

## Chapter 10.

# Cluster Analysis: Evaluation

Meng Jiang

CSE 40647/60647 Data Science Fall 2017

Introduction to Data Mining

# Cluster Analysis

- Cluster Analysis: An Introduction
- Partitioning Methods
- Density-based Methods
- **Evaluation of Clustering**

# Clustering Validation

- Clustering Validation: Basic Concepts
- Clustering Evaluation: Measuring Clustering Quality
- External Measures for Clustering Validation
  - I: Matching-Based Measures
  - II: Pairwise Measures
  - III: Entropy-Based Measures
- Internal Measures for Clustering Validation: BetaCV
- Relative Measures for Clustering Validation

# Clustering Validation and Assessment

- Major issues on clustering validation and assessment
  - Clustering evaluation
    - Evaluating the goodness of the clustering
  - Clustering stability
    - To understand the sensitivity of the clustering result to various algorithm parameters, e.g., # of clusters
  - Clustering tendency
    - Assess the suitability of clustering, i.e., whether the data has any inherent grouping structure

# Measuring Clustering Quality

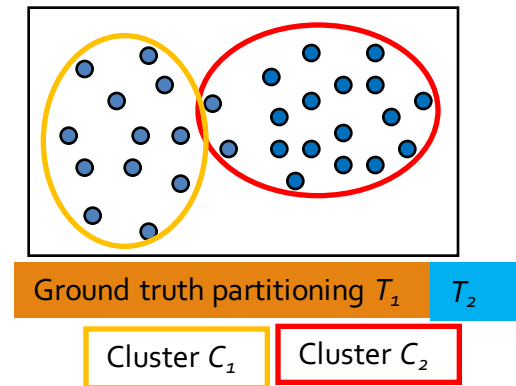
- **Clustering Evaluation:** Evaluating the goodness of clustering results
  - No commonly recognized best suitable measure in practice
- **Three categorization of measures:** External, internal, and relative
  - **External:** Supervised, employ criteria not inherent to the dataset
    - Compare a clustering against prior or expert-specified knowledge (i.e., the ground truth) using certain clustering quality measure
  - **Internal:** Unsupervised, criteria derived from data itself
    - Evaluate the goodness of a clustering by considering how well the clusters are separated and how compact the clusters are, e.g., silhouette coefficient
  - **Relative:** Directly compare different clusterings, usually those obtained via different parameter settings for the same algorithm

# Measuring Clustering Quality: External Methods

- Given the **ground truth**  $T$ ,  $Q(C, T)$  is the **quality measure** for a clustering  $C$
- $Q(C, T)$  is good if it satisfies the following **four** essential criteria
  - **Cluster homogeneity**
    - The purer, the better
  - **Cluster completeness**
    - Assign objects belonging to the same category in the ground truth to the same cluster
  - **Rag bag better than alien**
    - Putting a heterogeneous object into a pure cluster should be penalized more than putting it into a *rag bag* (i.e., “miscellaneous” or “other” category)
  - **Small cluster preservation**
    - Splitting a small category into pieces is more harmful than splitting a large category into pieces

# Commonly Used External Measures

- **Matching-based measures**
  - Purity, maximum matching, F-measure
- **Pairwise measures**
  - Four possibilities: True positive (TP), FN, FP, TN
  - Jaccard coefficient, Rand statistic, Fowlkes-Mallow measure
- **Entropy-Based Measures**
  - Conditional entropy
  - Normalized mutual information (NMI)
  - Variation of information



# Matching-Based Measures (I): Purity vs. Maximum Matching

- **Purity:** Quantifies the extent that cluster  $C_i$  contains points only from one (ground truth) partition:

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

– Total purity of clustering  $C$ : 
$$purity = \sum_{i=1}^r \frac{n_i}{n} purity_i = \frac{1}{n} \sum_{i=1}^r \max_{j=1}^k \{n_{ij}\}$$

