# ProBin Hackathon: Conventional k-means vs Bayesian Nonparametric k-means& Clustering Validation Criteria

Umer Zeeshan Ijaz

# Introduction(1)

- Three main types of clustering techniques
  - Overlapping
  - Partitional
  - Hierarchical
- **Hierarchical clustering** is nested sequence of hard **Partional clustering**, each of which is a partition of data set into a different number of mutually disjoint subsets

# Introduction(2)

Dataset: $\mathbf{X} = \{\mathbf{x}(1), ..., \mathbf{x}(N)\}$

where $\mathbf{x}(j)$ is *n*-dimensional feature or attribute vectors

Collection: $\mathbf{S} = \{\mathbf{S}_1, ..., \mathbf{S}_k\}$

of *k* non-overlapping data subsets $\mathbf{S}_i$ (clusters)

such that

$\mathbf{S}_1 \cup \mathbf{S}_2 \cup ... \cup \mathbf{S}_k = \mathbf{X}$

$\mathbf{S}_i \neq \varnothing$, and $\mathbf{S}_i \cap \mathbf{S}_l = \varnothing$ for $i \neq l$

In **Overlap clustering**, you relax the criteria $\mathbf{S}_i \cap \mathbf{S}_l = \varnothing$

# Problem statement

- Estimation of number of clusters contained in data
- Most algorithms require that the number of clusters be defined *a priori* or *a posteriori* by user, e.g., **k-means**, **EM** (expectation maximization), and **hierarchical clustering algorithms**
- Conventional solution is to get data partition with different number of clusters and choose the best result according to specific criteria (may be **AIC**, **BIC**, etc.)
- Would it then be possible to have a **quantitative** criteria for evaluating the quality of clustering?

# k-means (via Bayesian Nonparametrics) (1)

- In **Gaussian mixture model**, data arises from the distribution

$$p(\mathbf{x}) = \sum_{c=1}^{k} \pi_c N(\mathbf{x} \mid \mu_c, \Sigma_c)$$

where *k* is fixed number of components, $\pi_c$ are the mixing coefficients, and $\mu_c$ and $\Sigma_c$ are the means and covariances, respectively, of the *k* Gaussian distributions.

- In the non-Bayesian setting, we use the EM algorithm to perform maximum likelihood given a set of observations $x_1, \ldots, x_n$

B. Kulis, M. I. Jordan, **Revisiting k-means: New Algorithms via Bayesian Nonparametrics**, arXiv:1111.0352v2. http://arxiv.org/pdf/1111.0352.pdf

# k-means (via Bayesian Nonparametrics) (2)

- E-step
  - Compute the following quantities for all $i = 1,\ldots,n$ and for all $c = 1,\ldots,k$

$$\gamma(z_{ic}) = \frac{\pi_c N(\mathbf{x}_i \mid \mu_c, \Sigma_c)}{\sum_{j=1}^{c} \pi_j N(\mathbf{x}_i \mid \mu_j, \Sigma_j)}$$

- M-step
  - Re-estimate the parameters using the values of $\gamma(z_{ic})$

$$\mu_c^{new} = \frac{1}{n_c} \sum_{i=1}^{n} \gamma(z_{ic}) \mathbf{x}_i$$

$$\Sigma_c^{new} = \frac{1}{n_c} \sum_{i=1}^{n} \gamma(z_{ic})(\mathbf{x}_i - \mu_c^{new})(\mathbf{x}_i - \mu_c^{new})^T$$

$$\pi_c^{new} = \frac{n_c}{n}, \quad n_c = \sum_{i=1}^{n} \gamma(z_{ic})$$

  - EM converges to local optimum of log likelihood function
  - $\gamma(z_{ic})$ are the probabilities of assigning $\mathbf{x}_i$ to cluster c

# k-means (via Bayesian Nonparametrics) (3)

- k-means objective function
  - Given set of data points $x_1, \ldots, x_n$ , the k-means objective function to find clusters $l_1, \ldots, l_k$ to minimize the following objective function:

$$\min_{\{l_c\}_{c=1}^{k}} \sum_{c=1}^{k} \sum_{\mathbf{x} \in l_c} \left\| \mathbf{x} - \mu_c \right\|_2^2$$

$$\text{where } \mu_c = \frac{1}{|l_c|} \sum_{\mathbf{x} \in l_c} \mathbf{x}$$

  - Minimizing this function is done by k-means by computing the squared Euclidean distance from each point to cluster mean, and find the minimum by computing $l^*(i) = \text{argmin}_c \left\| \mathbf{x} - \mu_c \right\|_2^2$ .
  - Each point is then reassigned to the clusters index by $l^*(i)$.
  - The centroid update step of the algorithm recomputes the mean of each cluster, updating $\mu_c$ for all c.

