Evaluation of Clustering Algorithms

Dr. Sharu Theresa Jose

University of Birmingham

February 26, 2024

Learning Outcomes

- Understand the importance of cluster analysis
- Familiarize with some commonly used cluster validation criteria

Overview of Lecture

- Introduction to Cluster Evaluation
- Cluster Validation Criteria
  - Unsupervised, Supervised and Relative Validation Criteria
- Unsupervised and Supervised Validation Criteria

Introduction

- Supervised learning has well-accepted evaluation measures and procedures (e.g., accuracy, cross-validation).
- In contrast, cluster evaluation (or validation) is not trivial.
- Nevertheless, cluster validation is important: every clustering algorithm will find clusters in a dataset, even if data has no natural clustering structure.

Cluster Validation Can Help Answer...

- Is there a clustering tendency in the observed data, i.e., determine whether non-random structure (or natural grouping) exists in the data?
- Can we evaluate how well the results of a clustering algorithm fit the data (or natural grouping) **without** external information?
- Can we evaluate how well the results of a clustering algorithm fit the data **with** external information?
- Can we compare two sets of clusters to determine which is better?
- Can we determine the correct number of clusters?

What is Cluster Validation?

- Goal: Evaluate in a quantitative and objective manner the cluster structure found by an algorithm according to a validation criterion
- Validation criterion: Index used to measure the adequacy of the found cluster structures.
- Adequacy refers to the sense in which the found cluster structure provides true information about the data or reflect the intrinsic character of the data.

Types of Cluster Validation Criteria

- **Unsupervised (Internal Indices)**
  - Measures goodness of a clustering structure **without** reference to external information
  - Example: WCSS
  - Can be further divided into two classes: intra-cluster and inter-cluster similarity indices.
  - Can also be used to estimate the optimal number of clusters (e.g., elbow method).
- **Supervised (External Indices)**
  - Measures the extent to which a clustering algorithm matches some externally supplied information.
  - Example: Entropy (how well cluster labels match with externally supplied class labels), also recall classes-to-cluster evaluation in Weka
- **Relative**
  - Compares two different sets of clusters or algorithms.
  - Can be a supervised or unsupervised criteria used for the purpose of comparison.
  - Example: Two K-means clusterings can be compared using either WCSS or entropy.

Unsupervised Validation Criteria

The following are examples of unsupervised validation criteria for partitional and hierarchical clustering algortithms.

- Partitional Clustering Algorithms
  - Variability and Separation-Based
  - Silhouette Coefficient
- Hierarchical Clustering Algorithms
  - Cophenetic Correlation

Variability and Separation Criteria

- Quantifies the inter-cluster (separation) and intra-cluster (variability) dissimilarity.
- Separation/variability criteria can be then used to define an overall validation criterion for a clustering structure $\mathcal{C}$.
- We have previously seen one measure of intra-cluster dissimilarilty: Inertia.
- Recall: Inertia is the sum of the (squared) Euclidean distance of each example in the cluster to its centroid.

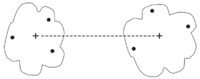
- More generally, we can define the following measures of **centroid-based variability and separation** criteria:
  - Centroid-based variability of cluster C:

$$variability_c(C) = \sum_{e \in C} d(e, centroid(C))$$

    where $d(\cdot,\cdot)$ is **any** distance measure. In particular, if $d(\cdot,\cdot)$ is the squared Euclidean distance, then this measure coincides with inertia.

  - Centroid-based inter-cluster separation between clusters C_1 and C_2:

$$separation_c(C_1, C_2) = d(centroid(C_1), centroid(C_2))$$



  - Centroid-based separation of cluster C_1 with respect to the whole data:

$$separation_c(C_1) = d(centroid(C_1), centroid(data)).$$

Validation criteria for Clustering Structure

- Consider $d(\cdot, \cdot)$ to be **squared Euclidean distance**.
- Variability and separation criteria under the squared Euclidean distance can be used to define the following two overall validity criteria for a clustering structure ($\mathcal{C}$):
  - Withing Cluster Sum of Squares (WCSS):

$$WCSS(\mathcal{C}) = \sum_{C \in \mathcal{C}} inertia(C)$$

  - Between Cluster Sum of Squares (BCSS):

$$BCSS(\mathcal{C}) = \sum_{C \in \mathcal{C}} |C| separation_c(C),$$

  where $|C|$ denotes the number of examples in cluster C.
- Importantly, for a clustering structure $\mathcal{C}$, the following relation holds:

$$WCSS(\mathcal{C}) + BCSS(\mathcal{C}) = constant,$$

  whereby minimizing WCSS ensures maximizing BCSS.
- Moreover, the validation criteria WCSS can be used to estimate the number of clusters via elbow method.

