e-PGPathshala

Subject : Computer Science

Paper: Machine Learning

Module: Cluster Analysis and Cluster Validity

Module No: CS/ML/25

Quadrant I – e-text

Welcome to the e-PG Pathshala Lecture Series on Machine Learning.In this module we will discuss important issues associated with clustering through the analysis and validation of clusters.

Learning Objectives:

The learning objectives of this module are as follows:

- To understand the Cluster Validity and Cluster Validation Process
- To discuss the different Measures of Cluster Validity
- To understand External, Internal and the Statistical Framework for determining Cluster Validity

25.1 Cluster Validity

Let us first give a definition of validation of clustering. “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.”-[Jain & Dubes]

For cluster analysis, the question is how to evaluate the “goodness” of the resulting clusters? But as it is said the clusters can be defined from different perspectives. However validation is needed to compare clustering algorithms, solve the problem of determining the number of clusters, comparing two clusters, comparing two sets of clusters, and for avoiding finding patterns in noise.

Now let us understand a simple evaluation of clustering using precision and recall. Let us see the example given in Figure 25.1. Here we see that is a total of 5 oranges and 5 apples. The clustering produces two clusters, cluster 1 with 5 oranges and 2 apples and cluster 2 with 3 apples. Precision in this context is

defined as the number of items of a particular category (oranges or apples) obtained in the cluster divided by total number of items .of that category. Recall on the other hand is defined as the number of items of a particular category cluster. Accordingly precision and recall of oranges and are given in Figure 25.1. Note that these measures can be determined only if the ground truth is given. We will discuss these measures in detail later in the module.

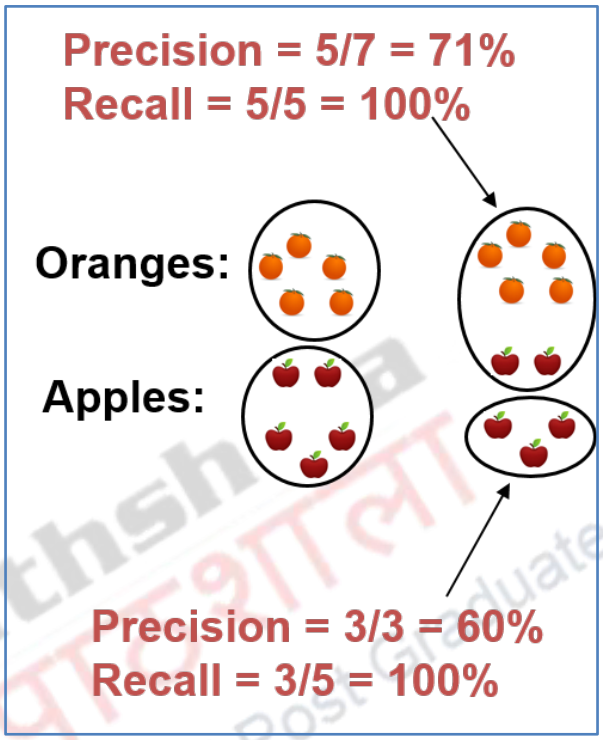

Figure 25.1 Precision and Recall of Clustering

25.2 Good Clustering

We will now discuss the issues of good clustering. First we have the internal criterion which defines good clustering as a clustering that produces high quality clusters in which the intra-class similarity is high and the inter-class similarity is low. The measured quality of a clustering depends on both the representation of the items that are to be clustered and the similarity measure used. On the other hand external criterion measures the quality of a clustering by its ability to discover some or all of the hidden patterns or latent classes available, where we compare the clustering results to gold standard data.

25.3 Aspects of Cluster Validation

Cluster validation involves evaluation of the clustering using external index by comparing the clustering results to *ground truth* (externally known results). Evaluation of the quality of clusters *without* reference to external information using only the data is called as evaluation using internal index. Another aspect is determining the *reliability* of clusters, that is the confidence level that the

clusters are not formed by chance, normally determined using a statistical framework.

25.3.1 Cluster Validation process

Now let us discuss the process of cluster validation. The following are the steps involved:

1. The first step is to find out whether the set of data has the clustering tendency that is distinguishing whether non-random structures actually exist in the data or whether the set of data is actually one cluster (Figure 25.2). Figure 25.2 shows the clusters obtained for different algorithms for random points.

