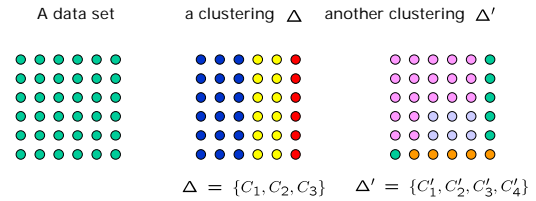


## Comparing clusterings

Marina Meila  
University of Washington  
Department of Statistics

[www.stat.washington.edu/mmp](http://www.stat.washington.edu/mmp)

## The problem



- How different are  $\Delta, \Delta'$  ?

## Why do we compare clusterings?

- How different are  $\Delta, \Delta'$  ?
- Which of  $\Delta, \Delta'$  is closer to a ground truth  $\Delta^*$ ?
- Several algorithms produce clusterings of the same data set. Which algorithm is better (assuming the correct clustering is known)?
- Two algorithms cluster several different data sets. Which algorithm is better (assuming the correct clusterings are known)?
- Given a large set of clusterings of the same data, find a small "diverse" subset of clusterings.

## How should the "distance" be?

- It depends... ☺ (on the application)
- A "reasonable" set of requirements
  - Applies to any two partitions of the same data set
  - Makes no assumptions about how the clusterings are obtained
  - Values of the "distance" for two pairs of clusterings comparable under the weakest possible assumptions
  - Metricity

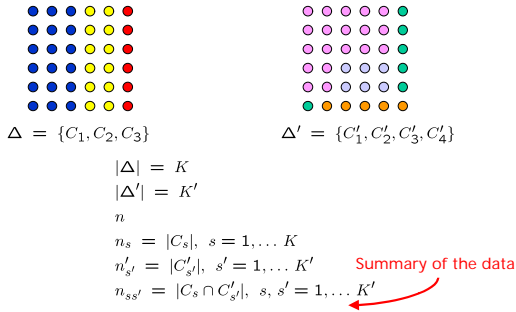
## Outline

- Overview of clustering comparison criteria
- The Variation of Information (VI) distance
- Properties of the VI distance
- Extensions
- Other criteria in perspective

## Criteria for comparing clusterings. An overview

- Type of criterion by "principle"
  - Criteria based on counting pairs (or triples)
    - Rand, Fowlkes-Mallows
  - Criteria based on cluster matching
    - "classification error", Van Dongen, Larsen
  - Other criteria
    - Based on mutual information
- Type of criterion by mathematical properties
  - Metric (distance) = 0 for identical clusterings
  - "index" = 1 for identical clusterings

### Notation



### Criteria based on counting pairs

$N_{11}$  = # pairs in the same cluster in both  $\Delta, \Delta'$   
 $N_{00}$  = # pairs in different clusters in  $\Delta, \Delta'$   
 $N_{01}$  = # pairs in same cluster in  $\Delta'$  different in  $\Delta$   
 $N_{10}$  = # pairs in same cluster in  $\Delta$  different in  $\Delta'$

$$N_{11} + N_{00} + N_{01} + N_{10} = n(n-1)/2$$

- Wallace I, II (asymmetric)

$$W_I(\Delta, \Delta') = \frac{N_{11}}{N_{11} + N_{01}}$$

Similar to precision, recall

$$W_{II}(\Delta, \Delta') = \frac{N_{11}}{N_{11} + N_{10}}$$

- Rand

$$\mathcal{R}(\Delta, \Delta') = \frac{N_{11} + N_{00}}{n(n-1)/2}$$

- Fowlkes-Mallows

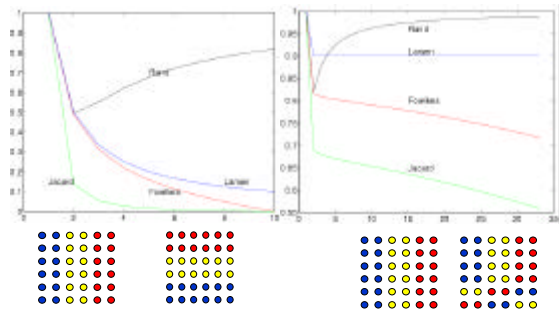
$$\mathcal{F}(\Delta, \Delta') = \frac{N_{11}}{\sqrt{(N_{11} + N_{10})(N_{11} + N_{01})}}$$

