ECE 657A: Data and Knowledge Modelling and Analysis

Evaluation Measures for Clustering

Mark Crowley

Evaluation Measures for Clustering

Internal

- Density of clusters
- Inherent properties of the data, classes

External

- Evaluated based on correct clusters

External Evaluation Measures for Clustering

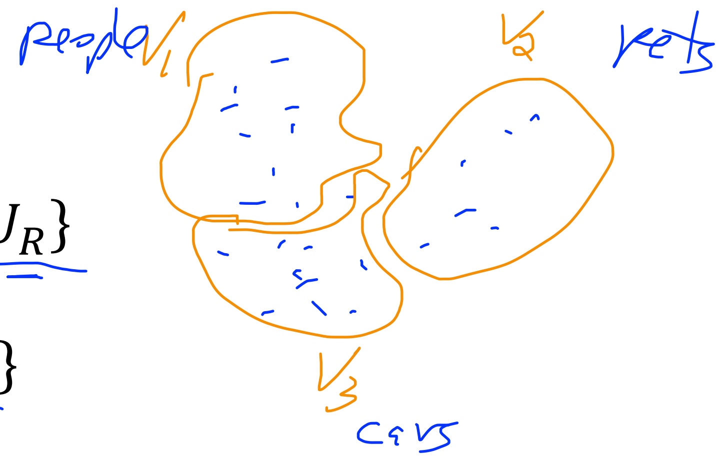
- If we know the class labels of the true clusters, accuracy and F-measure can be calculated by constructing the contingency table using the known clustering to compare to resulting clustering

Suppose we know the classes are

$$U = \{U_1, U_2, \ldots, U_R\}$$

and the resulting clustering is

$$V = \{V_1, \ldots, V_k\}$$



Notation for Clustering Success

Indicator Variable

Characteristic Functions

$$I_u(i,j) = \begin{cases} \underline{1} \text{ if } x_i \in U_r \text{ and } x_j \in U_r, 1 \le r \le R \\ \underline{0} \text{ otherwise} \end{cases}$$

R classes
"true"

$$I_v(i,j) = \begin{cases} 1 \text{ if } \mathrm{x_i} \in V_\mathrm{s} \text{ and } x_j \in V_\mathrm{s} \ 1 \le \mathrm{s} \le K \\ 0 \end{cases}$$

K clusters
"assigned"

Evaluation Measures

TP : True Positive
FN : False Negative

	same cluster	different cluster
same class	**a (TP)**	**b (FN)**
different class	**c (FP)**	**d (TN)**

Count the number of points that satisfy each pairing

- a+d : number of agreements between U and V
- b+c : number of disagreements between U and V

Contingency Table

		I_v		
		1	0	
I_U	1	**a**	**b**	m_1
	0	**c**	**d**	$M - m_1$
		m_2	$M - m_2$	

total number of pairs of objects

$$M = a + b + c + d = \frac{n(n-1)}{2} \quad \binom{n}{2}$$

Evaluation Measures

class

	T	F
T	a	b
F	c	d

cluster

as $U \leftrightarrow V \rightarrow 1$

Larger Means Closer

$$\text{Rand Index} = \frac{a+d}{\binom{n}{2}}$$

- Probability that U and V agree on a random pair

$$\text{Jaccard} = \frac{a}{a+b+c}$$

- Measures amount of overlap in U and V

$$\text{Fowlkes \& Mallows} = \sqrt{\frac{a}{a+c}\frac{a}{a+b}}$$

- Works well even when U and V very unrelated

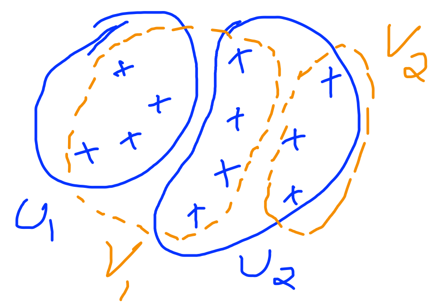
$\rightarrow 0$

Evaluation Measures

- **Rand:** focused on relative accuracy, number of agreements normalized, valid even if labels not available.
- **Jaccard:** measures similarity between the two clusterings, size of intersection divided by size of union.
- **F&M:** corrects data to be normalized by “bunchiness” of data, how easy is it to cluster. Also works well if U and V are very different, approaches zero (Rand approaches 1).

F-measure

- If we can't specify the class of the cluster we can compute the F-measure of the cluster with regard to each class
- **Class** is inherent to data
- **Cluster** is *our* current grouping of the data.