# k-means (via Bayesian Nonparametrics) (4)

- EM algorithm for mixtures of Gaussians is quite similar to the k-means algorithm.
  - When all Gaussians have fixed covariance equal to $\sigma I$ , the covariances need not be re-estimated during the M-step. The E-step takes the following form:

$$\gamma(z_{ic}) = \frac{\pi_c . \exp\left(-\frac{1}{2\sigma}\left\|\mathbf{x}_i - \mu_c\right\|_2^2\right)}{\sum_{j=1}^{c} \pi_j . \exp\left(-\frac{1}{2\sigma}\left\|\mathbf{x}_i - \mu_j\right\|_2^2\right)}$$

  - In the limit $\sigma \to 0$ the value of $\gamma(z_{ic})$ approaches zero for all c except for the one corresponding to the smallest distance $\left\|\mathbf{x}_i - \mu_c\right\|_2^2$ .
  - In this case, E-step is equivalent to the reassignment step of k-means and also the M-step exactly recomputes the center of the new clusters.

# k-means (via Bayesian Nonparametrics) (5)

- **Dirichlet Process Mixture Models**
  - We place a Dirichlet prior of dimension *k*, $\mathrm{Dir}(k, \pi_0)$ and assume the covariance of the Gaussians are fixed to $\sigma I$ and that the means are drawn from prior distribution $G_0$, we get the following model

$$\mu_1, \ldots, \mu_k \sim G_0$$
$$\pi \sim \mathrm{Dir}(k, \pi_0)$$
$$z_1, \ldots, z_n \sim \mathrm{Discrete}(\pi)$$
$$x_1, \ldots, x_n \sim N(\mu_{z_i}, \sigma I)$$

  - Gibbs sampling for inference in a DP mixture model is algorithm 2 in the following paper

R. M. Neal. **Markov chain sampling methods for Dirichlet process mixture models**. Journal of Computational and Graphical Statistics, 9:249–265, 2000.

# k-means (via Bayesian Nonparametrics) (6)

- **Algorithm 2**
  - Loop repeatedly through each of the data points and perform Gibbs moves on the cluster indicators for each point.
  - For $i = 1,...,n$, reassign $x_i$ to existing cluster c with probability $n_{-i,c}.N(x_i \mid \mu_c, \sigma I)/Z$ where $n_{-i,c}$ is the number of data points(excluding $x_i$) that are assigned to cluster c.
  - Start a new cluster with the probability
$$\frac{\alpha}{Z}\int N(x_i \mid \mu, \sigma I)\, dG_0(\mu)$$
where *Z* is the normalizing constant
  - If end up choosing to start a new cluster, select its mean from the posterior distribution obtained from the prior $G_0$ and the single sample $x_i$
  - After sampling all clusters, perform Gibbs moves on the means: sample $\mu_c$ given all points currently assigned to cluster c for all c.

# k-means (via Bayesian Nonparametrics) (7)

- **Adapting DP mixture model to Gaussian mixture scenario**
  - Write the model as

$$G \sim \mathrm{DP}(\alpha, G_0)$$
$$\phi_i \sim G \qquad \text{for } i = 1,\ldots,n$$
$$x_i \sim N(\phi_i, \sigma I) \quad \text{for } i = 1,\ldots,n$$

  - Think of a draw from G as choosing one of the infinite means $\mu_c$ drawn from $G_0$, with the property that the means are chosen with probability equal to the corresponding mixing weights. As a result, each $\phi_i$ is equal to $\mu_c$ for some c.

# k-means (via Bayesian Nonparametrics) (8)

- **DP mixture model connection to k-means(1)**
  - Take $G_0$, the prior distribution over means to be a zero-mean Gaussian with covariance, i.e., $\mu \sim N(0, \rho I)$ .
  - Given this, probability of starting a new cluster is equal to

$$\frac{\alpha}{Z}(2\pi(\rho+\sigma))^{-d/2}.\exp\left(-\frac{1}{2(\rho+\sigma)}\|x_i\|^2\right)$$

  - The probability of being assigned to cluster c equal to

$$\frac{n_{-i,c}}{Z}(2\pi\sigma)^{-d/2}.\exp\left(-\frac{1}{2\sigma}\|x_i-\mu_c\|_2^2\right)$$

  - Let $\sigma \to 0$, and write $\alpha = (1+\rho/\sigma)^{d/2}.\exp\left(\frac{-\lambda}{2\sigma}\right)$ for some $\lambda$