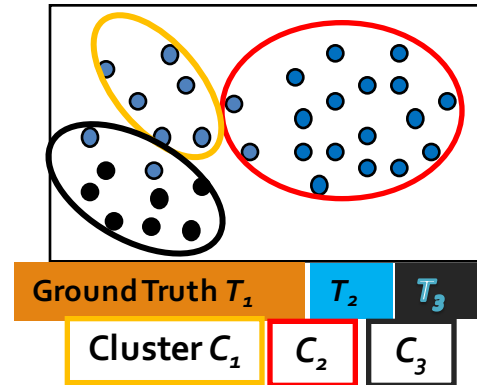
- Perfect clustering if purity = 1 and  $r = k$  (the number of clusters obtained is the same as that in the ground truth)

- Ex. 1 (green or orange):  $purity_1 = 30/50$ ;  $purity_2 = 20/25$ ;  $purity_3 = 25/25$ ;  $purity = (30 + 20 + 25)/100 = 0.75$

- Two clusters may share the same majority partition

- **Maximum matching:** Only one cluster can match one partition

- Maximum weight matching: Paire-wise
- Ex2. (green)  $match = purity = 0.75$ ; (orange)  $match = 0.65 > 0.6$



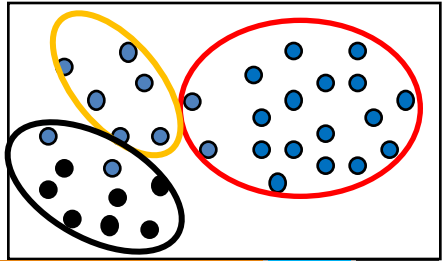
$C \setminus T$	$T_1$	$T_2$	$T_3$	Sum
$C_1$	0	20	30	50
$C_2$	0	20	5	25
$C_3$	25	0	0	25
$m_j$	25	40	35	100

$C \setminus T$	$T_1$	$T_2$	$T_3$	Sum
$C_1$	0	30	20	50
$C_2$	0	20	5	25
$C_3$	25	0	0	25
$m_j$	25	50	25	100



# Matching-Based Measures (II): F-Measure

- **Precision:** The fraction of points in  $C_i$  from the majority partition  $T_{j_i}$  (i.e., the same as purity), where  $j_i$  is the partition that contains the maximum # of points from  $C_i$ 
  - Ex. For the green table
    - $prec_1 = 30/50$ ;  $prec_2 = 20/25$ ;  $prec_3 = 25/25$
- **Recall:** The fraction of point in partition shared in common with cluster  $C_i$ , where  $m_{j_i} = |T_{j_i}|$ 
  - Ex. For the green table
    - $recall_1 = 30/35$ ;  $recall_2 = 20/40$ ;  $recall_3 = 25/25$
- **F-measure** for  $C_i$ : The harmonic means of  $prec_i$  and  $recall_i$ :  $F_i = \frac{2n_{ij_i}}{n_i + m_{j_i}}$
- F-measure for clustering C: average of all clusters:
  - Ex. For the green table
    - $F_1 = 60/85$ ;  $F_2 = 40/65$ ;  $F_3 = 1$ ;  $F = 0.774$



The diagram shows three clusters of points: a black cluster (bottom left), a blue cluster (top left), and a red cluster (top right). Below the diagram is a confusion matrix with columns for Ground Truth partitions  $T_1$  (orange),  $T_2$  (blue), and  $T_3$  (black), and rows for Cluster partitions  $C_1$  (orange),  $C_2$  (red), and  $C_3$  (black). The matrix is a 4x4 table with the following values:

$C \setminus T$	$T_1$	$T_2$	$T_3$	Sum
$C_1$	0	20	30	50
$C_2$	0	20	5	25
$C_3$	25	0	0	25
$m_j$	25	40	35	100

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

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

# Pairwise Measures: Four Possibilities for Truth Assignment

- **Four possibilities** based on the agreement between cluster label and partition label
  - *TP*: true positive—Two points  $\mathbf{x}_i$  and  $\mathbf{x}_j$  belong to the same partition  $T$ , and they also in the same cluster  $C$

$$TP = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j \text{ and } \hat{y}_i = \hat{y}_j\}|$$

where  $y_i$ : the true partition label, and  $\hat{y}_i$ : the cluster label for point  $\mathbf{x}_i$

