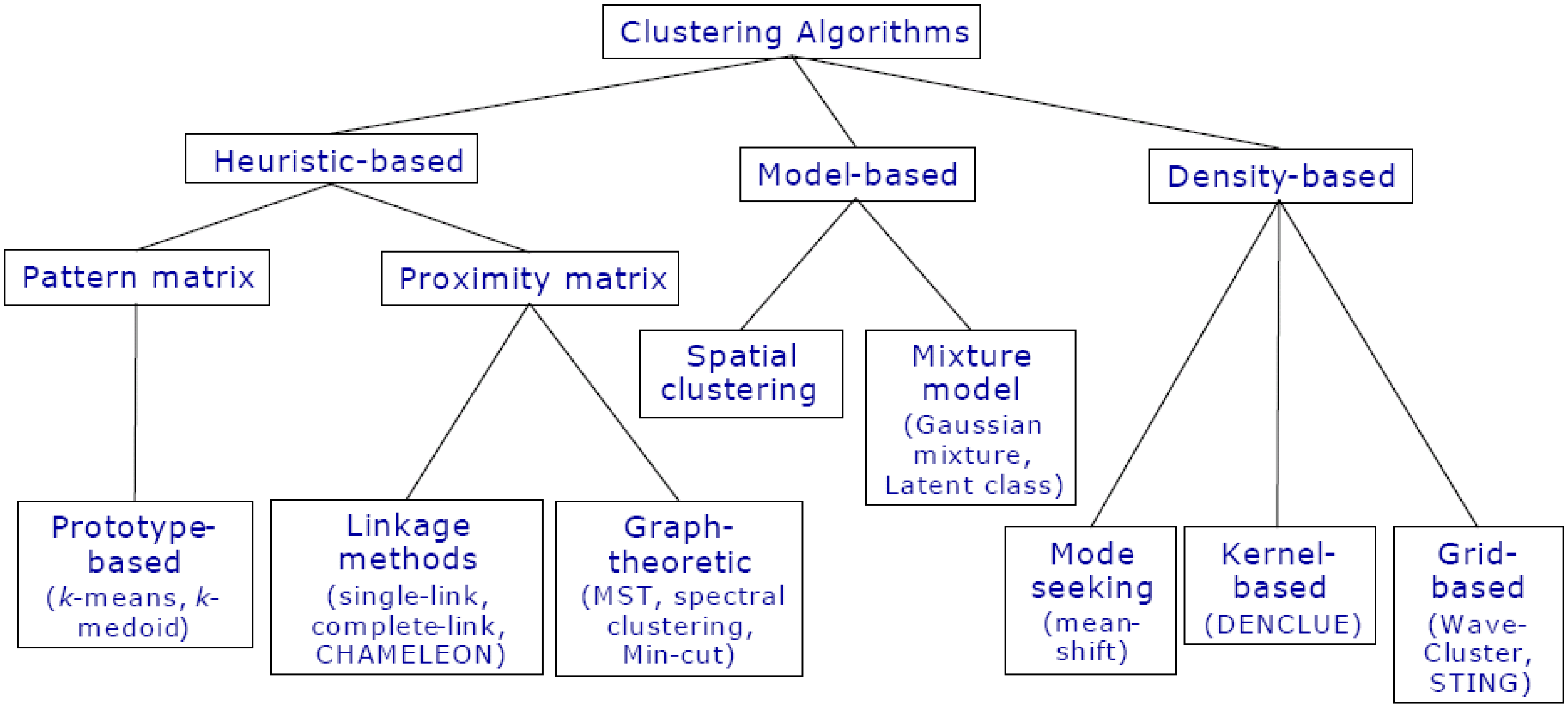
Select Clustering Algorithm

Visually Inspect the clusters

Interpret the clusters



Too Many Clustering Algorithms? → Evaluation



Clustering Algorithms

Heuristic-based

Model-based

Density-based

Pattern matrix

Proximity matrix

Spatial clustering

Mixture model (Gaussian mixture, Latent class)

Prototype-based (*k*-means, *k*-medoid)

Linkage methods (single-link, complete-link, CHAMELEON)

Graph-theoretic (MST, spectral clustering, Min-cut)

Mode seeking (mean-shift)

Kernel-based (DENCLUE)

Grid-based (Wave-Cluster, STING)

Cluster Validity

- For supervised classification we have a variety of measures to evaluate how good our model is
  - Accuracy, precision, recall

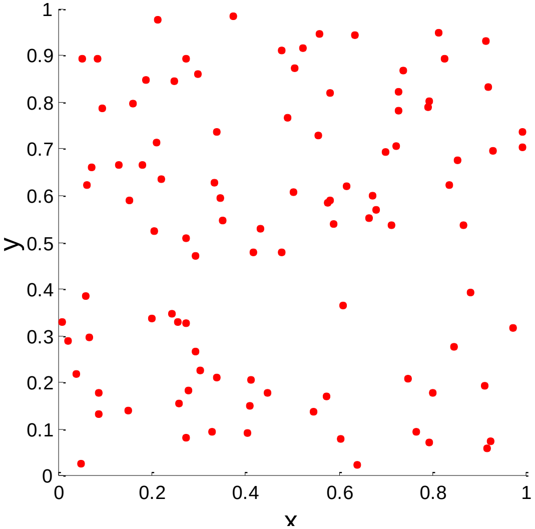
- For cluster analysis, the analogous question is how to evaluate the “goodness” of the resulting clusters?