# k-means (via Bayesian Nonparametrics) (9)

- **DP mixture model connection to k-means(2)**
  - Let $\hat{\gamma}(z_{ic})$ be the posterior probability of point *i* assigned to cluster c and let $\hat{\gamma}(z_{ic,new})$ be the probability of starting a new cluster. We then obtain

$$\hat{\gamma}(z_{ic}) = \frac{n_{-i,c}.\exp\left(-\frac{1}{2\sigma}\left\|x_i-\mu_c\right\|^2\right)}{\exp\left(-\frac{\lambda}{2\sigma}-\frac{\left\|x_i\right\|^2}{2(\rho+\sigma)}\right)+\sum_{j=1}^{k} n_{-i,j}.\exp\left(-\frac{1}{2\sigma}\left\|x_i-\mu_j\right\|^2\right)}$$

$$\hat{\gamma}(z_{ic,new}) = \frac{\exp\left(-\frac{\lambda}{2\sigma}-\frac{\left\|x_i\right\|^2}{2(\rho+\sigma)}\right)}{\exp\left(-\frac{\lambda}{2\sigma}-\frac{\left\|x_i\right\|^2}{2(\rho+\sigma)}\right)+\sum_{j=1}^{k} n_{-i,j}.\exp\left(-\frac{1}{2\sigma}\left\|x_i-\mu_j\right\|^2\right)}$$

# k-means (via Bayesian Nonparametrics) (10)

- **DP mixture model connection to k-means(3)**
  - We can write $\hat{\gamma}(z_{ic,new})$ as

$$\exp\left(-\frac{1}{2\sigma}\left[\lambda+\frac{\sigma}{\rho+\sigma}\|x_i\|^2\right]\right)$$

  - If we set $\sigma \to 0$ with fixed $\rho$
    - the term $\lambda$ dominates
    - Probabilities defined before become binary
    - $\hat{\gamma}(z_{ic})$ and $\hat{\gamma}(z_{ic,new})$ become increasingly dominated by $\left\{\|x_i-\mu_1\|^2,\ldots,\|x_i-\mu_k\|^2,\lambda\right\}$
    - The smallest of these values receives a non-zero $\hat{\gamma}$ value.
    - The resulting form is **ANALOGOUS** to k-means
    - **Reassign a point to the cluster corresponding to the closest mean, unless the closest cluster has squared Euclidean distance greater than $\lambda$**

# k-means (via Bayesian Nonparametrics) (11)

- **DP mixture model connection to k-means(4)**
  - If we choose to start a new cluster, final step is to sample a new mean from the posterior based on the prior $G_0$ and single observation $x_i$
  - Prior and likelihood are Gaussian, so the posterior is also Gaussian
  - Let $\overline{x}_c$ be the mean of the points currently assigned to cluster c and $n_c$ be the number of points assigned to cluster c, then the posterior is a Gaussian with the mean and covariance as

$$\tilde{\mu}_c = \left(1 + \frac{\sigma}{\rho n_c}\right)^{-1} \overline{x}_c, \quad \tilde{\Sigma}_c = \frac{\sigma\rho}{\sigma + \rho n_c} I$$

  - As $\sigma \to 0$, the mean of the Gaussian approaches $\overline{x}_c$, covariance goes to zero and mass of distribution becomes concentrated at $\overline{x}_c$
  - **Thus algorithm similar to k-means is obtained with the exception that a new cluster is formed whenever a point is farther than $\lambda$ away from every cluster centroid**
  - DP-mean algorithm shown on next page is the end result **BUT IT DEPENDS** on ordering of the data (Come up with ordering?)

# k-means (via Bayesian Nonparametrics) (12)

- Objective function is simply k-means with an additional penalty based on the number of clusters with $\lambda$ controlling the tradeoff between traditional k-means and cluster term:

$$\min_{\{l_c\}_{c=1}^{k}} \sum_{c=1}^{k} \sum_{\mathbf{x} \in l_c} \left\| \mathbf{x} - \mu_c \right\|^2 + \lambda k$$

$$\text{where } \mu_c = \frac{1}{|l_c|} \sum_{\mathbf{x} \in l_c} \mathbf{x}$$

**Algorithm 1** DP-means

**Input:** $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$: input data, $\lambda$ : cluster penalty parameter
**Output:** Clustering $\ell_1, ..., \ell_k$ and number of clusters $k$

1. Init. $k = 1, \ell_1 = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ and $\boldsymbol{\mu}_1$ the global mean.
2. Init. cluster indicators $z_i = 1$ for all $i = 1, ..., n$.
3. Repeat until convergence
   - For each point $\boldsymbol{x}_i$
     - Compute $d_{ic} = \|\boldsymbol{x}_i - \boldsymbol{\mu}_c\|^2$ for $c = 1, ..., k$
     - If $\min_c d_{ic} > \lambda$, set $k = k + 1$, $z_i = k$, and $\boldsymbol{\mu}_k = \boldsymbol{x}_i$.
     - Otherwise, set $z_i = \text{argmin}_c d_{ic}$.
   - Generate clusters $\ell_1, ..., \ell_k$ based on $z_1, ..., z_k$: $\ell_j = \{\boldsymbol{x}_i \mid z_i = j\}$.
   - For each cluster $\ell_j$, compute $\boldsymbol{\mu}_j = \frac{1}{|\ell_j|} \sum_{\boldsymbol{x} \in \ell_j} \boldsymbol{x}$.