- *FN*: false negative:  $FN = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}|$
- *FP*: false positive  $FP = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j \text{ and } \hat{y}_i = \hat{y}_j\}|$
- *TN*: true negative  $TN = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}|$

- Calculate the four measures:

$$TP = \sum_{i=1}^r \sum_{j=1}^k \binom{n_{ij}}{2} = \frac{1}{2} \left( \sum_{i=1}^r \sum_{j=1}^k n_{ij}^2 \right) - n \quad FN = \sum_{j=1}^k \binom{m_j}{2} - TP \quad N = \binom{n}{2} \quad \text{Total \# of pairs of points}$$

$$FP = \sum_{i=1}^r \binom{n_i}{2} - TP \quad TN = N - (TP + FN + FP) = \frac{1}{2} \left( n^2 - \sum_{i=1}^r n_i^2 - \sum_{j=1}^k m_j^2 + \sum_{i=1}^r \sum_{j=1}^k n_{ij}^2 \right)$$

# Pairwise Measures: Jaccard Coefficient and Rand Statistic

- Jaccard coefficient: Fraction of true positive point pairs, but after ignoring the true negatives (thus asymmetric)
  - Jaccard =  $TP / (TP + FN + FP)$  [i.e., denominator ignores TN]
  - Perfect clustering: Jaccard = 1
- Rand Statistic:
  - Rand =  $(TP + TN) / N_{total}$
  - Symmetric; perfect clustering: Rand = 1
- Fowlkes-Mallow Measure:
  - Geometric mean of precision and recall
$$FM = \sqrt{prec \times recall} = \frac{TP}{\sqrt{(TP + FN)(TP + FP)}}$$
- Using the above formulas, one can calculate all the measures for the green table (leave as an exercise)

$C \setminus T$	$T_1$	$T_2$	$T_3$	Sum
$C_1$	0	20	30	50
$C_2$	0	20	5	25
$C_3$	25	0	0	25
$m_j$	25	40	35	100

# Entropy-Based Measures (I): Conditional Entropy

- **Entropy of clustering  $\mathcal{C}$ :**  $H(\mathcal{C}) = - \sum_{i=1}^r p_{C_i} \log p_{C_i}$   $p_{C_i} = \frac{n_i}{n}$  (i.e., the probability of cluster  $C_i$ )
- **Entropy of partitioning  $\mathcal{T}$ :**  $H(\mathcal{T}) = - \sum_{j=1}^k p_{T_j} \log p_{T_j}$
- **Entropy of  $\mathcal{T}$  with respect to cluster  $C_i$ :**  $H(\mathcal{T}|C_i) = - \sum_{j=1}^k \left(\frac{n_{ij}}{n_i}\right) \log\left(\frac{n_{ij}}{n_i}\right)$
- **Conditional entropy of  $\mathcal{T}$  with respect to clustering  $\mathcal{C}$ :**

$$H(\mathcal{T}|\mathcal{C}) = - \sum_{i=1}^r \left(\frac{n_i}{n}\right) H(\mathcal{T}|C_i) = - \sum_{i=1}^r \sum_{j=1}^k p_{ij} \log\left(\frac{p_{ij}}{p_{C_i}}\right)$$

- The more a cluster's members are split into different partitions, the higher the conditional entropy
- For a perfect clustering, the conditional entropy value is 0, where the worst possible conditional entropy value is  $\log k$

$$\begin{aligned} H(\mathcal{T}|\mathcal{C}) &= - \sum_{i=1}^r \sum_{j=1}^k p_{ij} (\log p_{ij} - \log p_{C_i}) = - \sum_{i=1}^r \sum_{j=1}^k p_{ij} \log p_{ij} + \sum_{i=1}^r (\log p_{C_i} \sum_{j=1}^k p_{ij}) \\ &= - \sum_{i=1}^r \sum_{j=1}^k p_{ij} \log p_{ij} + \sum_{i=1}^r (p_{C_i} \log p_{C_i}) = H(\mathcal{C}, \mathcal{T}) - H(\mathcal{C}) \end{aligned}$$

