CSCI 550: Adv. Data Mining

13- Clustering Validation

MONTANA STATE UNIVERSITY

Clustering Validation and Evaluation

Cluster validation and assessment encompasses three main tasks:

- Clustering evaluation seeks to assess the goodness or quality of the clustering
- clustering stability seeks to understand the sensitivity of the clustering result to various algorithmic parameters, for example, the number of clusters
- clustering tendency assesses the suitability of applying clustering in the first place, that is, whether the data has any inherent grouping structure.

Clustering Validation and Evaluation

Validity measures can be divided into three main types:

- External: External validation measures employ criteria that are not inherent to the dataset. This can be in form of prior or expert-specified knowledge about the clusters e.g., class labels.
- Internal: Internal validation measures employ criteria that are derived from the data itself. For instance, we can use intracluster and intercluster distances to obtain measures of cluster compactness (e.g., how similar are the points in the same cluster?) and separation (e.g., how far apart are the points in different clusters?).
- Relative: Relative validation measures aim to directly compare different clusterings, usually those obtained via different parameter settings for the same algorithm.

External Measures

- External measures assume that the correct or ground-truth clustering is known *a priori*, which is used to evaluate a given clustering.
- Let D be a dataset consisting of n points in a d-dimensional space, partitioned into k clusters. Let $y_i \in \{1,2, . . . ,k\}$ denote the ground-truth cluster membership or label information for each point.
- The ground-truth clustering is given as $T = \{T_1,T_2, . . . ,T_k\}$, where the cluster T_j consists of all the points with label j , i.e., $T_j = \{x_i \in D|y_i = j\}$. We refer to T as the ground-truth partitioning, and to each T_i as a partition.
- Let C = {C1, . . . ,Cr} denote a clustering of the same dataset into r clusters,
- obtained via some clustering algorithm, and let $\hat{y}_i \in \{1,2, . . . , r\}$ denote the cluster label for x_i .
- Because the ground truth is assumed to be known, typically clustering methods will be run with the correct number of clusters, that is, with r = k.
- However, to keep the discussion more general, we allow r to be different from k

External Measures

- External evaluation measures try capture the extent to which points from the same partition appear in the same cluster, and the extent to which points from different partitions are grouped in different clusters.
- There is usually a trade-off between these two goals

- All of the external measures rely on the r × k contingency table N that is induced by a clustering C and the ground-truth partitioning T , defined as follows: $N(i , j) = n_{ij} = |C_i \cap T_j |$
- The count n_{ij} denotes the number of points that are common to cluster C_i and ground-truth partition T_j .
- Let $n_i = |C_i|$ denote the number of points in cluster C_i ,and let $m_j = |T_j |$ denote the number of points in partition T_j .

Matching Based Measures: Purity or Precision and Max Matching or Recall

- Purity quantifies the extent to which a cluster C_i contains entities from only one partition:

$$purity_i = \frac{1}{n_i} \max_{j=1}^{k} \{n_{ij}\}$$

- The purity of clustering C is defined as the weighted sum of the clusterwise purity values:

$$purity = \sum_{i=1}^{r} \frac{n_i}{n} purity_i = \frac{1}{n} \sum_{i=1}^{r} \max_{j=1}^{k} \{n_{ij}\}$$

- where the ratio n_i / n denotes the fraction of points in cluster Ci .

The recall of cluster Ci is defined a

$$recall_i = \frac{n_{ij_i}}{|T_{j_i}|} = \frac{n_{ij_i}}{m_{j_i}}$$

where $m_{ji} = |T_{ji}|$.

Matching Based Measures: Purity or Precision and Max Matching or Recall

iClicker

1- What is the total purity of a clustering if each data point falls in a separate cluster:

A. 0

B. 1

C. between 0 and 1

D. none of above

2- What is total recall when all points are clustered in one cluster:

A. 0

B. 1

C. between 0 and 1

D. none of above

Matching Based Measures: Purity or Precision and Max Matching or Recall

Discussion:

Is there a trade-off between purity and recall?

Matching Based Measures: F-measure

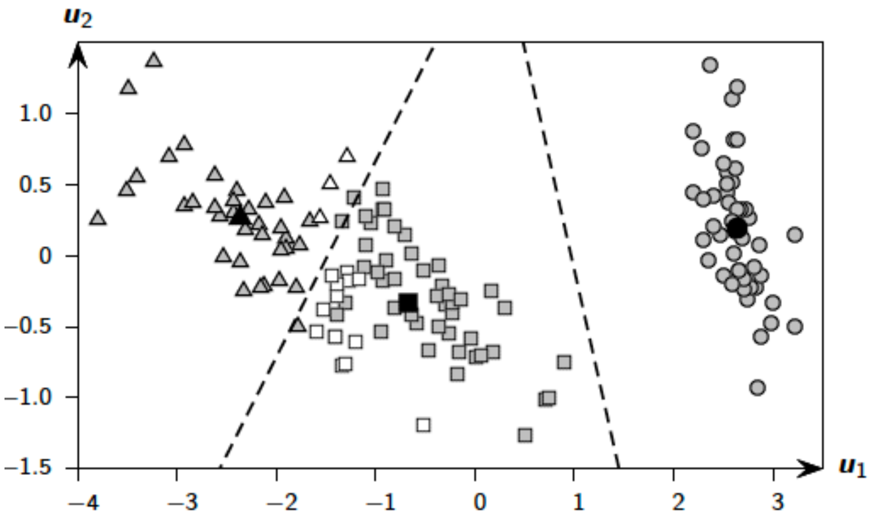
- The F-measure is the harmonic mean of the precision and recall values for each C_i

$$F_i = \frac{2}{\frac{1}{prec_i} + \frac{1}{recall_i}} = \frac{2 \cdot prec_i \cdot recall_i}{prec_i + recall_i} = \frac{2\, n_{ij_i}}{n_i + m_{j_i}}$$