# Validity measures

- Davies-Bouldin Index

  D. L. Davies and D. W. Bouldin, **A cluster separation measure**, IEEE Trans. Pattern Analysis and Machine Intelligence, 1 (1979), pp. 224–227.

- Variance Ratio Criterion – VRC (Calinski-Harabasz Index)

  R. B. Calinski and J. Harabasz, **A dendrite method for cluster analysis**, Comm. in Statistics, 3 (1974), pp. 1–27.

- Dunn's Index

  J. C. Dunn, **Well separated clusters and optimal fuzzy partitions**, J. of Cybernetics, 4 (1974), pp. 95–104.
  M. Halkidi, Y. Batistakis, and M. Vazirgiannis, **On clustering validation techniques**, J. of Intelligent Information Systems, 17 (2001), pp. 107–145.

- Silhouette Width Criterion of Kaufman and Rousseeuw

  L. Kaufman and P. Rousseeuw, **Finding Groups in Data**, Wiley, 1990.

- Adjusted Rand Index

  L. Hubert and P. Arabie, **Comparing partitions**, J. of Classification, 2 (1985), pp. 193–218.

- Jaccard Index

  M. Halkidi, Y. Batistakis, and M. Vazirgiannis, **On clustering validation techniques**, J. of Intelligent Information Systems, 17 (2001), pp. 107–145.

# Calinski-Harabasz Index (VRC)(1)

Dataset: $\mathbf{X} = \{\mathbf{x}(1),...,\mathbf{x}(N)\}$ where $\mathbf{x}(j) \in \Re^n$

and a partition of data into *k* mutually disjoined clusters

$$VRC = \frac{\text{trace}(\mathbf{B})}{\text{trace}(\mathbf{W})} \times \frac{N-k}{k-1}$$

where **W** and **B** are the within-group and between-group dispersion matrices, respectively, defined as:

$$\mathbf{W} = \sum_{i=1}^{k} \sum_{l=1}^{N_i} (\mathbf{x}_i(l) - \overline{\mathbf{x}}_i)(\mathbf{x}_i(l) - \overline{\mathbf{x}}_i)^T, \ \mathbf{B} = \sum_{i=1}^{k} N_i (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})^T$$

where $N_i$ is the number of objects assigned to *i*th cluster, $\mathbf{x}_i(l)$ is *l*th object assigned to that cluster, $\overline{\mathbf{x}}_i$ is *n*-th dimensional vector of sample means within that cluster (cluster centroid), $\overline{\mathbf{x}}$ is *n*-th dimensional vector of overall sample means (data centroid).

# Calinski-Harabasz Index (VRC)(2)

Within-group and between-group matrices sum up to the scatter matrix of the data set:

$$\mathbf{T} = \mathbf{W} + \mathbf{B}$$

where $\mathbf{T} = \sum_{l=1}^{N} (\mathbf{x}(l) - \bar{\mathbf{x}})(\mathbf{x}(l) - \bar{\mathbf{x}})^T$

- Trace of W is the sum of the within-cluster variances.
- Trace of B is the sum of the between-cluster variances.
- Compact and separated clusters are expected to have small values of trace(W) and large values of trace(B).
- Better the data partition, the greater the value of the ratio between trace(B) and trace(W).
- Normalization term (N-k)/(k-1) prevents this ratio to increase monotonically with the number of clusters, thus making VRC and optimization (maximization) criteria with respect to k.

# Davies-Bouldin Index(1)

- Related to VRC as based on ratio involving within-group and between-group distances

$$DB = \frac{1}{k}\sum_{i=1}^{k} D_i$$

where $D_i = \max_{j \neq i}\left\{D_{i,j}\right\}$,

$D_{i,j}$ is the within-to-between cluster spread for the *i*th and *j*th clusters, i.e.

$$D_{i,j} = (\bar{d}_i + \bar{d}_j) / d_{i,j}$$

$\bar{d}_i$ and $d_{i,j}$ are the average within-group distance for the *i*th cluster and the inter-group distance between clusters *i* and *j*, respectively. Also,

$$\bar{d}_i = (1/N_i)\sum_{l=1}^{N_i}\left\|\mathbf{x}_i(l) - \bar{\mathbf{x}}_i\right\|, \quad d_{i,j} = \left\|\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j\right\|$$

where $\|.\|$ is a norm (e.g., Euclidean)

# Davies-Bouldin Index(2)

- $D_i$ represents the worst-case within-to-between cluster spread involving the *i*th cluster.
- Minimizing $D_i$ for all clusters minimizes the Davies-Bouldin Index.
- Good partitions, composed of compact and separated clusters, are distinguished by small values of *DB*.