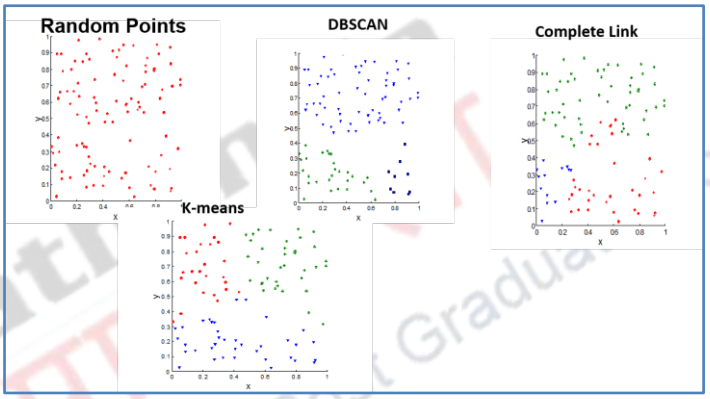

Figure 25.2 Clustering of Random Points

2. The next step is the comparison of the results of a cluster analysis to the externally known results,that is to externally given class labels.
3. The third step is the evaluation how well the results of a cluster analysis fit the data *without* reference to external information. Use only the data.
4. Then we carry out the comparison of the results of two different sets of cluster analyses to determine which is better.
5. Finally we have the important task of determining the ‘correct’ number of clusters for the given data set.

In cluster validation, the steps 2, 3, and 4, can be further distinguished dependingon whether we want to evaluate the entire clustering or just individual clusters.

According toJain & Dubes, cluster validation refers to procedures that evaluate the results of clustering in a **quantitative** and **objective** fashion. In this context evaluation is quantitative when we employ the measures for evaluation while it is objectivewhen wevalidate of the measures.

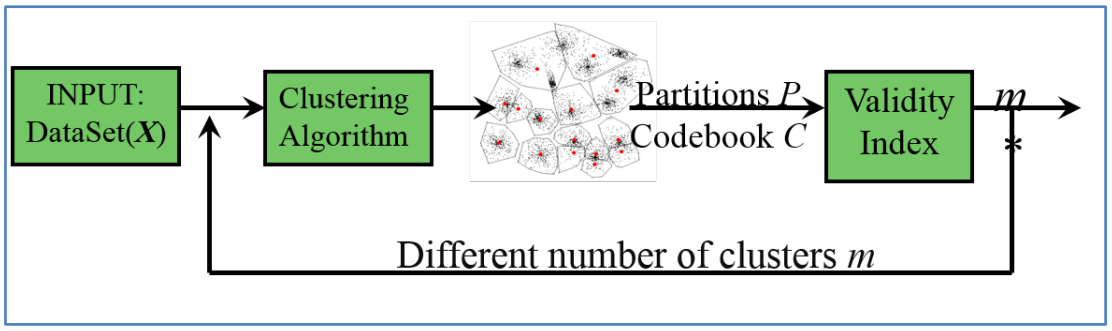

Figure 25.3 Clustering and Evaluation

Figure 25.3 shows that for a given data set X, we can apply the clustering algorithm and given the gold standard in terms of different partitions P and the associated codebook C, determine the validity index. The process is repeated by trying out with different number of clusters.

25.3.2 Measures of Cluster Validity

As we have already discussed, external index is where we validate against ground truth, or compare two clusters to find out the similarity between them (Figure 25.4). Figure 25.4 also illustrates the determination of the internal index where we validate *without*any external information, and if we do the same with different number of clusters we can solve the problem of determining the number of clusters.