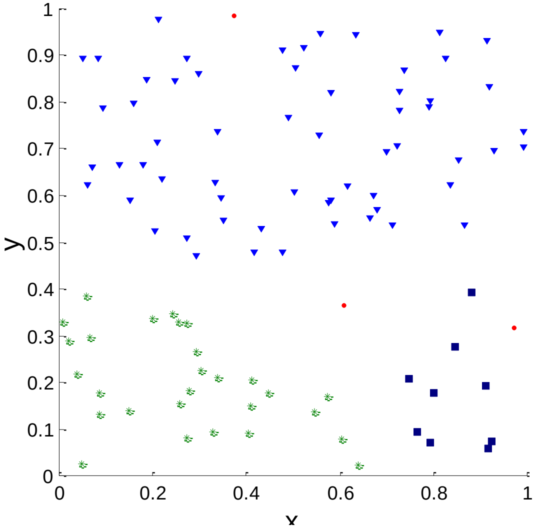
- But “clusters are in the eye of the beholder”!

- Then why do we want to evaluate them?
  - To avoid finding patterns in noise
  - To compare clustering algorithms
  - To compare two sets of clusters
  - To compare two clusters

Clusters found in Random Data

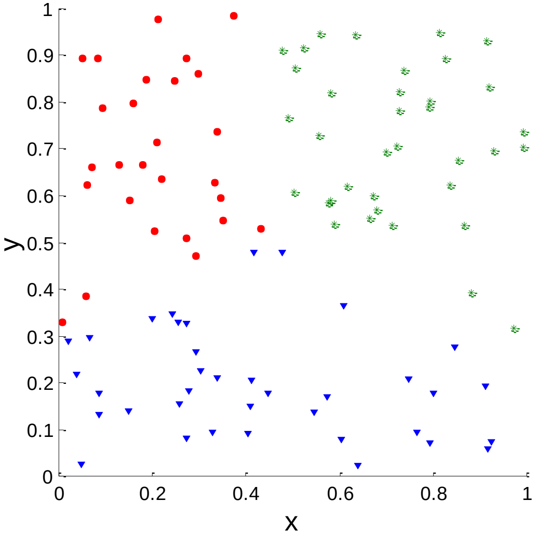
Random Points

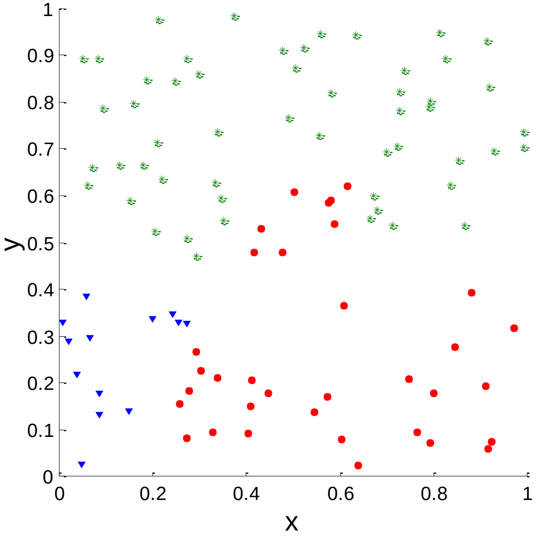




DBSCAN

K-means





Complete Link

Different Aspects of Cluster Validation

distinguishing whether non-random structure actually exists in the data.

2. Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.
3. Evaluating how well the results of a cluster analysis fit the data *without* reference to external information.
   - Use only the data
4. Comparing the results of two different sets of cluster analyses to determine which is better.
5. Determining the ‘correct’ number of clusters.

For 2, 3, and 4, we can further distinguish whether we want to evaluate the entire clustering or just individual clusters.

Measures of Cluster Validity

of cluster validity, are classified into the following three types.

- External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
    - Entropy
- Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information.
    - Sum of Squared Error (SSE)
- Relative Index: Used to compare two different clusterings or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy

- Sometimes these are referred to as criteria instead of indices
    - However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

Measuring Cluster Validity Via Correlation

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

- 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.

	1	2	3
1	0	0.5	1.0
2		0	2.0
3			0

	1	2	3
1		1	0
2			1
3			

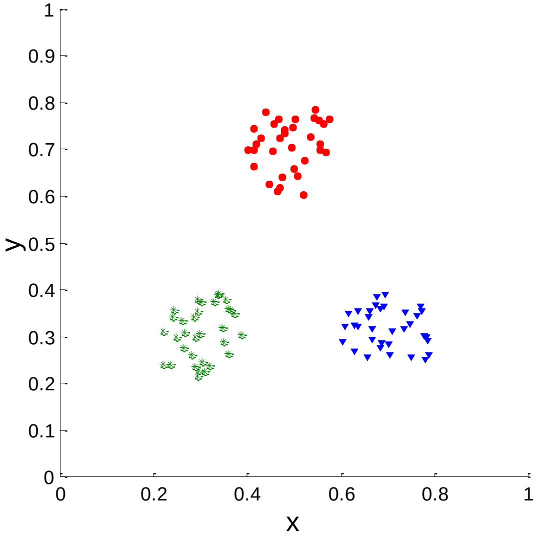
Measuring Cluster Validity Via Correlation

same cluster are close to each other.

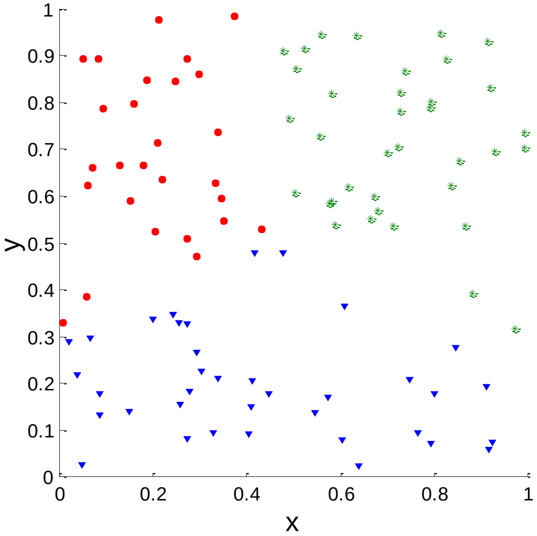
- Not a good measure for some density or contiguity based clusters.