# Dunn's Index

- It is based on geometrical measure of cluster compactness and separation and defined as

$$DN = \min_{\substack{p,\, q \in \{1,\ldots,k\} \\ p \neq q}} \left\{ \frac{\delta_{p,q}}{\max_{l \in \{1,\ldots,k\}} \Delta_l} \right\}$$

where $\Delta_l$ is the diameter of the *l*th cluster, defined as

$$\Delta_l = \max_{i \neq j} \left\| \mathbf{x}_l(i) - \mathbf{x}_l(j) \right\|$$

Set distance $\delta_{p,q}$ is defined as the minimum distance between a pair of objects across clusters *p* and *q*.

$$\delta_{p,q} = \min_{i \neq j} \left\| \mathbf{x}_p(i) - \mathbf{x}_q(j) \right\|$$ **(single linkage)**

- Partitions composed of compact and separated clusters are distinguished by large values of *DN*.

# Variants of Dunn's Index(1)

- **Set distance** and **diameter** (previous slide) were generalized to give 17 variants of original Dunn's index (combination of 5 set distances and 3 diameters)

J. C. Bezdek and N. R. Pal, **Some new indexes of cluster validity**, IEEE Trans. Systems, Man and Cybernetics –B, 28 (1998), pp. 301–315.

$$\delta_{p,q} = \max_{i,j} \left\| \mathbf{x}_p(i) - \mathbf{x}_q(j) \right\|$$ **(complete linkage)**

$$\delta_{p,q} = \frac{1}{N_p N_q} \sum_{i=1}^{N_p} \sum_{j=1}^{N_q} \left\| \mathbf{x}_p(i) - \mathbf{x}_q(j) \right\|$$ **(average linkage)**

$$\delta_{p,q} = \left\| \bar{\mathbf{x}}_p - \bar{\mathbf{x}}_q \right\|$$ (same as inter-group distance in Davis-Bouldin Index)

$$\delta_{p,q} = \frac{1}{N_p + N_q} \left( \sum_{i=1}^{N_p} \left\| \mathbf{x}_p(i) - \bar{\mathbf{x}}_q \right\| + \sum_{j=1}^{N_q} \left\| \mathbf{x}_q(i) - \bar{\mathbf{x}}_p \right\| \right)$$ (Hybrid of **average linkage**)

$$\delta_{p,q} = \max \left\{ \max_i \min_j \left\| \mathbf{x}_p(i) - \mathbf{x}_q(j) \right\|, \max_j \min_i \left\| \mathbf{x}_p(i) - \mathbf{x}_q(j) \right\| \right\}$$ **(Hausdorff metric)**

# Variants of Dunn's Index(2)

- Alternative definitions for diameter

$$\Delta_l = \frac{2}{N_l(N_l-1)} \sum_{i=1}^{N_l} \sum_{j=1}^{i} \left\| \mathbf{x}_l(i) - \mathbf{x}_l(j) \right\|$$

(average distance among all $N_l(N_l-1)/2$ pairs of the *l*th cluster)

$$\Delta_l = \frac{2}{N_l} \sum_{i=1}^{N_l} \left\| \mathbf{x}_l(i) - \bar{\mathbf{x}}_l \right\|$$

(two times the cluster radius, estimated as the average distance among the objects of the *l*th cluster and it's prototype)

# Silhouette Width Criterion(1)

- It is based on geometrical considerations about compactness and separation of clusters.
- Let *j*th object of the dataset, $\mathbf{x}(j)$, belong to a given cluster $p \in \{1,...,k\}$
- Let the average distance of this object to all other objects in cluster *p* be denoted by $a_{p,j}$
- Let the average distance of this object to all objects in another cluster *q*, $q \neq p$, be called $d_{q,j}$
- Let $b_{p,j}$ be the minimum $d_{q,j}$ computed over $q = 1,...,k$ ,which represents the average dissimilarity of object $\mathbf{x}(j)$ to its closest neighboring cluster

# Silhouette Width Criterion(2)

- Silhouette of the individual object $\mathbf{x}(j)$ is defined as

$$s_{\mathbf{x}(j)} = \frac{b_{p,j} - a_{p,j}}{\max\{a_{p,j}, b_{p,j}\}}$$

where denominator is just a normalization term

- The higher $s_{\mathbf{x}(j)}$, the better the assignment of $\mathbf{x}(j)$ to cluster *p*.
- If *p* is singleton, i.e., only unique $\mathbf{x}(j)$ then $s_{\mathbf{x}(j)} = 0$. This prevents the Silouette Width Criterion, defined as the average of $s_{\mathbf{x}(j)}$ over $j = 1, 2, ..., N$, i.e., $SWC = \frac{1}{N}\sum_{j=1}^{N} S_{\mathbf{x}(j)}$

to elect the trivial solution *k*=*N*, with each object of the data set forming a cluster on its own, as the best one.