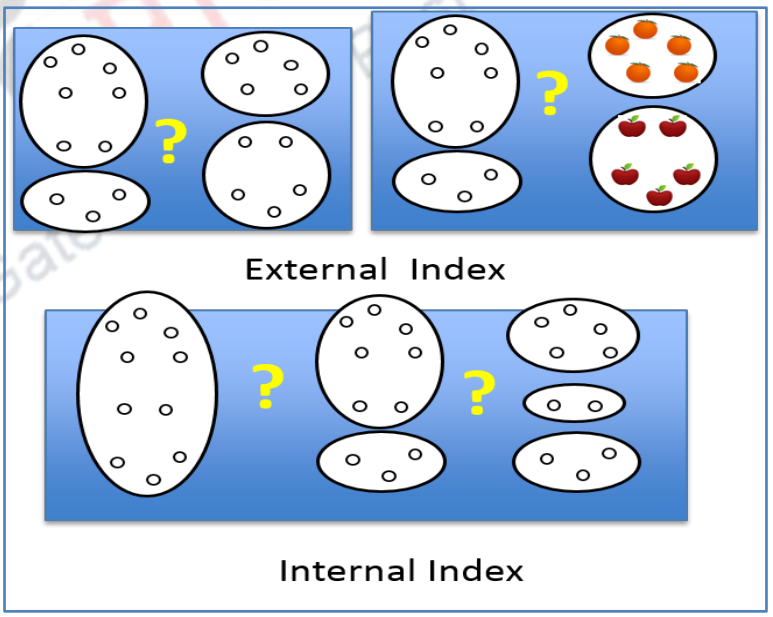

Figure 25.4External Index and Internal Index

25.4Comparing with Ground Truth

Now let notation for comparing with ground truth be as follows:
N: number of objects in the data set;

$P=\{P_1,\ldots,P_m\}$: the set of "ground truth" clusters;
$C=\{C_1,\ldots,C_n\}$: the set of clusters reported by a clustering algorithm.
Now we create the "incidence matrix", which is a $N \times N$ matrix where both the rows and columns correspond to objects. $P_{ij} = 1$ if object O_i and object O_j belong to the same "ground truth" cluster in P; $P_{ij}=0$ otherwise. Similarly $C_{ij} = 1$ if object O_i and object O_j belong to the same cluster in C; $C_{ij}=0$ otherwise.

A pair of data object (O_i,O_j) falls into one of the following categories :

- SS: $C_{ij}=1$ and $P_{ij}=1$; (agree)
- DD: $C_{ij}=0$ and $P_{ij}=0$; (agree)
- SD: $C_{ij}=1$ and $P_{ij}=0$; (disagree)
- DS: $C_{ij}=0$ and $P_{ij}=1$; (disagree)

Now we define two evaluation measures based on this incidence matrix, Rand index and Jaccard coefficient. Rand Index may be dominated by DD, which is about objects not belonging to the same cluster in either the ground truth case or in the clusters obtained through the clustering algorithm.

$$Rand = \frac{|Agree|}{|Agree|+|Disagree|} = \frac{|SS|+|DD|}{|SS|+|SD|+|DS|+|DD|}$$

$$Jaccard\ \ coefficient = \frac{|SS|}{|SS|+|SD|+|DS|}$$

25.5 Purity Based Measures

Now let us discuss another method used to find external index. Here we assume that the class of each object is available. The confusion matrix is given in Figure 25.5

	Class 1	Class 2	Class 3	
Cluster 1	n_{11}	n_{12}	n_{13}	m_1
Cluster 2	n_{21}	n_{22}	n_{23}	m_2
Cluster 3	n_{31}	n_{32}	n_{33}	m_3
	c_1	c_2	c_3	n

	Class 1	Class 2	Class 3	
Cluster 1	p_{11}	p_{12}	p_{13}	m_1
Cluster 2	p_{21}	p_{22}	p_{23}	m_2
Cluster 3	p_{31}	p_{32}	p_{33}	m_3
	c_1	c_2	c_3	n

Figure 25.5 Confusion Matrix

given in Figure 25.5. Here the details are as follows:

- n = the number of points
- m_i = the points in cluster i
- c_j = the points in class j
- n_{ij} = the points in cluster i coming from class j
- $p_{ij} = n_{ij}/m_i$ = probability of element from cluster i to be assigned in class j

25.5.1 Purity Based Measures

We can define two measures entropy and purity. Both these measures can be measured for a cluster or for a clustering which is the evaluation of the complete set of clusters. Entropy of a cluster is based on probability p_{ij} of element from cluster i being assigned to class j. Here L is the total number of classes and K the total number of clusters.

Entropy of a cluster i: $\boldsymbol{e_i = -\sum_{j=1}^{L} p_{ij}\, log\, p_{ij}}$.

Entropy of a clustering: $\boldsymbol{e = \sum_{i=1}^{K} \frac{m_i}{n} e_i}$

Entropy of a clustering is based on average entropy of each cluster as well as the total number of objects in the data set n. This value is highest when objects belong uniformly across clusters and zero when objects belong to a single cluster.