Measuring Cluster Validity Via Correlation

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



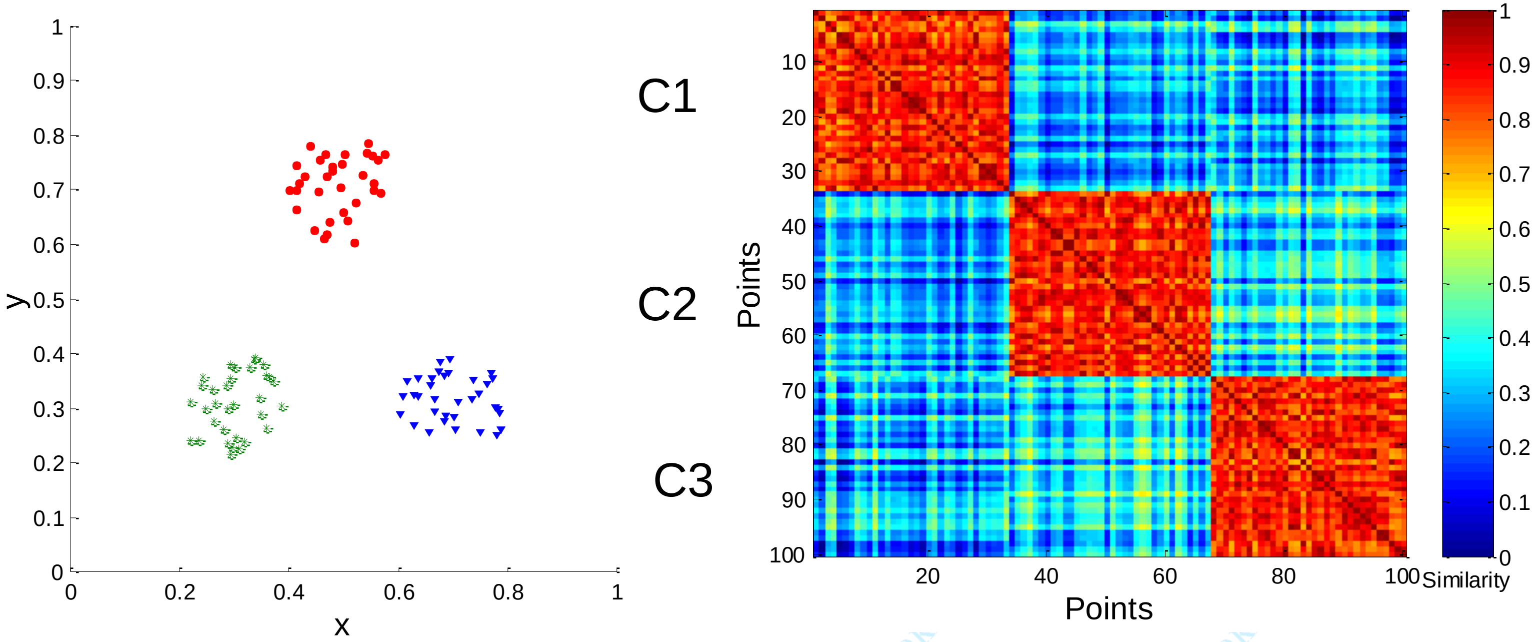
Corr = -0.9235



Corr = -0.5810

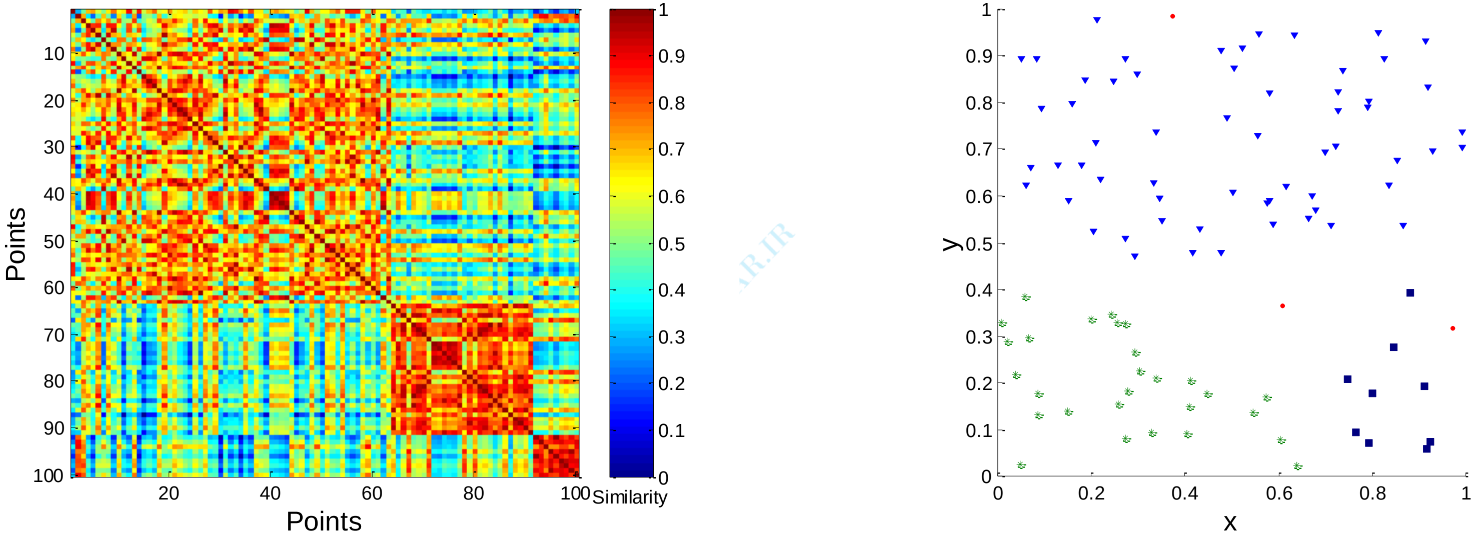
Using Similarity Matrix for Cluster Validation