# Silhouette Width Criterion(3)

- The best partition is achieved when SWC is maximized, which implies minimizing the intra-group distance ( $a_{p,j}$ ) while maximizing the inter-group distance ( $b_{p,j}$ ).

# Silhouette Variations(1)

- **Simplified:** The computation of all distances among all objects can be replaced with a simplified one based on distance among objects and cluster centroids (idea is to replace average distances with distances to the mean points).
  - $a_{p,j}$ is redefined as the dissimilarity of *i*th object to the centroid of its cluster *p*.
  - $d_{q,j}$ is computed as the dissimilarity of the *i*th object to the centroid of the cluster *q*, $q \neq p$.
  - $b_{p,j}$ becomes the dissimilarity of the *i*th object to the centroid of its closest neighboring cluster

# Silhouette Variations(2)

- **Alternative Silhouette:** We have an alternative definition of the silhouette of an individual object:

$$s_{\mathbf{x}(j)} = \frac{b_{p,j}}{a_{p,j} + \varepsilon}$$

  where $\varepsilon$ is a small constant (e.g. 10^-6 for normalized data) used to avoid division by zero when $a_{p,j} = 0$.

- Both definition of silhouette are intended to favour larger values of $b_{p,j}$ and lower values of $a_{p,j}$, with previous definition in linear and this new definition as non-linear case.

- **Hybrid Silhouette:** Combining **alternative silhouette** with **simplified silhouette**.

# Adjusted Rand Index(1)

- One of the main difficulties in classification problems consists on the correct evaluation of the classifier performance.
- Conventionally measures like **Mean Squared Error (MSE)** or the **Classification Correct Rate** are used.
- Other measures like **AUC** (area in percentage under the ROC curve), Sensitivity and Specificity are also used.
- All these measures compare the labeled outcome of the supervised classification algorithm with the known labeled target
- This leads to poor results due to the fact that the output labels could be switched even if the classes are well identified.
- We want a measure which **evaluates how well the algorithm split the input data in different classes by looking at the relationship between elements of each class AND NOT THE LABELS**
- **Solution:** Use ARI. It is a measure of agreement between two partitions: one given by the **clustering process** and the other defined by **external criteria**.

# Adjusted Rand Index(2)

- Lower value of ARI for bad classification results and higher value of ARI for good classification results
- It can be used to perform feature selection if we split each feature in non-overlapping equal intervals and compare the partition derived from the split with the one given by the targets.
- Thus we evaluate each feature's discriminant power and rank the features according to the computed ARI value.
- Finally, we select the most discriminant feature to apply in classification algorithm

J. M. Santos, and M. Embrechts. **On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification**. ICANN'09 Proceedings of the 19th International Conference on Artificial Neural Networks: Part II, Lecture Notes in Computer Science, Volume 5769, 2009, pp. 175-184

# Adjusted Rand Index(3)

- Dataset: A set of *n* objects $S = \{O_1, O_2, ..., O_n\}$.
- We have two different partitions of the objects in *S*, $U = \{u_1, u_2, ..., u_R\}$ and $V = \{v_1, v_2, ..., v_C\}$ such that $\bigcup_{i=1}^{R} u_i = S = \bigcup_{i=1}^{C} v_j$ and $u_i \bigcap u_{i'} = \varnothing = v_j \bigcap v_{j'}$ for $1 \le i \ne i' \le R$ and $1 \le j \ne j' \le C$.
- Contingency table:

| Partition | | $V$ | | | | |
|---|---|---|---|---|---|---|
| | Group | $v_1$ | $v_2$ | $\cdots$ | $v_C$ | Total |
| $U$ | $u_1$ | $t_{11}$ | $t_{12}$ | $\cdots$ | $t_{1C}$ | $t_{1.}$ |
| | $u_2$ | $t_{21}$ | $t_{22}$ | $\cdots$ | $t_{2C}$ | $t_{2.}$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| | $u_R$ | $t_{R1}$ | $t_{R2}$ | $\cdots$ | $t_{RC}$ | $t_{R.}$ |
| Total | | $t_{.1}$ | $t_{.2}$ | $\cdots$ | $t_{.C}$ | $t_{..} = n$ |

# Adjusted Rand Index(4)

- $t_{rc}$ in the contingency table represents the number of objects that were classified in the *r*th subset of partition *R* and in the *c*th subset of the partition *C*.
- From the total number of possible combination of pairs $\binom{n}{2}$ from a given set we can represent the results in four different types of pairs:
  - **a** – objects in a pair are placed in the same group in *U* and in the same group in *V*
  - **b** – objects in a pair are placed in the same group in *U* and in different groups in *V*
  - **c** – objects in a pair are placed in the same group in *V* and in different groups in *U*
  - **d** – objects in a pair are placed in different groups in *U* and in different groups in *V*