- Jacard

$$\mathcal{J}(\Delta, \Delta') = \frac{N_{11}}{N_{11} + N_{10} + N_{01}}$$

### Problem with indices that count pairs

- Unclear how to compare values for different numbers of clusters

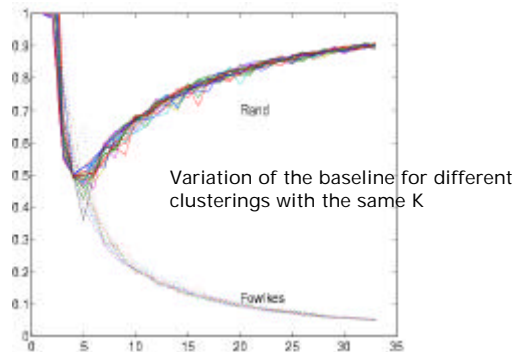


### Adjusted indices

- Idea (Huber + Arabie)

$$\frac{\text{Index} - E[\text{Index}]}{\max[\text{Index}] - E[\text{Index}]}$$

- Implies distributional assumption
  - $\Delta, \Delta'$  independent given marginal counts  $n_1, n_2, \dots, n_K, n'_1, n'_2, \dots, n'_{K'}$
  - Assumption violated in practice
  - $\max[\text{Index}]$  given marginal counts hard to compute
    - In practice used upper bounds



### The Mirkin metric

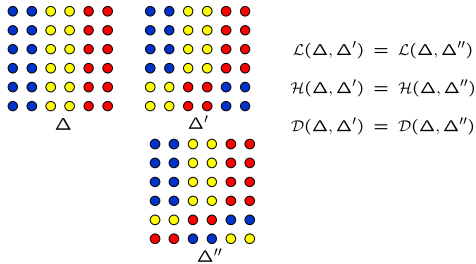
- The Mirkin metric
 
$$\mathcal{K}(\Delta, \Delta') = \sum_{k=1}^K n_k^2 + \sum_{k'=1}^{K'} n_{k'}^2 - \sum_{k=1}^K \sum_{k'=1}^{K'} n_{kk'}^2$$
  - Relationship to Rand index
 
$$\mathcal{K}(\Delta, \Delta') = n(n-1)[1 - \mathcal{R}(\Delta, \Delta')]$$
  - Is a Hamming distance

### Criteria based on cluster matching

- Larsen
  - asymmetric
 
$$\mathcal{L}(\Delta, \Delta') = \frac{1}{K} \sum_{k=1}^K \max_{k'} \frac{2n_{kk'}}{n_k + n_{k'}}$$
- "classification error"
  - (E.g. Meila-Heckerman)
 
$$\mathcal{H}(\Delta, \Delta') = \frac{1}{n} \sum_{k'=match(k)} n_{kk'}$$
- Van Dongen
  - Is a metric
 
$$\mathcal{D}(\Delta, \Delta') = 2n - \sum_k \max_{k'} n_{kk'} - \sum_{k'} \max_k n_{kk'}$$

### Problem with cluster matching criteria

- Cluster matching criteria discard information
  - Acceptable if clusterings very close
  - Bad if clusterings very different

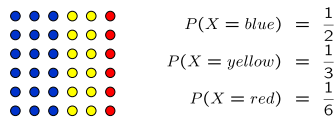


### Outline

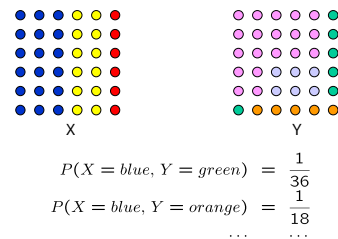
- Overview of clustering comparison criteria
- The Variation of Information (VI) distance
- Properties of the VI distance
- Extensions
- Other criteria in perspective

### The RV associated to a clustering

- Idea: any clustering uniquely determines a random variable
  - Pick a data point at random
  - The RV represents the cluster index of the selected point



### The joint distribution defined by 2 clusterings



- We compare two clusterings by measuring the information loss/gain between the associated RV's

### Entropy and mutual information

- For any RV

$$H(X) = - \sum_{s=1}^K P_X(s) \log P_X(s)$$

$$I(X, Y) = \sum_{s=1}^K \sum_{s'=1}^{K'} P_{XY}(s, s') \log \frac{P_{XY}(s, s')}{P_X(s)P_Y(s')}$$

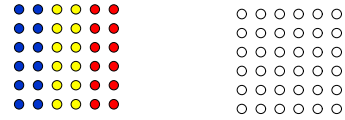
- Entropy = 0 iff RV deterministic
- Mutual information = 0 iff RV's independent

$$0 \leq I(X, Y) \leq H(X), H(Y)$$

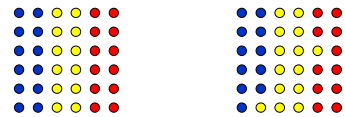
- For two clusterings
  - $H(\Delta), H(\Delta')$  = the entropy of the associated RV's
  - $I(\Delta, \Delta')$  = the mutual information of the associated RV's