Purity of a cluster i is the class j for which the probability p_{ij} that is the probability that an element from cluster i is assigned to class j.

Purity of a cluster i: $\boldsymbol{p_i = max_j\, p_{ij}}$

Purity of a clustering: $\boldsymbol{p(C) = \sum_{i=1}^{K} \frac{m_i}{n} p_i}$

The purity of the clustering C is based on average of the purity of each cluster p_i, m_i the points in cluster i and the total number of objects in the data set n.

25.5.2 Precision and Recall

Precision of cluster i with respect to class j: $\boldsymbol{Prec(i,j) = p_{ij}}$ where $\boldsymbol{p_{ij}}$ is the probability of element from cluster i to be assigned in class j

Recall of cluster i with respect to class j: $\boldsymbol{Rec(i,j) = \frac{n_{ij}}{c_j}}$where n_{ij} is the number of data points in cluster i coming from class j and c_j is the total number of data points in class j

F-measure is defined as the Harmonic Mean of Precision and Recall:

$$\boldsymbol{F(i,j) = \frac{2 * Prec(i,j) * Rec(i,j)}{Prec(i,j) + Rec(i,j)}}$$

25.5.3 Precision/Recall for Clusters and Clustering

We normally assign to cluster i the class $k_i$which is the class j such that cluster *i has the maximum number of data* points coming from class j

$$\boldsymbol{k_i = arg\,max_j\, n_{ij}}$$

Now we define the **Precision of cluster i as** $\boldsymbol{Prec(i) = \frac{n_{ik_i}}{m_i}}$where n_{ik_i} is the number of data points in cluster i coming from class $\boldsymbol{k_i}$and $m_i$is the total number of data points in cluster i.

Precision of the clustering is defined as $\boldsymbol{Prec(C) = \sum_i \frac{m_i}{n} Prec(i)}$where $m_i$is the total number of data points in cluster i, n is total number of data points and $\boldsymbol{Prec(i)}i$s the precision of each cluster i.

Recall of cluster is defined as $\boldsymbol{Rec(i) = \frac{n_{ik_i}}{c_{k_i}}}$where n_{ik_i} is the number of data points in cluster i coming from class $\boldsymbol{k_i}$and$\boldsymbol{c_{k_i}}$is the total number of data points in class $\boldsymbol{k_i}$

Recall of the clustering defined as$\boldsymbol{Rec(C) = \sum_i \frac{m_i}{n} Rec(i)}$where $m_i$is the total number of data points in cluster i, n is total number of data points and $\boldsymbol{Rec(i)}i$s the precision of each cluster i.As normally defined F-measure is the harmonic mean of precision and recall.

25.5.4 Good and Bad Clustering

Now let us understand the evaluation using the examples given in Figure 25.6. We had discussed that the p**urity of a cluster i** is the class j for which the probability of p_{ij}, an element from cluster i being assigned to class j. Purity of a cluster i: $\boldsymbol{p_i = max_j\, p_{ij}}$**.** In Figure 25.6 (a), the purity of cluster 1 is Max value of cluster divided by the total number of data points in class 1 i.e. 85/90=0.94, that

of cluster 2 is 90/110=0.81, and that of cluster 3 is 85/100=0.85. The overall purity of the clustering is 0.86. Similarly we can find the precision and recall for each cluster and for the overall clustering. Figure 25.6 (b) shows another example with the calculated evaluation measures. As we can see in Figure 25.6 (a) a good clustering has high values of purity, precision and recall while a bad clustering. of Figure 25.6 (b) has low values of purity, precision and recall is

	Class 1	Class 2	Class 3	
Cluster 1	2	3	85	90
Cluster 2	90	12	8	110
Cluster 3	8	85	7	100
	100	100	100	300

Purity: (0.94, 0.81, 0.85) - overall 0.86
Precision: (0.94, 0.81, 0.85) - overall 0.86
Recall: (0.85, 0.9, 0.85) - overall 0.87

(a)

	Class 1	Class 2	Class 3	
Cluster 1	20	35	35	90
Cluster 2	30	42	38	110
Cluster 3	38	35	27	100
	100	100	100	300

Purity: (0.38, 0.38, 0.38) - overall 0.38
Precision: (0.38, 0.38, 0.38) – overall 0.38
Recall: (0.35, 0.42, 0.38) – overall 0.39

(b)

Figure 25.6 Good and Bad Clustering

25.6 Internal Measure

As discussed previously internal measures are used to measure the goodness of a clustering structure without comparing with external information. There are basically two aspects that are considered namely **cohesion**whichmeasures how closely related the data points in a cluster are and **separation**which measures how distinct or well-separated the data points of a cluster are from the data points of other clusters (Figure 25.7). These aspects can be measured using sum of squares and sum of squared error.