# Adjusted Rand Index(5)

- Simplified 2 X 2 contingency table

| Partition $U$ | $V$ | |
|---|---|---|
| | Pair in same group | Pair in different groups |
| Pair in same group | $a$ | $b$ |
| Pair in different groups | $c$ | $d$ |

- The values are given on next slide

# Adjusted Rand Index(6)

$$a = \sum_{r=1}^{R}\sum_{c=1}^{C}\binom{t_{rc}}{2} = \left(\sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^2 - n\right)/2$$

$$b = \sum_{r=1}^{R}\binom{t_{r.}}{2} - a = \left(\sum_{r=1}^{R} t_{r.}^2 - \sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^2\right)/2$$

$$c = \sum_{c=1}^{C}\binom{t_{.c}}{2} - a = \left(\sum_{c=1}^{C} t_{.c}^2 - \sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^2\right)/2$$

$$d = \binom{n}{2} - a - b - c = \binom{n}{2} - \sum_{r=1}^{R}\binom{t_{r.}}{2} - \sum_{c=1}^{C}\binom{t_{.c}}{2} + a$$

$$= \left(\sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^2 + n^2 - \sum_{r=1}^{R} t_{r.}^2 - \sum_{c=1}^{C} t_{.c}^2\right)/2$$

# Adjusted Rand Index(7)

- **Rand Index** can then be computed as

$$RI = \frac{a+d}{a+b+c+d}$$

  and it basically weights those objects that were classified together in both *U* and *V*.

- Problem with *RI* is that the expected value of the *RI* of two random partitions does not take a constant value (say zero) or that the Rand statistic approaches its upper limit of unity as the number of cluster increases.

# Adjusted Rand Index(8)

- With the intention to overcome limitations with *RI*, we have **Fowlkes-Mallows Index**

$$FMI = \frac{a}{\sqrt{(a+b)(a+c)}}$$

- Another one is **Adjusted Rand Index** (ARI) which has become one of the most successful cluster validation indices and is recommended as the index of choice for measuring agreement between two partitions in clustering analysis with different number of clusters.

# Adjusted Rand Index(9)

- ARI can then be computed by

$$ARI = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]}$$

or

$$ARI = \frac{\binom{n}{2}\sum_{r=1}^{R}\sum_{c=1}^{C}\binom{t_{rc}}{2} - \left[\sum_{r=1}^{R}\binom{t_{r.}}{2}\sum_{c=1}^{C}\binom{t_{.c}}{2}\right]}{\frac{1}{2}\binom{n}{2}\left[\sum_{r=1}^{R}\binom{t_{r.}}{2}+\sum_{c=1}^{C}\binom{t_{.c}}{2}\right] - \left[\sum_{r=1}^{R}\binom{t_{r.}}{2}\sum_{c=1}^{C}\binom{t_{.c}}{2}\right]}$$

with expected value zero and maximum value 1.

# WGK Correlation Index(1)

- Goodman-Kruksal (GK) index measures rank correlation between two sequences $A = \{a_1, \ldots, a_n\}$ and $B = \{b_1, \ldots, b_n\}$ in terms of the numbers of *concordant* and *discordant* pairs in *A* and *B*.
- Pairs $(a_i, a_j)$ and $(b_i, b_j)$ are concordant if either $a_i < a_j$ and $b_i < b_j$ or $a_i > a_j$ and $b_i > b_j$.
- They are discordant if either $a_i < a_j$ and $b_i > b_j$ or $a_i > a_j$ and $b_i < b_j$.
- The index is then defined as $\gamma = \frac{S_+ - S_-}{S_+ + S_-}$

  where $S_+$ and $S_-$ are the numbers of concordant and discordant pairs in *A* and *B*, respectively. $\gamma \in [-1, 1]$.

L. A. Goodman and W. H. Kruskal, **Measures of association for cross-classifications**, J. of the Am. Statistical Association, 49 (1954), pp. 732–764.

# WGK Correlation Index(2)

- GK is insensitive to the element values of sequences *A* and *B* because only ranks of these elements are considered.
- A weighted version of GK can bring together both the sensitivity of the original index to the rankings of sequences *A* and *B* and the sensitivity of the classic Pearson Coefficient to the values of these sequences by rewriting

$$\gamma = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_{ij}}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} |w_{ij}|}$$

where $|.|$ stands for absolute value.

# WGK Correlation Index(3)

- The weight $w_{ij}$ is given by:

$$w_{ij} = \begin{cases} w_{ij}^A / w_{ij}^B & \text{if } w_{ij}^B \neq 0 \\ 1 & \text{if } w_{ij}^A = 0 \text{ and } w_{ij}^B = 0 \\ 0 & \text{otherwise} \end{cases}$$

with $w_{ij}^A = \text{sign}(a_i - a_j)$ and $w_{ij}^B = \text{sign}(b_i - b_j)$.