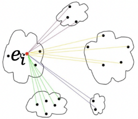
Silhouette Coefficient (SC)

- SC can be evaluated for an individual example, for a cluster, as well as for a clustering structure of K clusters.
- SC combines the ideas of variability and separation.
- **Computing SC for an example**: Let example e_i belongs to cluster C.
  - Calculate a_i = average distance of ith example to all other examples in its cluster, i.e.,

$$a_i = \frac{\sum_{e \in C, e \neq e_i} d(e_i, e)}{|C| - 1}.$$

  - Calculate b_i = minimum (over clusters) of the average distances of ith example to examples in another cluster, i.e.,

$$b_i = \min_{k=1,\ldots,K, C_k \neq C} \frac{\sum_{e \in C_k} d(e_i, e)}{|C_k|}$$



  - SC for ith example is

$$s_i = \frac{b_i - a_i}{\max\{b_i, a_i\}}$$

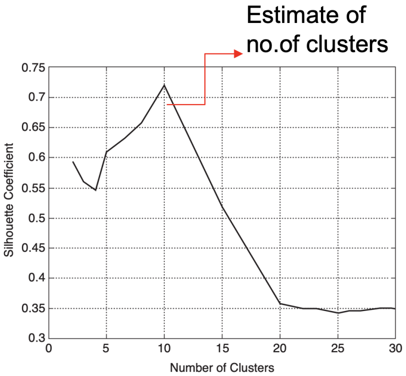
Properties of SC

- SC can vary between -1 and 1.
  - $SC = -1$: $(a_i > b_i = 0) \implies$ data is better fit to a neighboring cluster
  - $SC = 0$: $(a_i = b_i) \implies$ data is on the border between two clusters
  - $SC = 1$: $(0 = a_i < b_i) \implies$ data is well-matched to the cluster
- SC of a cluster = average of SCs of examples in the cluster
- SC of a clustering = average of SC of all examples in the dataset.

SC to Estimate the Number of Clusters

Average SC of a clustering structure can be used to estimate the optimal number of clusters in the data set.

- Plot the average SC of clustering as a function of number of clusters
- Peak in the plot gives an estimate of the number of clusters.



Supervised Validity Criteria for Partitional Clustering Algorithms

- Supervised validation criteria make use of access to to external information in the form of externally derived class labels for data objects.
- For partitional clustering algorithms, there are two classes of supervised validation criteria:
  - **Classification-oriented**
    - Uses measures from classification
    - Quantifies the extent to which a cluster contains objects of a single class
    - Examples include: Entropy, Purity, Precision, Recall, F-measure
  - **Similarity-oriented**
    - Related to similarity measures for binary data
    - Quantifies the extent to which two objects in the same class are in the same cluster and vice versa
    - Examples include: Jaccard measure, Rand statistic

Classification-Oriented Validity Measures

- Uses externally derived class labels for data examples.
- Example: Confusion matrix output of classes to clusters evaluation in WEKA on LA Times dataset. (Each entry corresponds to number of objects in a cluster that belongs to the corresponding class)

Predicted Cluster labels

True Class labels (externally supplied)

	Cluster 1	Cluster 2	Cluster 3	Total
Entertainment	10	11	50	71
Finance	15	60	13	88
Foreign	20	21	9	50
Metro	3	15	2	20
National	45	2	11	58
Sports	12	28	56	96
Total	105	137	141	383

Confusion Matrix for the output of Clustering Algorithm on LA Times Dataset

- Let L denote the number of classes and K denote the number of clusters.
- Probability that an example of cluster i belongs to class j is given by

$$p_{i,j} = \frac{\text{number of examples of class j in cluster i}}{|C_i|}$$

Precision, Recall and F-Measure

- Precision of cluster i with respect to class j:

$$precision(i,j) = p_{i,j}$$

  - Measures the extent to which a cluster contains objects of a single class.
- Recall of cluster i with respect to class j:

$$recall(i,j) = \frac{\text{number of objects of class j in cluster i}}{\text{number of objects in class j}}$$

  - Determines the fraction of class j contained in cluster i
- F-measure of cluster i with respect to class j:

$$F(i,j) = \frac{2 * precision(i,j) * recall(i,j)}{precision(i,j) + recall(i,j)}$$

  - Measures the extent to which a cluster contains only objects of a particular class and all objects of that class.
  - Combination of both precision and recall.

Entropy

- Degree to which each cluster consists of examples of a single class
- Entropy of ith cluster:

$$ent(C_i) = -\sum_{j=1}^{L} p_{i,j} \log_2(p_{i,j})$$

- Total entropy of a set of clusters:

$$ent = \sum_{i=1}^{K} \frac{|C_i|}{\text{total number of examples}} ent(C_i)$$

- Low entropy $\implies$ clusters consists mostly of examples of same class.