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



Using Similarity Matrix for Cluster Validation

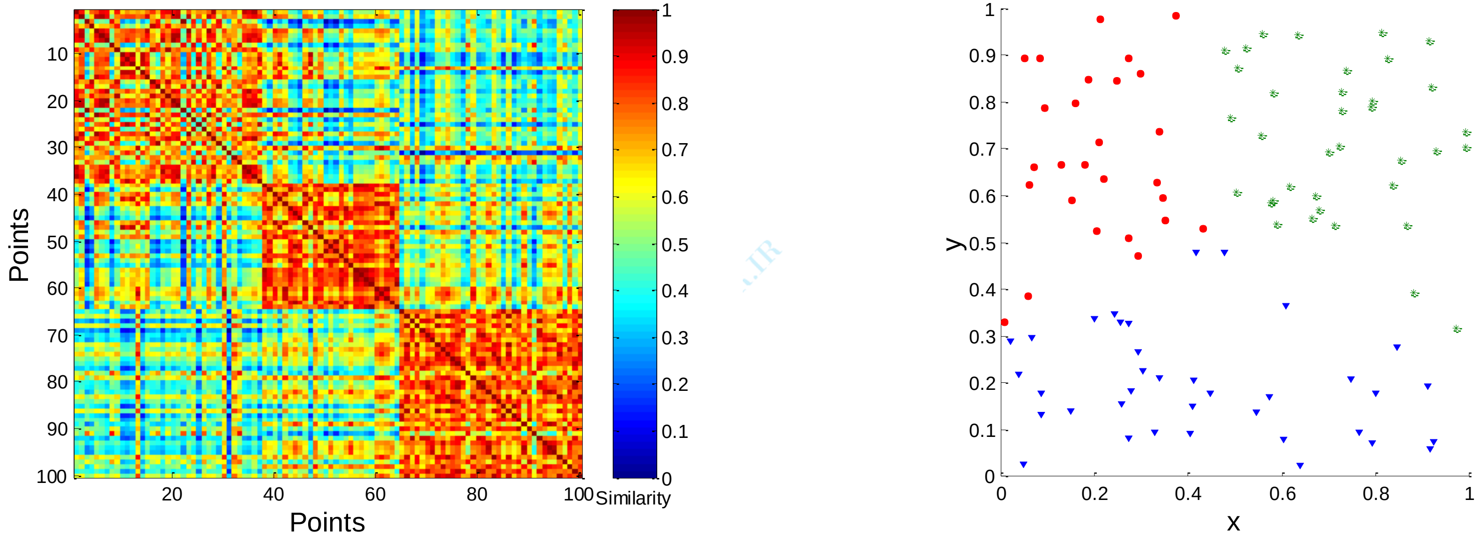
- Clusters in random data are not so crisp



DBSCAN

Using Similarity Matrix for Cluster Validation

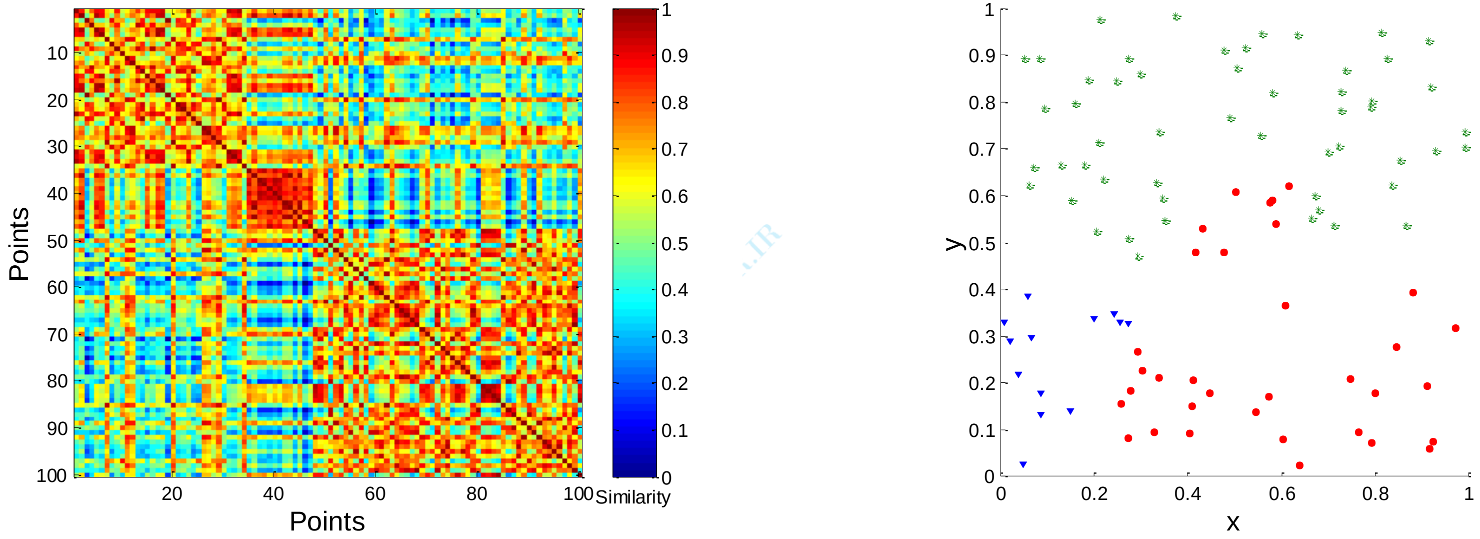
- Clusters in random data are not so crisp