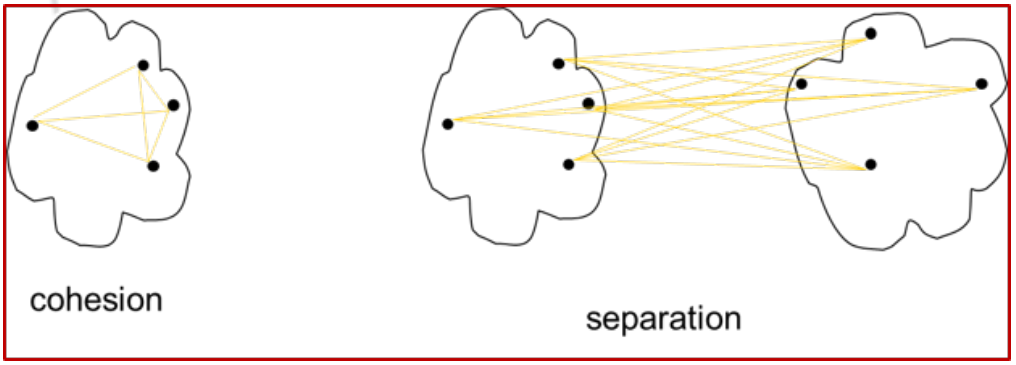

Figure 25.7 Cohesion and Separation

25.6.1 Internal Measure -Sum of Squares

When we measure cluster cohesion, we measure homogeneity measured by the within cluster sum of squares

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

Exactly the objective function of K-means.

Cluster Separation is measured by the between cluster sum of squares

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

Where |Ci| is the size of cluster i, m is the centroid of the whole data set.

The sum of BSS + WSS = constant. In general, a larger number of clusters tend to result in smaller WSS as shown in the example (Figure 25.8)

