Clustering

(Introduction, Evaluation Metrics)

Dr. JASMEET SINGH

ASSISTANT PROFESSOR

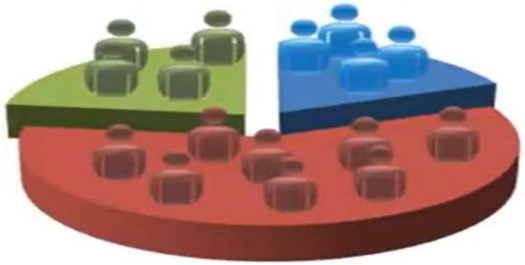
CSED, TIET

Clustering-Introduction

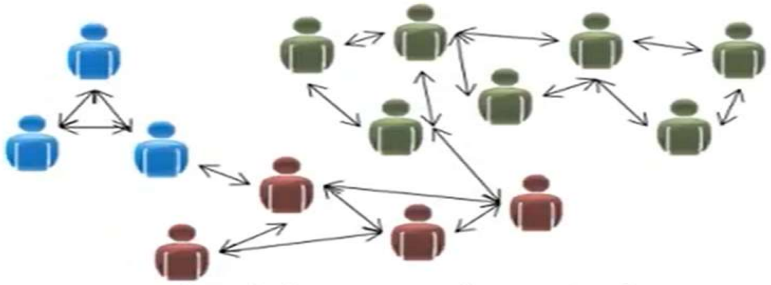
- **Cluster** analysis, or **clustering**, is an unsupervised **machine learning** task. It involves automatically discovering natural grouping in data.
- **Clustering** is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar (high intra-cluster similarity) to other data points in the same group than those in other groups (low inter-cluster similarity).
- In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

Applications of Clustering

Applications of clustering



Market segmentation



Social network analysis

Organize computing clusters

Astronomical data analysis

Applications of Clustering (Contd….)

- Clustering algorithms are widely used in a number of applications such as:
  - Market Segmentation / Targeted Marketing / Recommender Systems
  - Document / News / Article Clustering
  - Biology / Genome Clustering
  - City Planning
  - Speech Recognition
  - Social Network Analysis
  - Organize Computing Clusters
  - Astronomical Data Analysis

Evaluation Metrics

In order to evaluate the quality of clusters produced by a clustering algorithms, following evaluation metrics are used:

1. Silhouette Coefficient
2. Dunn's Index
3. Rand Index (RI)
4. Adjusted Rand Index (ARI)
5. Purity

Metrics 1 and 2 are used when we don't have any ground truth (unsupervised; only data points) where as metrics 3,4 and 5 are used when we have ground truth (supervised; data points and labels)

Silhouette Coefficient

- The Silhouette Coefficient is defined for each sample and is composed of two scores:
  a: The mean distance between a sample and all other points in the same cluster.
  b: The mean distance between a sample and all other points in the next nearest cluster.

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

- The Silhouette Coefficient for a set of samples is given as the mean of the Silhouette Coefficient for each sample.
- The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering.
- Scores around zero indicate overlapping clusters.
- The score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster.

Dunn Index

- The **Dunn index** is another internal clustering validation measure which can be computed as follow:

1. For each cluster, compute the distance between each of the objects in the cluster and the objects in the other clusters
2. Use the minimum of this pairwise distance as the inter-cluster separation (*min.separation*)
3. For each cluster, compute the distance between the objects in the same cluster.
4. Use the maximal intra-cluster distance (i.e maximum diameter) as the intra-cluster compactness
5. Calculate *Dunn index (D)* is computed as follows:

$$D = \frac{min.sepeartion}{maximum\ diameter}$$

If the data set contains compact and well-separated clusters, the diameter of the clusters is expected to be small and the distance between the clusters is expected to be large. Thus, Dunn index should be maximized. The value of Dunn Index lies between 0 and infinity.

Rand Index

- Rand Index is a measure of how similar clustering results or groupings are to the ground truth.
- Let C denotes the ground truth class labeling and K be the clustering assignment.
  - A be the number of element pairs that lie in the same set of C and K,
  - B be the number of element pairs that lie in different sets of both C and K.

Then RI is given by:

$$RI = \frac{A+B}{{}^{n}_{2}C}$$

where n are the total number of samples.

- RI can never exceed 1 and its possible lowest value is 0. More closer the score is to 1, better is the algorithm.

Rand Index- Example

- Say we have five examples. The clustering method groups examples A, B, and C into one group and examples D and E into another group. But according to ground truth groups A and B *are* together and C, D, and E together.
- To compute RI for this example, lets first list all possible unordered pairs of five examples at hand. We have 10 (n*(n-1)/2) such pairs. These are: $\{A, B\}$, $\{A, C\}$, $\{A, D\}$, $\{A, E\}$, $\{B, C\}$, $\{B, D\}$, $\{B, E\}$, $\{C, D\}$, $\{C, E\}$, and $\{D, E\}$.
- Examining these pairs, we notice that the pair $\{A, B\}$ and $\{D, E\}$ are always grouped together (both by clustering algorithm and ground truth). Thus, the value of A is two.
- We also notice that four pairs, $\{A, D\}$, $\{A, E\}$, $\{B, D\}$, and $\{B, E\}$, never occur together. Thus, the value of b is four.

$$RI = \frac{2 + 4}{10} = 0.6$$

Adjusted Rand Index (ARI)

- RI suffers from one drawback; it yields a high value for pairs of random partitions of a given set of examples.
- To counter this drawback, an adjustment is made to the calculations by taking into consideration grouping by chance.
- In this, we create a contingency table, as below the rows denote clusters made by clustering algorithm and columns denote clusters given by ground truth (For example, if the total clusters returned by ground truth and clustering method is 3, then contingency table is as shown below).