- The F-measure for the clustering C is the mean of clusterwise F-meaure values:

$$F = \frac{1}{r} \sum_{i=1}^{r} F_i$$

K-means: Iris Principal Components Data

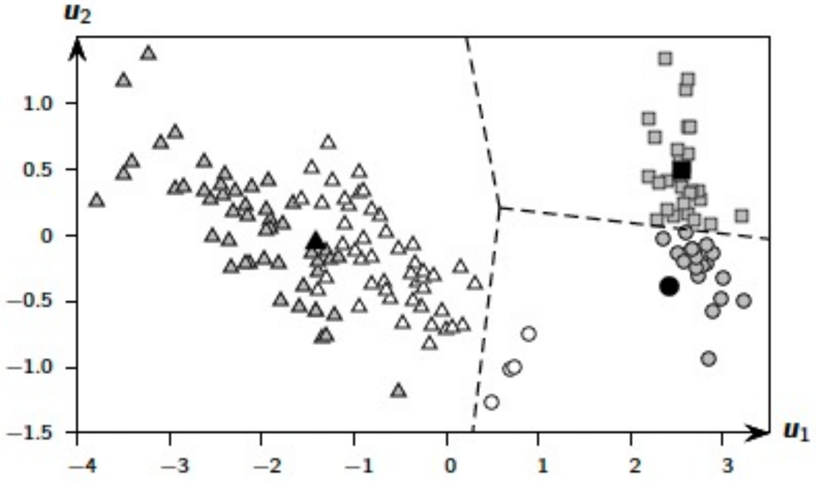


Contingency table:

	iris-setosa	iris-versicolor	iris-virginica	
	T_1	T_2	T_3	n_i
C_1(squares)	0	47	14	61
C_2(circles)	50	0	0	50
C_3(triangles)	0	3	36	39
m_j	50	50	50	$n = 100$

$purity = 0.887$, $match = 0.887$, $F = 0.885$.

K-means: Iris Principal Components Data



Contingency table:

	iris-setosa	iris-versicolor	iris-virginica	
	T_1	T_2	T_3	n_i
$C_1(squares)$	30	0	0	30
$C_2(circles)$	20	4	0	24
$C_3(triangles)$	0	46	50	96
m_j	50	50	50	$n = 150$

$purity = 0.667,\ match = 0.560,\ F = 0.658$

Other External Validating Techniques

- Entropy-based Measures:
  - Conditional Entropy
  - Normalized Mutual Information
  - Variation of Information
- Pairwise Measures:
  - Jaccard Coefficient
  - Rand Statistic
  - Fowlkes–Mallows (FM) Measure
- Correlation Measures:
  - Hubert statistic

Internal Measures

- Internal evaluation measures do not have recourse to the ground-truth partitioning, which is the typical scenario when clustering a dataset.
- To evaluate the quality of the clustering, internal measures therefore have to utilize notions of intracluster similarity or compactness, contrasted with notions of intercluster separation, with usually a trade-off in maximizing these two aims.

- BetaCV and C-index
- Normalized Cut and Modularity
- Dunn Index
- Davies-Bouldin Index

Relative Measures: Silhouette Coefficient

- Relative measures are used to compare different clusterings obtained by varying different parameters for the same algorithm, for example, to choose the number of clusters k.
- The silhouette coefficient is a clustering validation metric that measures the similarity of an object to its own cluster (cohesion) compared to other clusters (separation).
- For a data point Xj, the silhouette coefficient can be computed as follows. Let aj be the mean distance between data point Xj and all the other points in the cluster that Xj belongs to. And let bj be the mean distance between data point Xj and all the points in the next nearest cluster. Then the silhouette coefficient for point Xj is given by:

$$s_j = \frac{b_j - a_j}{\max(a_j, b_j)} . \quad SC = \frac{1}{n} \sum_{i=1}^{n} s_i .$$

Relative Measures: Silhouette Coefficient

- When this is averaged over all the points in a given cluster, it measures how tightly grouped the points in that cluster are. When it is averaged over all points in the data set, it measures how well the data have been clustered.

iClicker

- What is the range of values for the silhouette coefficient?

A. [0, 1]
B. [-1,1]
C. [0, ∞]
D. [-∞, ∞]

- When will the silhouette coefficient of a point be close to 1? When will it be close to 0? What does it mean if the Si value is negative? Answer all these questions for SC of each cluster and total SC value

Other Relative Measures

- Calinski–Harabasz Index
- Gap Statistic