K-means

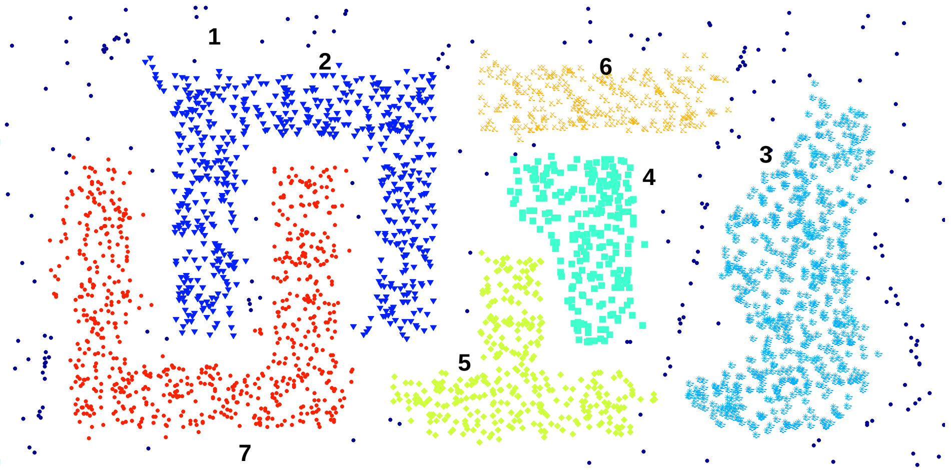
Using Similarity Matrix for Cluster Validation

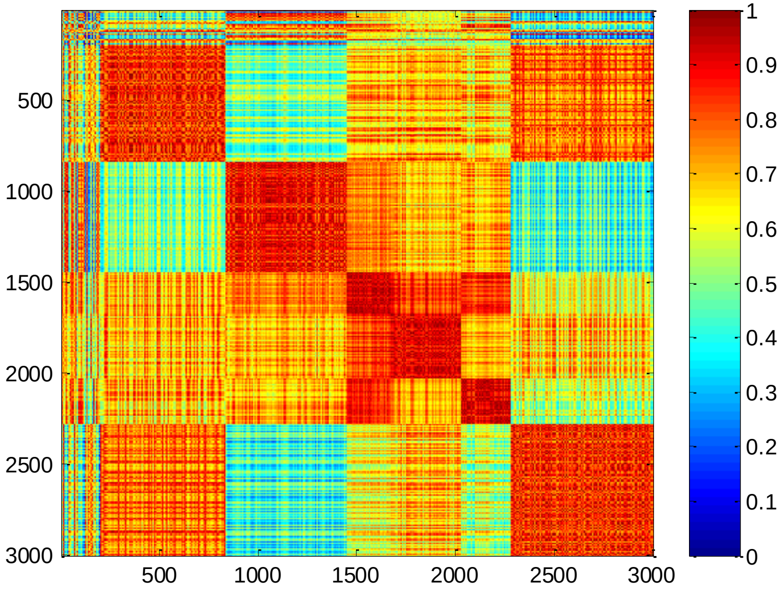
- Clusters in random data are not so crisp



Complete Link

Using Similarity Matrix for Cluster Validation

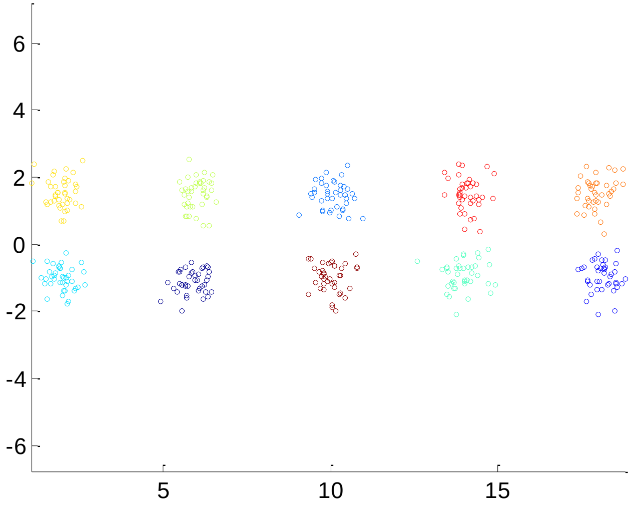


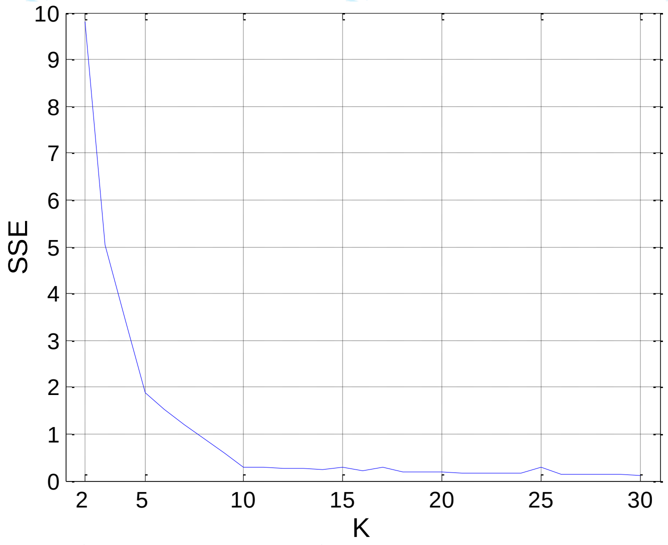


DBSCAN

Internal Measures: SSE

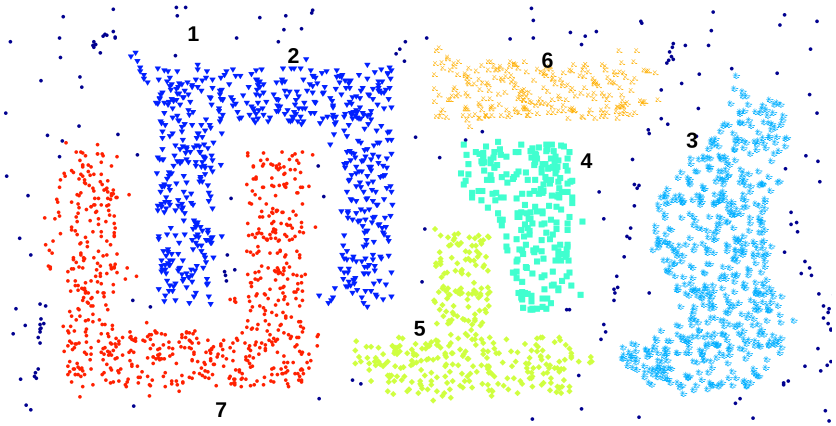
- Internal Index: Used to measure the goodness of a clustering structure without respect to external information
  - SSE
- SSE is good for comparing two clusterings or two clusters (average SSE).
- Can also be used to estimate the number of clusters