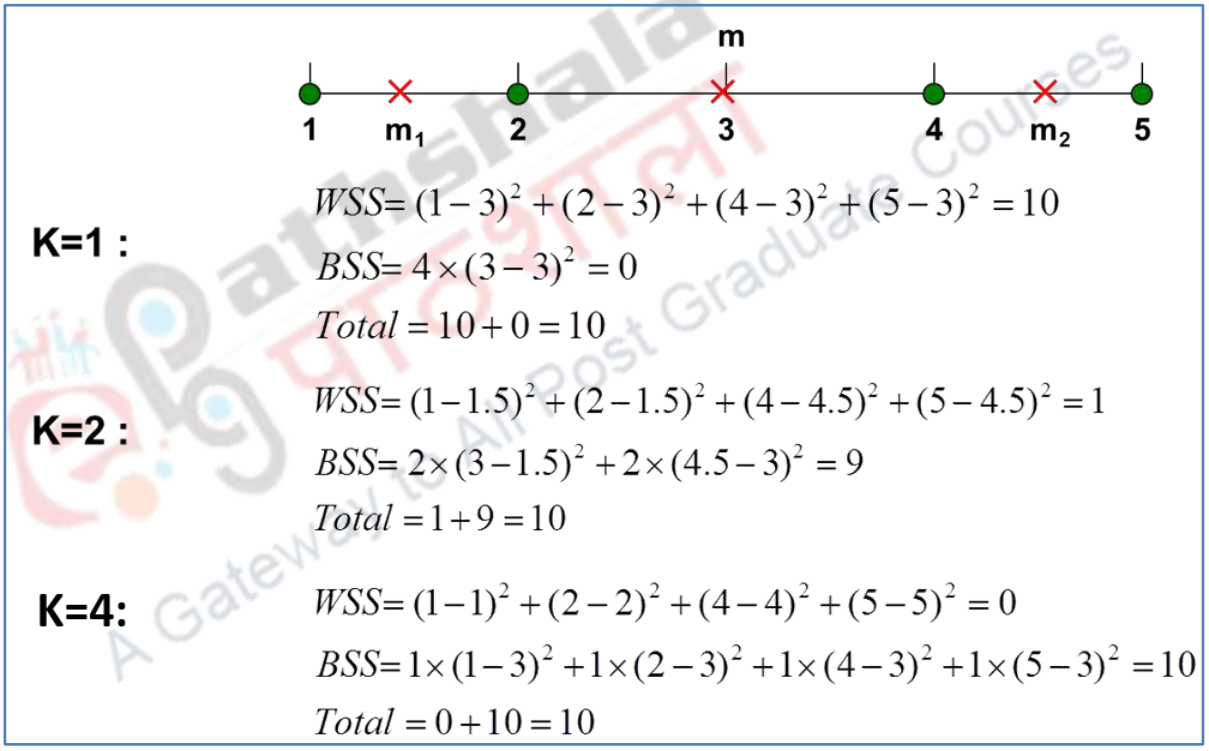

Figure 25.8 Example of WSS an BSS with Different Number of Clusters

Internal Measure SSE curve for a more complicated data set is shown in Figure 25.9.

.

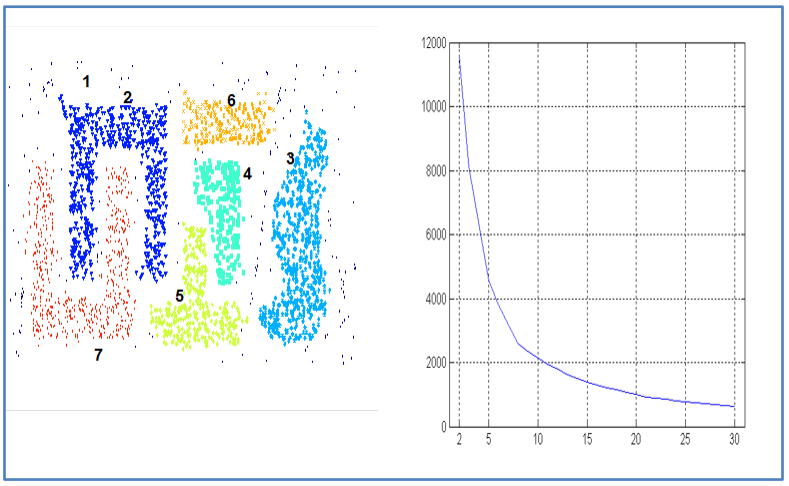

Figure 25.9 SSE for a Complex Data Set

25.6.2 Internal Measure –Average Distance

In the same spirit as the sum of squareswe alsodefine ***Cohesion a*(*x*)**as theaverage distance of *x* to all other vectors in the same cluster. We also define ***Separation b*(*x*)**as theaverage distance of *x* to the vectors in other clusters. We find the minimum among the clusters

25.6.3 Internal Measure -Silhouette coefficient

Another important internal measure is the**SilhouetteCoefficients(x)**which is defined below. We first define silhouette as:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

s(*x*) = [-1, +1]: -1=bad, 0=indifferent, 1=good

We use the above to define the Silhouette coefficient (SC) to find the silhouette of all the N data points :

$$SC = \frac{1}{N}\sum_{i=1}^{N} s_i(x)$$

25.7 Correlation with Distance Matrix

We will down define two matrices, the proximity/distance matrix and the incidence matrix. The proximity matrix /distance matrix is a matrix having rows and columns equal to the number of objects where each entry D_{ij} is the similarity between object O_i and O_j. The incidence matrix has one row and one column for each data point and the entry is 1 if the associated pair of points belong to the same cluster and 0 if the associated pair of points belong to different clusters.

Now we need to compute the correlation between the two matrices. Here we need to calculate only n(n-1)/2 entries. High correlation indicates good clustering.

Given Distance Matrix D = $\{d_{11}, d_{12}, \ldots, d_{nn}\}$ and Incidence Matrix $C = \{c_{11}, c_{12}, \ldots, c_{nn}\}$.

Correlation *r* between D and C is given by

$$r = \frac{\sum_{i=1,j=1}^{n} (d_{ij} - \bar{d})(c_{ij} - \bar{c})}{\sqrt{\sum_{i=1,j=1}^{n} (d_{ij} - \bar{d})^2} \sqrt{\sum_{i=1,j=1}^{n} (c_{ij} - \bar{c})^2}}.$$

This is not a good measure for some density or contiguity based clusters.