# Entropy-Based Measures (II): Normalized Mutual Information (NMI)

- **Mutual information:**

- Quantifies the amount of shared info between  $I(C, T) = \sum_{i=1}^r \sum_{j=1}^k p_{ij} \log\left(\frac{p_{ij}}{p_{C_i} \cdot p_{T_j}}\right)$  the clustering  $C$  and partitioning  $T$
- Measures the dependency between the observed joint probability  $p_{ij}$  of  $C$  and  $T$ , and the expected joint probability  $p_{C_i} \cdot p_{T_j}$  under the independence assumption
- When  $C$  and  $T$  are independent,  $p_{ij} = p_{C_i} \cdot p_{T_j}$ ,  $I(C, T) = 0$ . However, there is no upper bound on the mutual information

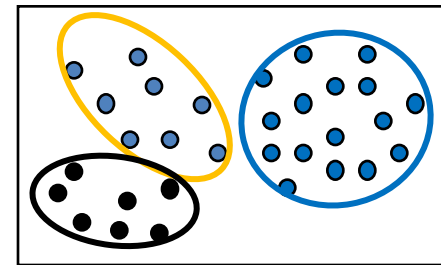
- **Normalized mutual information (NMI)**

$$NMI(C, T) = \sqrt{\frac{I(C, T)}{H(C)} \cdot \frac{I(C, T)}{H(T)}} = \frac{I(C, T)}{\sqrt{H(C) \cdot H(T)}}$$

- Value range of NMI:  $[0, 1]$ . Value close to 1 indicates a good clustering

# Internal Measures: BetaCV Measure

- A trade-off in maximizing intra-cluster compactness and inter-cluster separation
- Given a clustering  $C = \{C_1, \dots, C_k\}$  with  $k$  clusters, cluster  $C_i$  containing  $n_i = |C_i|$  points
  - Let  $W(S, R)$  be sum of weights on all edges with one vertex in  $S$  and the other in  $R$
  - The sum of all the intra-cluster weights over all clusters:  $W_{in} = \frac{1}{2} \sum_{i=1}^k W(C_i, C_i)$
  - The sum of all the inter-cluster weights:  $W_{out} = \frac{1}{2} \sum_{i=1}^k W(C_i, \overline{C_i}) = \sum_{i=1}^{k-1} \sum_{j>i}^k W(C_i, C_j)$
  - The number of distinct intra-cluster edges:  $N_{in} = \sum_{i=1}^k \binom{n_i}{2}$
  - The number of distinct inter-cluster edges:  $N_{out} = \sum_{i=1}^{k-1} \sum_{j=i+1}^k n_i n_j$
- **Beta-CV measure:**
  - The ratio of the mean intra-cluster distance to the mean inter-cluster distance
  - The smaller, the better the clustering



$$BetaCV = \frac{W_{in} / N_{in}}{W_{out} / N_{out}}$$

# Summary

- Cluster Analysis: An Introduction
- Partitioning Methods
- Density-based Methods
- Evaluation of Clustering

# References: (IV) Evaluation of Clustering

- M. J. Zaki and W. Meira, Jr.. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, 2014
- L. Hubert and P. Arabie. Comparing Partitions. Journal of Classification, 2:193–218, 1985
- A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Printice Hall, 1988
- M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On Clustering Validation Techniques. Journal of Intelligent Info. Systems, 17(2-3):107–145, 2001
- J. Han, M. Kamber, and J. Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011
- H. Xiong and Z. Li. Clustering Validation Measures. in (Chapter 23) C. Aggarwal and C. K. Reddy (eds.), Data Clustering: Algorithms and Applications. CRC Press, 2014