### Examples

- Clusterings with 0 mutual information



- Clusterings with high mutual information

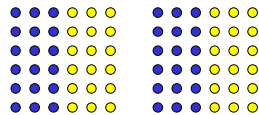


- Mutual information is not enough

### Mutual information is not enough

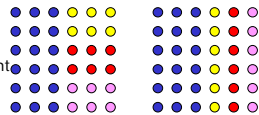
- $I(\Delta, \Delta') = 1$  bit

- Clusterings identical



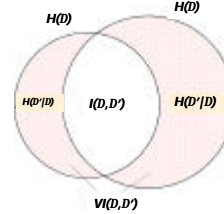
- $I(\Delta, \Delta') = 1$  bit

- Clusterings very different



### The variation of information (VI)

- VI measures the sum of information lost + gained between the two clusterings



$$VI(\Delta, \Delta') = H(\Delta) + H(\Delta') - 2I(\Delta, \Delta')$$

$$VI(\Delta, \Delta') = H(\Delta | \Delta') + H(\Delta' | \Delta)$$

$$VI(\Delta, \Delta') = 2H(\Delta, \Delta') - H(\Delta) - H(\Delta')$$

### Outline

- Overview of clustering comparison criteria
- The Variation of Information (VI) distance
- Properties of the VI distance
- Extensions
- Other criteria in perspective

### Fundamental properties

- Theorem** The VI is a metric

- Non-negative
- symmetric
- triangle inequality

$$VI(\Delta_1, \Delta_3) + VI(\Delta_2, \Delta_3) \geq VI(\Delta_1, \Delta_2)$$

- Advantages of metrics

- More intuitive
- Local neighborhood
- Algorithms operating on sets of clusterings
  - E.g clusters of clusterings

### The local neighborhood

- Splitting a cluster/ Merging two clusters

$$\Delta = \{C_1, C_2, \dots, C_K\}$$

$$\Delta' = \{C'_1, C''_1, C_2, \dots, C_K\}$$

$$VI(\Delta, \Delta') = P(C_1)H(C'_1, C''_1)$$

- Theorem** Nearest neighbor is clustering obtained by splitting/merging
- Radius of smallest nontrivial neighborhood  $\geq \frac{2}{n}$

### Upper bounds

$$VI(\Delta, \Delta') \leq \log n$$

- The bound is attained for  $|\Delta| = 1, |\Delta'| = n$
- If the number of clusters cannot exceed  $K^* \leq \sqrt{n}$ , then
  - $VI(\Delta, \Delta') \leq 2 \log K^*$ 
    - This bound is approached asymptotically in the limit of large  $n$  and is attained exactly for every  $n = m \times (K^*)^2$

### Extensions. Weighted data points

- Data point  $j$  has "weight"  $P(j), \sum_{j=1}^n P(j) = 1$

- Redefine the RV

$$P(s) = \sum_{j \in C_s} P(j)$$

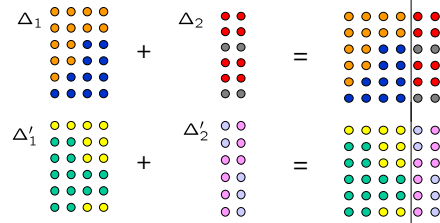
$$P(s, s') = \sum_{j \in C_s \cap C'_{s'}} P(j)$$

- VI defined as before
- Properties
  - Upper bound  $\log n$  not tight
  - Lower bounds can be arbitrarily small
  - All other properties are preserved

### Scale invariance and "convex additivity"

- Scale invariance
  - The VI distance does not depend directly on the number of points

- Linearity



$$VI(\Delta, \Delta') = \frac{n_1}{n_1 + n_2} VI(\Delta_1, \Delta'_1) + \frac{n_2}{n_1 + n_2} VI(\Delta_2, \Delta'_2)$$

### Outline

- Overview of clustering comparison criteria
- The Variation of Information (VI) distance
- Properties of the VI distance
- Extensions
- Other criteria in perspective

### Extensions. Soft clusterings

- A data point has a probability of belonging to each cluster

$$Pr[i \in C_s] = \gamma_i(s)$$

- Redefine the associated RV's

$$P(s) = \sum_{i \in D} P(i) \gamma_i(s) \quad P'(s') = \sum_{i \in D} P(i) \gamma_i(s')$$

$$P(s, s') = \sum_{i=1}^n P(i) \gamma_i(s) \gamma_i(s')$$

- Assumes
  - Is this  $acP(k|i, k) = P(k|i)$
- Obeys triangle inequality but it is not a metric
  - $VI(\Delta, \Delta) > 0$

### Outline

- Overview of clustering comparison criteria
- The Variation of Information (VI) distance
- Properties of the VI distance
- Extensions
- Other criteria in perspective