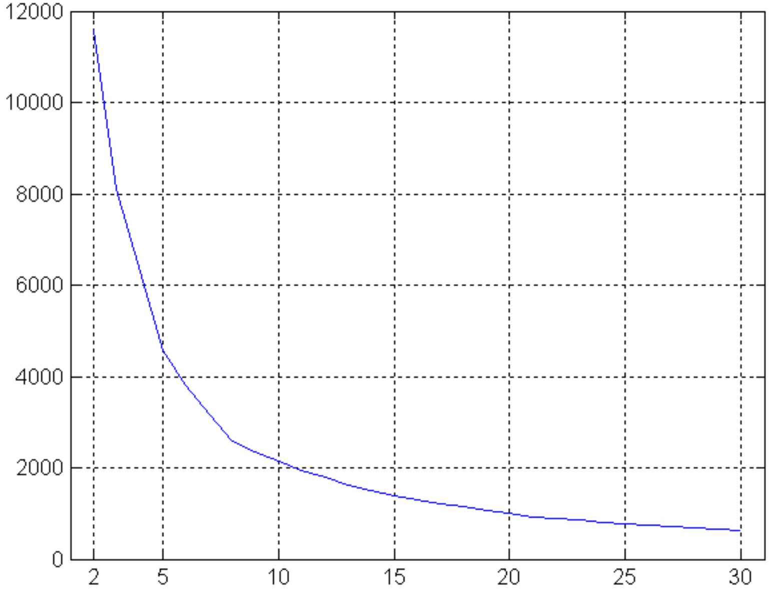




Internal Measures: SSE

- SSE curve for a more complicated data set





SSE of clusters found using K-means

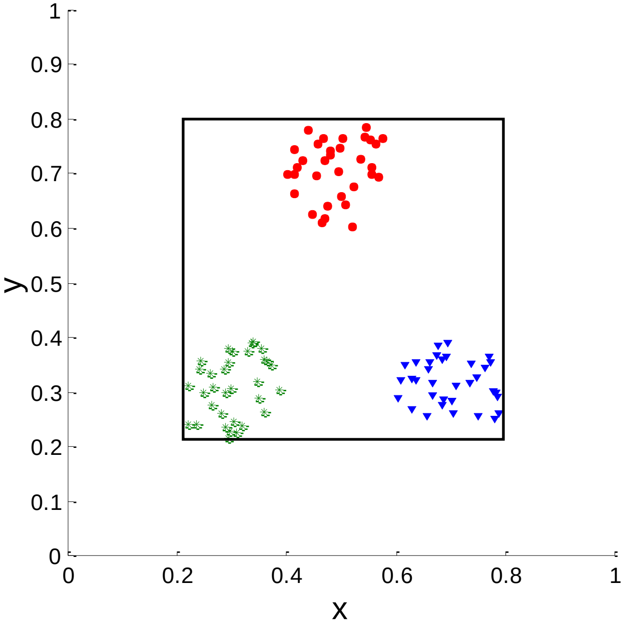
Framework for Cluster Validity

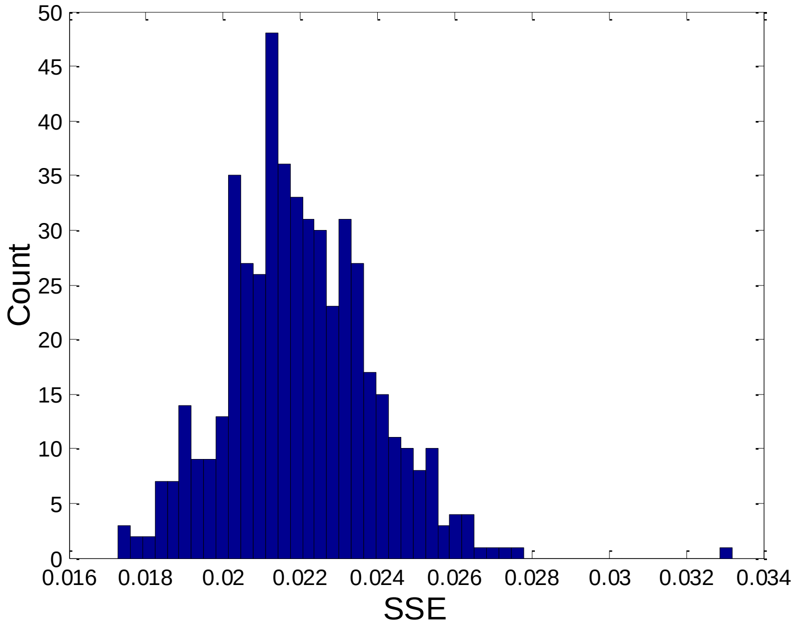
- For example, if our measure of evaluation has the value, 10, is that good, fair, or poor?