	C1	C2	C3
C1			
C2			
C3			

- Any $(ij)^{th}$ entry is the number of common objects belonging to clustering algorithm cluster Ci and ground truth cluster c_j

Adjusted Rand Index (ARI)- Contd….

n_{ij} = Number of examples common to cluster i and cluster j

a_i = Sum of contingency cells in row i

b_j = Sum of contingency cells in column j

The ARI is then expressed as:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}$$

ARI-Example

Consider the same example as discussed for RI (in slide 9).

The contingency matrix for the example is given by:

	M2C1 {A,B}	M2C2 {C,D,E}	
M1C1 {A,B,C}	2	1	3
M1C2 {D,E}	0	2	2
	2	3	

$$\sum_{ij} \binom{n_{ij}}{2} = \binom{2}{2} + \binom{1}{2} + \binom{0}{2} + \binom{2}{2} = (1 + 0 + 0 + 1) = 2$$

$$\sum_{i} \binom{a_i}{2} = (\binom{3}{2} + \binom{2}{2}) = (3 + 1) = 4$$

$$\sum_{j} \binom{b_j}{2} = (\binom{2}{2} + \binom{3}{2}) = (1 + 3) = 4$$

$$ARI = \frac{2 - \frac{4 \times 4}{10}}{\frac{4+4}{2} - \frac{4 \times 4}{2}} = \frac{2 - 1.6}{4 - 1.6} = 0.1666$$

Purity

- Purity is also an external evaluation criterion of cluster quality.
- It is the percent of the total number of objects(data points) that were classified correctly.
- It also lies in the range 0 to 1. Higher the purity, better is the model

$$Purity = \frac{1}{N}\sum_{i=1}^{k} max_j |c_i \cap t_j|$$

where N = number of objects(data points), k = number of clusters, ci is a cluster in Clustering algorithm C, and tj is the cluster in the ground truth

Purity-Example

- For the example discussed for ARI (in slide 12), the contingency table is as below:

	M2C1 {A,B}	M2C2 {C,D,E}	
M1C1 {A,B,C}	2	1	3
M1C2 {D,E}	0	2	2
	2	3	

$$Purity = \frac{\max of\ each\ row\ in\ contigency\ matrix}{Total\ samples} = \frac{2+2}{5} = 0.8$$

Types of Clustering Algorithms

- The clustering algorithms are broadly classified into five categories:

1. Partitioning-based Methods
2. Hierarchical-based Methods
3. Density-based Methods
4. Grid-based Methods
5. Model-based Methods

Partitioning-based Methods

- These methods partition the objects into k clusters and each partition forms one cluster.
- This method is used to optimize an objective criterion similarity function.
- The quality of clustering is measured by an objective function. This objective function is designed to achieve high intra-cluster similarity and low inter-cluster similarity.
- Example *K-means, CLARANS (Clustering Large Applications based upon Randomized Search)* etc.

Hierarchical-based Methods

- These methods perform a hierarchical breakdown of a given dataset which can be classified as agglomerative and divisive.
- In agglomerative methods, initially, each object is regarded as a cluster on its own and they are then successively merged till they satisfy a termination condition.
- By contrast, in the divisive approach, initially, the set of objects is considered as a single large cluster and is successively split up into smaller clusters until a termination condition is satisfied.
- The former is also called the bottom-up approach whereas the latter is called the top-down approach.

Density-based Methods

- Density-based methods discover clusters based on density.
- These methods can find clusters of arbitrary shapes.
- Here, a cluster is kept growing as long as the number of data objects in the neighborhood exceeds some threshold value.
- These methods have good accuracy and ability to merge two clusters.
- Examples *DBSCAN (Density-Based Spatial Clustering of Applications with Noise)* , *OPTICS (Ordering Points to Identify Clustering Structure)* etc.

Grid-based Methods

- Here, each dimension is divided into several cells thus forming a grid structure between dimensions.
- Clustering operations are then performed on this quantized space.
- The processing time of these methods is independent of the number of objects. Rather, it is determined by the number of cells in the grid structure.
- STING, Wavecluster , CLIQUE and OptiGrid are well-known examples of this category.

Model-Based

- These methods perform clustering by first hypothesizing a mathematical model and then finding its best ft for a given dataset.
- For example, the EM algorithm performs an expectation-maximization analysis
- COBWEB performs a probability analysis and a neural network based method, Self-Organizing Maps (SOMs) , performs clustering by mapping high dimensional data onto a 2-D or 3-D feature map.

Clustering Algorithms