Cluster i Class j

$$\text{precision}(i, j) = m_{ij} / n_i$$

$$\text{recall}(i, j) = m_{ij} / m_j$$

$4/4 = 1$

$m_{ij} = \#$ of objects of Class j in Cluster i

$4/8 = \frac{1}{2}$

$n_i = \#$ of objects in Cluster i

$m_j = \#$ of objects in Class j

F-measure

F-measure of cluster i with respect to class j is

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

$$F = \sum_{j=1}^{\overset{\text{class}}{C}} \frac{m_j}{n} \max_{\underset{\text{cluster}}{i \in k}} F(i,j)$$

For k clusters and C classes

Entropy

The degree to which each cluster consists of objects of a single class

$$P_{ij} = m_{ij} / n_i = \text{probability that a member of cluster i belongs to class j}$$

$$e_i = -\sum_{j=1}^{C} P_{ij} \log_2 P_{ij} = \text{entropy of cluster i}$$

total entropy

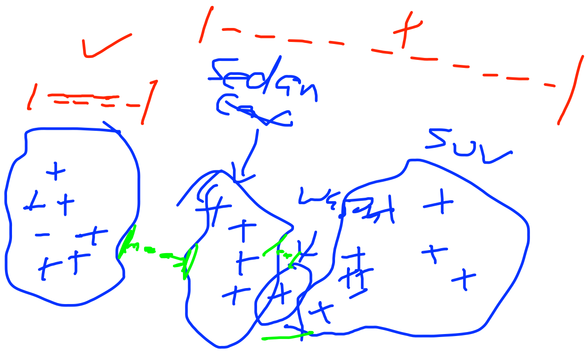
$$e = \sum_{i=1}^{k} \frac{n_i}{n} e_i$$

k = clusters

n = number of points

Internal Measures

Evaluating how well the results of a clustering algorithm perform without reference to external information



Two types of measures:

1. **Cluster Cohesion** (compactness, tightness): how closely related are the patterns in the same cluster?

2. **Cluster Separation** (isolation): how well separated is the cluster from other clusters?

Internal Measures

$$cohesion(C_i) = \sum_{\substack{x \in C_i \\ Y \in C_i}} proximity(x, y)$$

Proximity can be a function of similarity between patterns for protoype-based clusters

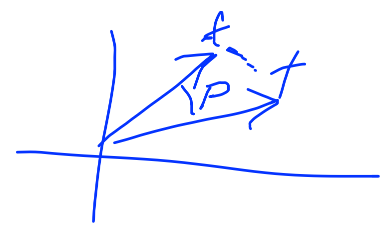
$$cohesion\ (C_i) = \sum_{x \in C_i} proximity(x, m_i)$$

m_i is the prototype (or center) of cluster i

Internal Measures

$$\text{overall cohesion} = \sum_{i=1}^{k} w_i \, \text{cohesion}(C_i)$$

Where k is the number of clusters and w_i is a weight function of the size of the cluster e.g. n_i / n

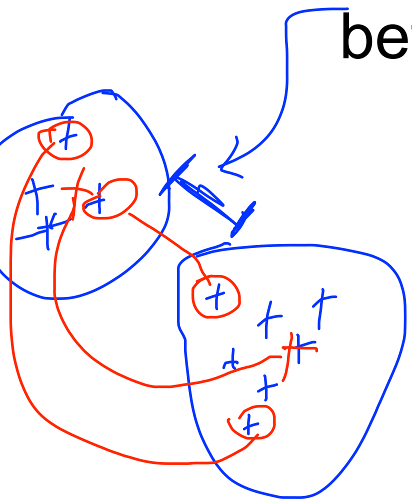


example of proximity is cos. similarity

another example is $\dfrac{1}{\max d}$ or $\dfrac{1}{\sum d}$ or $\dfrac{1}{\sum d^2}$

Internal Measures

Separation between 2 clusters can be the distance between the 2 clusters (can be minimum)



$$d(C_i, C_j) = \min_{\substack{x \in C_i \\ y \in C_j}} d(x, y) \qquad \underline{\text{single - link like}}$$

or max

$$d(C_i, C_j) = \max_{\substack{x \in C_i \\ y \in C_j}} d(x, y) \qquad \underline{\text{complete link like}}$$

overall separation can be the min pairwise separation between the clusters

$$\text{overall separation} = \min_{\substack{i=1 \ldots k \\ j=i+1, \ldots k}} d(C_i, C_j)$$

Ratios & Indices

Separation Index (SI)

Smaller indicates more separate

Within cluster scatter

$$SI = \sum_{i=1}^{k} \sum_{x_j \in C_i} d^2\left(x_j, m_i\right) \Big/ n \min_{C_r, C_s \in C} d\left(C_r, C_s\right)$$

Ratios & Indices

Separation Index (SI)

- d^2(x,y) is the Euclidean distance between points, or some other distance
- d(C_r,C_s) is a measure across sets, minimal distance or max distance, min squared error depending on what is being used.

Dunn - Index

$$D_C = \min_{i=1,\cdots K}\left\{ \min_{j=i+1,\cdots K} \frac{d(C_i, C_j)}{\max\limits_{\ell=1,\cdots k} diam(C_\ell)} \right\}$$

$$d(C_i, C_j) = \min_{\substack{x \in C_i \\ y \in C_j}} d(x, y)$$

$$diam(c) = \max_{x, y \in C} d(x, y)$$

Distance between clusters (separation)

Measures dispersion of the cluster (inverse of cohesion)

Large indicates compact & well separated clusters

Other measures
see Jain and Dubes book Chapter 4
see article Halkidi etal, J of Intelligent Info. Systems, Dec. 2001