- 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

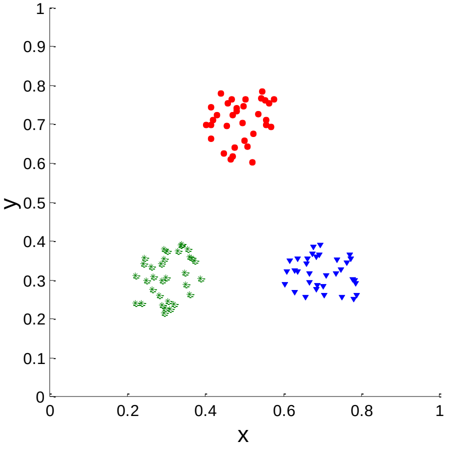
- Example
  - Compare SSE of 0.005 against three clusters in random data
  - Histogram shows SSE of three clusters in 500 sets of random data points of size 100 distributed over the range 0.2 – 0.8 for x and y values



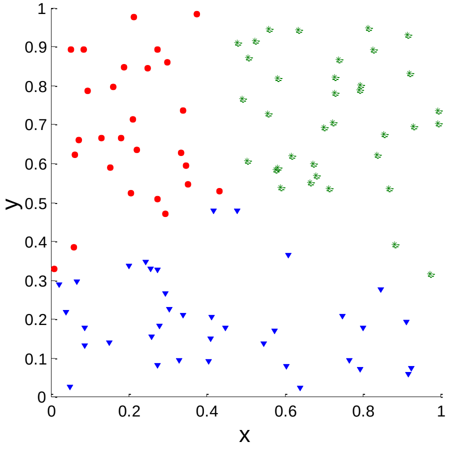


Statistical Framework for Correlation

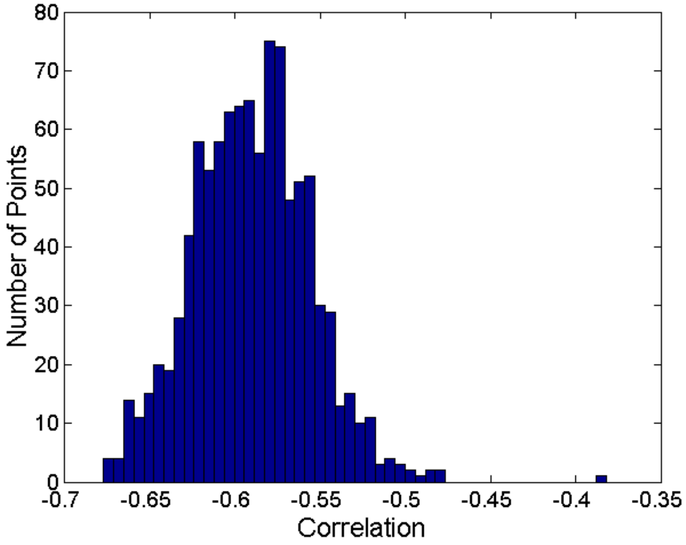
- Correlation of incidence and proximity matrices for the K-means clusterings of the following two data sets.



Corr = -0.9235



Corr = -0.5810



Internal Measures: Cohesion and Separation

- Cluster Cohesion: Measures how closely related are objects in a cluster
  - Example: SSE
- Cluster Separation: Measure how distinct or well-separated a cluster is from other clusters
- Example: Squared Error

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

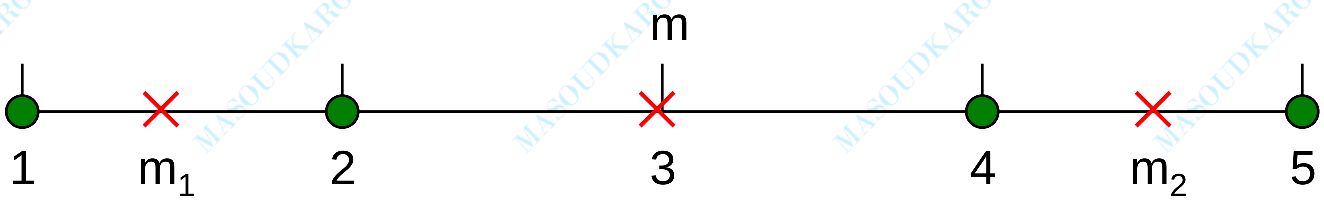
  - Cohesion is measured by the within cluster sum of squares (SSE)

$$BSS = \sum_{i} |C_i| (m - m_i)^2$$

  - Separation is measured by the between cluster sum of squares
    - Where $|C_i|$ is the size of cluster i

Internal Measures: Cohesion and Separation

- Example: SSE
  - BSS + WSS = constant



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:

$$WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

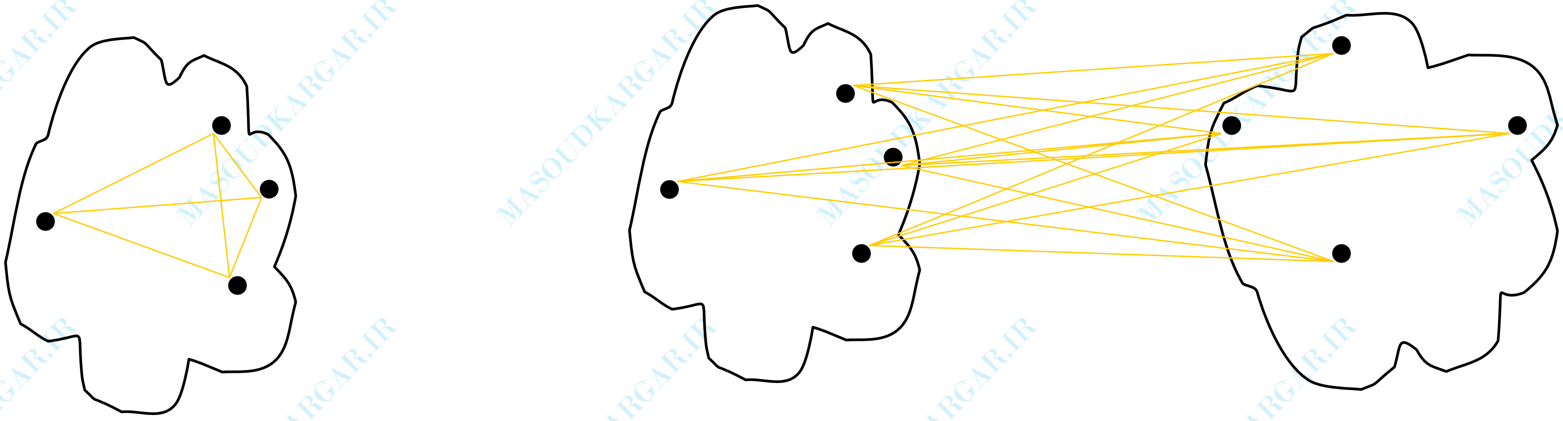
$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

$$Total = 1 + 9 = 10$$

Internal Measures: Cohesion and Separation

cohesion and separation.