### Scale invariance

- Scale invariant
  - VI, Larsen  $\mathcal{L}$  clustering error  $\mathcal{H}$
- Can be made invariant
  - Mirkin  $\mathcal{K}$  and Rand  $\mathcal{R}$  Van Dongen  $\mathcal{D}$
- Not invariant
  - Fowlkes-Mallows  $\mathcal{F}$ , Jacard  $\mathcal{J}$

### Scale invariance

$$VI(\Delta_1, \Delta_2) = VI(\Delta'_1, \Delta'_2)$$

$$\mathcal{K}_{inv}(\Delta_1, \Delta_2) = \mathcal{K}_{inv}(\Delta'_1, \Delta'_2)$$

$$\mathcal{R}_{inv}(\Delta_1, \Delta_2) = \mathcal{R}_{inv}(\Delta'_1, \Delta'_2)$$

$$\mathcal{D}_{inv}(\Delta_1, \Delta_2) = \mathcal{D}_{inv}(\Delta'_1, \Delta'_2)$$

$$\mathcal{F}(\Delta_1, \Delta_2) \neq \mathcal{F}(\Delta'_1, \Delta'_2)$$

$$\mathcal{J}(\Delta_1, \Delta_2) \neq \mathcal{J}(\Delta'_1, \Delta'_2)$$

### Linearity

$$VI(\Delta, \Delta') = \frac{n_1}{n_1 + n_2} VI(\Delta_1, \Delta'_1) + \frac{n_2}{n_1 + n_2} VI(\Delta_2, \Delta'_2)$$

$$\mathcal{D}(\Delta, \Delta') = \frac{n_1}{n_1 + n_2} \mathcal{D}_{inv}(\Delta_1, \Delta'_1) + \frac{n_2}{n_1 + n_2} \mathcal{D}_{inv}(\Delta_2, \Delta'_2)$$

$$\mathcal{K}_{inv}(\Delta, \Delta') = \frac{n_1}{(n_1 + n_2)^2} \mathcal{K}_{inv}(\Delta_1, \Delta'_1) + \frac{n_2}{(n_1 + n_2)^2} \mathcal{K}_{inv}(\Delta_2, \Delta'_2)$$

$$\mathcal{R}_{inv}(\Delta, \Delta') = \frac{n_1}{(n_1 + n_2)^2} \mathcal{R}_{inv}(\Delta_1, \Delta'_1) + \frac{n_2}{(n_1 + n_2)^2} \mathcal{R}_{inv}(\Delta_2, \Delta'_2)$$

$\mathcal{F}$  NO

$\mathcal{J}$  NO

### Changes in larger clusters result in larger "distances"

- True for all criteria

### Maximally distant clusterings

$$VI(\Delta_1, \Delta_2) = 2 \log K$$

$$\mathcal{K}_{inv}(\Delta_1, \Delta_2) \approx 2 \frac{K-1}{K^2}$$

$$\mathcal{R}_{inv}(\Delta_1, \Delta_2) \approx 1 - 2 \frac{K-1}{K^2}$$

$$\mathcal{D}_{inv}(\Delta_1, \Delta_2) = 2 \left(1 - \frac{1}{K}\right)$$

$$\mathcal{F}(\Delta_1, \Delta_2) = \frac{n - K^2}{K(n - K)}$$

$$\mathcal{J}(\Delta_1, \Delta_2) \approx \frac{1}{2K^2 - 1}$$

### Axioms (preliminary)

■ **Theorem.** VI is the unique criterion  $d$  that satisfies the following axioms

- A1.  $d$  is a metric
- A2. If  $\Delta_1 \subseteq \Delta_2 \subseteq \Delta_3$  then the triangle inequality is satisfied with equality

$$d(\Delta_1, \Delta_3) = d(\Delta_1, \Delta_2) + d(\Delta_2, \Delta_3)$$

- A3. Convex additivity

$$d(\Delta_1 \cup \Delta_2, \Delta'_1 \cup \Delta'_2) = \frac{n_1}{n_1 + n_2} d(\Delta_1, \Delta'_1) + \frac{n_2}{n_1 + n_2} d(\Delta_2, \Delta'_2)$$

- A4. The distance between 1 and a clustering with  $K$  equal clusters is  $\log K$

### Summary

- No optimal distance between clusterings. The context matters
- Values of "distance" should be
  - comparable under weak assumptions
  - Metric, scale invariant if possible
- The VI distance
  - Inspired by information theory
  - Based on relationship between a point and its cluster rather than on pairwise relationships
  - Matches intuition in some cases
  - Useful mathematical properties: is a METRIC!
    - Generates local neighborhoods in the lattice of partitions
- Future work
  - Exploit the local neighborhoods