Correlation of incidence and proximity matrices for the K-means clustering of two data sets with different values is shown in Figure 25.10.

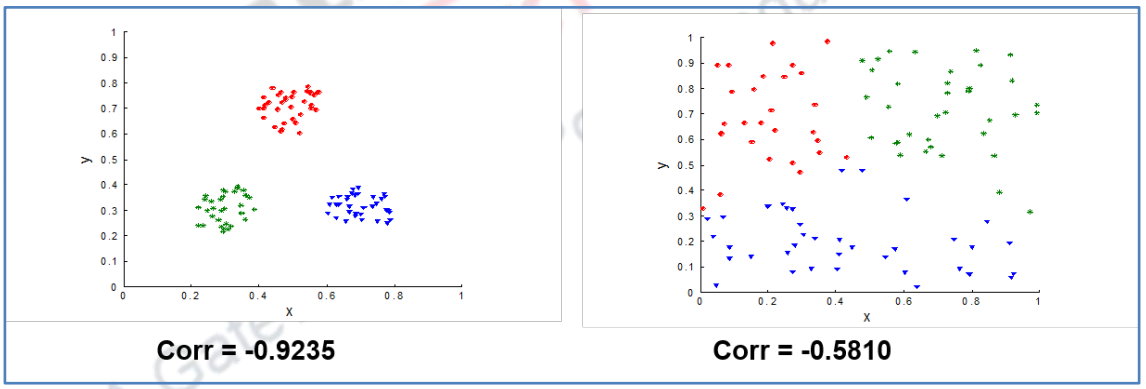

Figure 25.10. Correlation Values Using Distance Matrix

25.7.1 Using Similarity Matrix for Cluster Validation

We can order the similarity matrix with respect to cluster labels and inspect visually. This is shown in Figure 25.11 which is a good clustering. Clusters in random data are not so crisp and is shown in Figure 25.12.

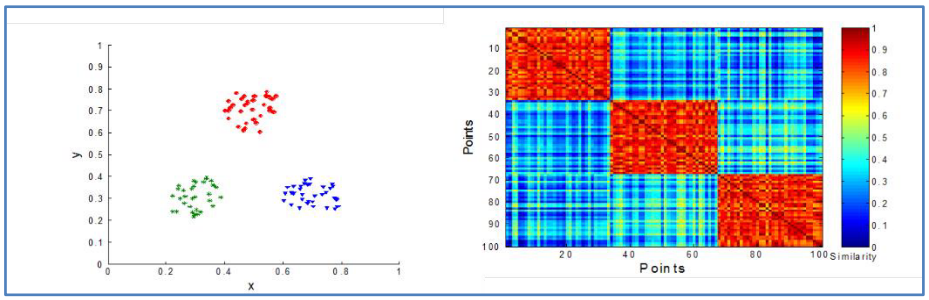

Figure 25.11 Good Clustering

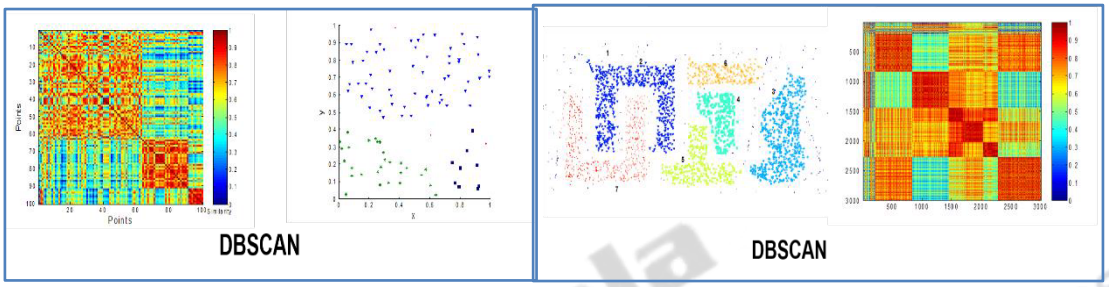

Figure 25.12Example of Clustering

25.8 Statistical Framework for SSE

Now let us discuss some statistical aspects. Now let us consider the example where we compare the SSE (discussed previously) of 0.005 against three clusters in random data. The figure shows the SSE Histogram of 500 sets of random data points of size 100 distributed over the range 0.2 – 0.8 for x and y values.