Purity

- Another measure of the extent to which a cluster consists of examples of a single class.
- Purity of ith cluster:

$$Purity(C_i) = \max_j p_{i,j}$$

- Overall purity of the clustering structure:

$$Purity = \sum_{i=1}^{K} \frac{|C_i|}{\text{total no.of examples}} Purity(C_i)$$

- Ideally, we require high purity (close to 1).

Example

Consider the output of K-means clustering as summarized in the following table. Compute the entropy and purity of Cluster 1.

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

	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$	$j = 6$
Cluster	Enter-tainment	Financial	Foreign	Metro	National	Sports
1	3	5	40	506	96	27
2	4	7	280	29	39	2
3	1	1	1	7	4	671
4	10	162	3	119	73	2
5	331	22	5	70	13	23
6	5	358	12	212	48	13
Total	354	555	341	943	273	738

$$p_{1,1} = \frac{3}{3+5+40+506+96+27} = \frac{3}{677} = 0.0044$$

$$p_{1,2} = \frac{5}{677} = 0.0073, \; p_{1,3} = \frac{40}{677} = 0.0590$$

$$p_{1,4} = \frac{506}{677} = 0.7474, \; p_{1,5} = \frac{96}{677} = 0.1418,$$

$$p_{1,6} = \frac{27}{677} = 0.0398$$

Purity of cluster 1
$= \max_j p_{1,j} = 0.7474$
Entropy of Cluster 1
$= \sum_j p_{1,j} \log_2(p_{1,j}) = 1.2270$

Similarity-Oriented Measures

- Measures the extent to which two examples in the same class belong to the same cluster and vice versa.
- Comparison of two $N \times N$ matrices (N is the number of examples):
  - **Ideal cluster similarity matrix** has 1 in the (i, j)th entry if two examples i and j are in the **same cluster**, and 0 otherwise.
  - **Ideal class similarity matrix** has 1 in the (i, j)th entry if two examples i and j are in the **same class**, and 0 otherwise.
- Two classes of similarity oriented measures:
  - Correlation based: Compute the correlation between the above two matrices
  - Binary similarity-based measures

Binary Similarity Based Measures

To evaluate binary similarity based measures, the following quantities need to be computed first:

$$f_{00} = \text{number of pairs of objects having a different class and a different cluster}$$
$$f_{01} = \text{number of pairs of objects having a different class and the same cluster}$$
$$f_{10} = \text{number of pairs of objects having the same class and a different cluster}$$
$$f_{11} = \text{number of pairs of objects having the same class and the same cluster}$$

	Same cluster	Different cluster
Same Class	f_{11}	f_{10}
Different Class	f_{01}	f_{00}

Rand Statistic

$$\text{Rand statistic} = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

- Rand statistic takes values in $[0, 1]$. Higher the better.
- Gives the ratio of object pairs that belong to same class-same cluster or different class-different cluster, among the set of all objects.

Jaccard Coefficient

$$\text{Jaccard coefficient} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

- Jaccard Coefficient takes values in $[0, 1]$. Higher the better.
- Gives the ratio of object pairs that belong to same cluster and same class, among all the object pairs that belong to at least same cluster/class.
- Ignores the set of object pairs not belonging to same class and to same cluster.

Example

Compute the Rand statistic and Jaccard coefficient based on the ideal cluster and ideal class similarity matrices given below.

Table 8.10. Ideal cluster similarity matrix.

Point	p1	p2	p3	p4	p5
p1	1	1	1	0	0
p2	1	1	1	0	0
p3	1	1	1	0	0
p4	0	0	0	1	1
p5	0	0	0	1	1

Table 8.11. Ideal class similarity matrix.

Point	p1	p2	p3	p4	p5
p1	1	1	0	0	0
p2	1	1	0	0	0
p3	0	0	1	1	1
p4	0	0	1	1	1
p5	0	0	1	1	1

Solution:

$$f_{00} = 4, f_{01} = 2, f_{10} = 2, f_{11} = 2$$
$$\text{Rand Statistic} = (2+4)/(4+2+2+2) = 0.6$$
$$\text{Jaccard Coefficient} = 2/(2+2+2) = 0.33$$

Final Remarks

- Measures of clustering tendency:
  - Evaluate whether a data set has clusters without clustering
  - Example: Hopkins Statistic
- There is more to cluster evaluation and is an active area of research.
  - Assessing the significance of cluster validity measures: The validity criteria discussed in this lecture give a single number as a measure of goodness of cluster. How to interpret the significance of this number?
  - Naïve solution – define the range of cluster validity criteria and use statistics to evaluate whether the value we have obtained is unusually low or high.

References

Introduction to Data Mining by Tan, Steinbach and Kumar - Chapter 8