- Cluster cohesion is the sum of the weight of all links within a cluster.
- Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.



cohesion

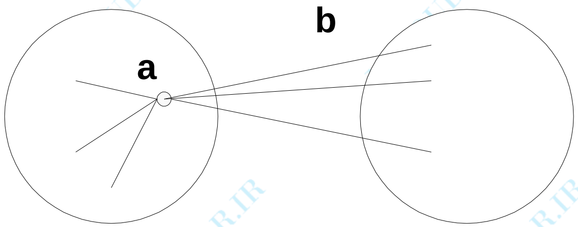
separation

Internal Measures: 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*** = average distance of *i* to the points in its cluster
  - Calculate ***b*** = min (average distance of *i* to points in another cluster)
  - The silhouette coefficient for a point is then given by

    s = 1 – a/b   if a < b,   (or s = b/a - 1   if a ≥ b, not the usual case)

  - Typically between 0 and 1.
  - The closer to 1 the better.



- Can calculate the Average Silhouette width for a cluster or a clustering

External Measures of Cluster Validity: Entropy and Purity

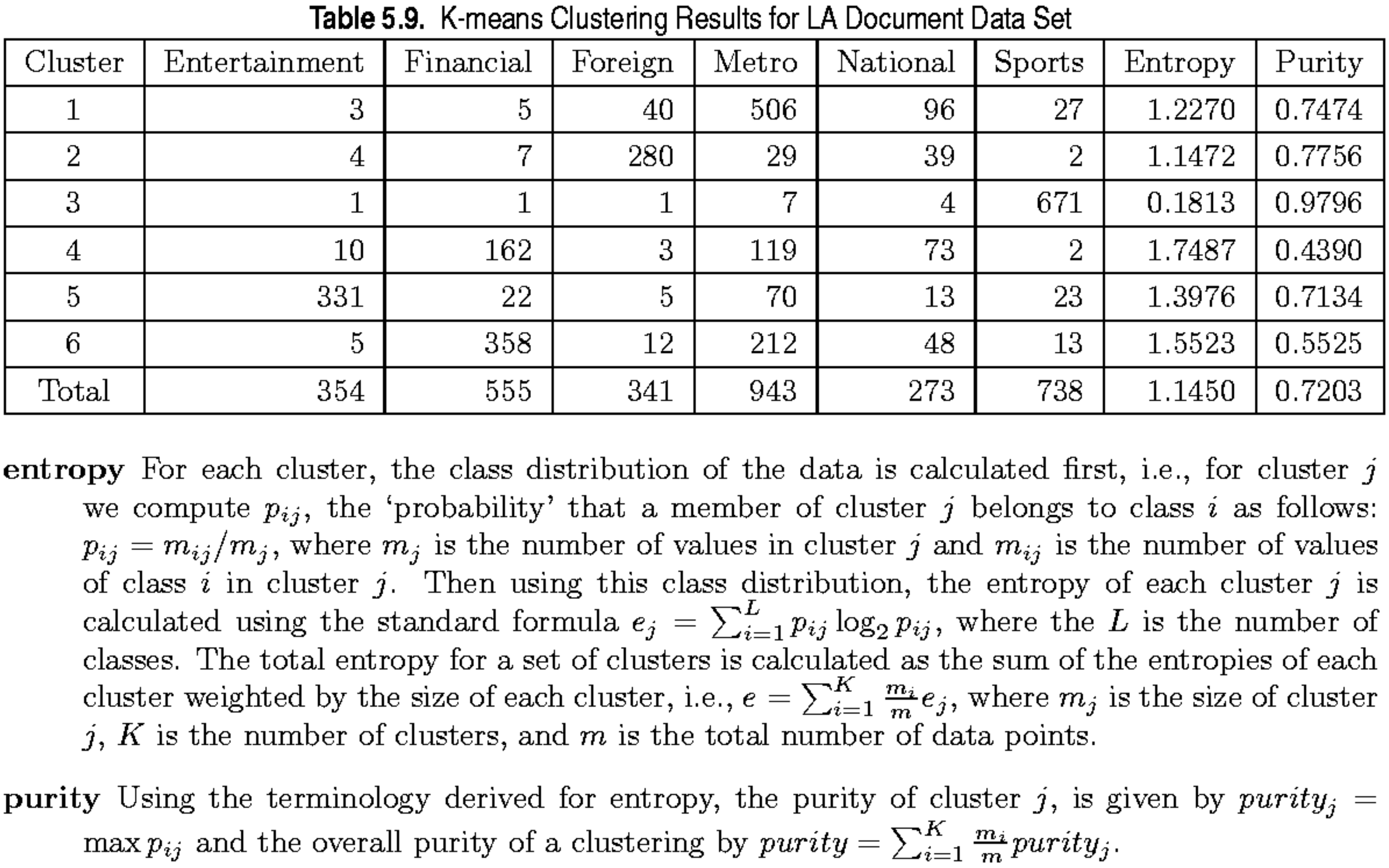


Table 5.9. K-means Clustering Results for LA Document Data Set

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

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

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

Summary

- Biclustering applications and concepts
- Available Biclustering algorithms
- Clustering algorithm evaluation
- Cluster validation

قدردانی

- Dr. Jianjun Hu
  http://mleg.cse.sc.edu/edu/csce822/
- University of South Carolina
- Department of Computer Science and Engineering