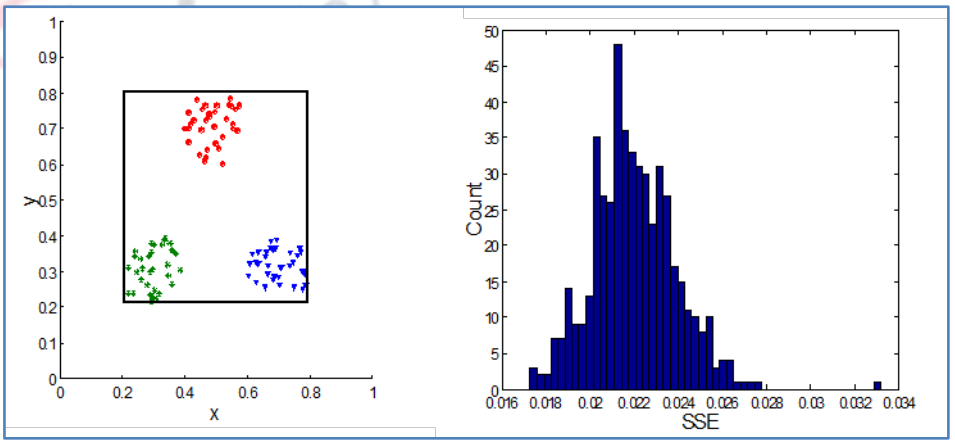

Figure 25.13 Histogram of Correlation with SSE

Correlation of incidence and distance matrices for the K-means of the following two data sets and the Histogram is shown in Figure 25.14.

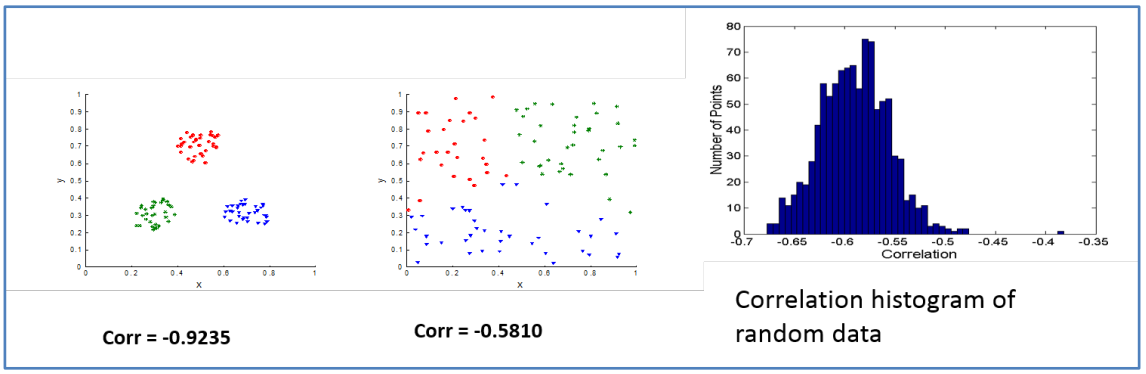

Figure 25.14 Histogram of Correlation with K means

25.8.1 Empirical p-value

Another measure of correlation is the empirical p-value. If we have a measurement v (e.g., the SSE value) and we have N measurements on random datasets, the empirical p-value is the fraction of measurements in the random data that have value less or equal than value v (or greater or equal if we want to maximize), in other words the value in the random dataset is at least as good as that in the real data. We usually require that p-value ≤ 0.05. However it is difficult to know the right notion of a random dataset.

25.8.2 Hyper Geometric Distribution

The next two measures of correlation or association are defined in terms of genes in a random data set. Given that the total number of genes in the data set associated with term T is M, if we randomly draw n genes from the data set N, the probability that m of the selected n genes will be associated with T is given as

$$\Pr(m \mid N, M, n) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$$

25.8.3 P-Value

Based on Hyper-geometric distribution given above, the probability of having m genes or fewer associated to T in N can be calculated by summing the probabilities of a random list of N genes having 1, 2, …, m genes associated to T. So the p-value of over-representation is as follows:

$$p = \sum_{i=m}^{\min(M,n)} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Summary

- Explained the different aspects of Cluster Validity and the Cluster Validation Process
- Discussed the roles of the different Measures in determining Cluster Validity
- Explained External, Internal and the Statistical Framework used for determining Cluster Validity