- With this definition GK is magnitude insensitive because the values of $w_{ij}$ are constrained to -1, 0, or +1 irrespective of the values in sequences *A* and *B*.

# WGK Correlation Index(4)

- Since concordance and discordance are both a matter of degree, a weighted version of this index can be obtained by replacing the terms in the numerator of $\gamma$ with continuous version of $w_{ij}, w_{ij}^{A}$, and $w_{ij}^{B}$.

$$\hat{w}_{ij} = \begin{cases} \min\{\hat{w}_{ij}^{A} / \hat{w}_{ij}^{B}, \hat{w}_{ij}^{B} / \hat{w}_{ij}^{A}\} & \text{if } \hat{w}_{ij}^{A} \text{ and } \hat{w}_{ij}^{B} \text{ have the same sign} \\ \max\{\hat{w}_{ij}^{A} / \hat{w}_{ij}^{B}, \hat{w}_{ij}^{B} / \hat{w}_{ij}^{A}\} & \text{if } \hat{w}_{ij}^{A} \text{ and } \hat{w}_{ij}^{B} \text{ have opposite signs} \\ 1 & \text{if } \hat{w}_{ij}^{A} = \hat{w}_{ij}^{B} = 0 \\ 0 & \text{otherwise} \end{cases}$$

with components shown on next slide.

# WGK Correlation Index(5)

$$\hat{w}_{ij}^{A} = \begin{cases} \dfrac{a_i - a_j}{a_{\max} - a_{\min}} & \text{if } a_{\max} \neq a_{\min} \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{w}_{ij}^{B} = \begin{cases} \dfrac{b_i - b_j}{b_{\max} - b_{\min}} & \text{if } b_{\max} \neq b_{\min} \\ 0 & \text{otherwise} \end{cases}$$

where $a_{\max}, a_{\min}, b_{\max},$ and $b_{\min}$ are the maximum and minimum elements of sequences *A and B*.

# WGK Correlation Index(6)

- $\hat{w}_{ij}^{A}$ and $\hat{w}_{ij}^{B}$ belong to [-1,+1] and represent the (signed) percentage differences between the values at *i*th and *j*th element of the corresponding sequence
- The weight $\hat{w}_{ij}$ is such that
  - It is positive if pairs $(a_i, a_j)$ and $(b_i, b_j)$ are concordant. ( $\hat{w}_{ij}^{A}$ and $\hat{w}_{ij}^{B}$ have the same sign or both are null)
  - It is negative if pairs $(a_i, a_j)$ and $(b_i, b_j)$ are disconcordant. ( $\hat{w}_{ij}^{A}$ and $\hat{w}_{ij}^{B}$ have opposite signs)
  - It is null in case of a neutral (either $\hat{w}_{ij}^{A}$ or $\hat{w}_{ij}^{B}$ is null)
  - Full concordance( $\hat{w}_{ij} = 1$) when $\hat{w}_{ij}^{A} = \hat{w}_{ij}^{B}$
  - Full discordance ( $\hat{w}_{ij} = -1$) when $\hat{w}_{ij}^{A} = -\hat{w}_{ij}^{B}$

# WGK Correlation Index(7)

- IT IS NOT CORRECT to assume that a larger difference between $\left|\hat{w}_{ij}^{A}\right|$ and $\left|\hat{w}_{ij}^{B}\right|$ should mean a greater discordance between pairs $(a_i, a_j)$ and $(b_i, b_j)$ . **Why?**
  - **Because**, increasing the abolute value of one of these weights is equivalent to reducing the absolute value of the other.
  - **But,** reducing the absolute value of a given weight means driving this weight towards zero and further changing sign.
  - This means reducing discordance towards neutrality and further turning to concordance

# WGK Correlation Index(8)

- Weighted Goodman-Kruskal index (WGK) is thus defined as

$$\gamma = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_{ij}}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} |w_{ij}|}$$

- It can be shown that, as it is in the case for Pearson, the extreme values for WGK ( $\hat{\gamma} = 1$ or $-1$ ) are obtained *iff A* is a linear (or affine) function of *B*.

# Wilcoxon Mann Whitney Test(1)

- In independent samples, we apply independent sample t-test when we had normality of the sample mean for each sample.
- If there is a violation of this assumption, we'll apply non-parametric test for independent samples – the Wilcoxon Mann Whitney test
- It can be used only when
  - Data is regraded as a random sample from their respective populations
  - Observations within each sample is independent of one another
  - The two samples are independent of one another
  - The efficiency of the Wilcoxon Mann Whitney test is 0.95 with respect to parametric tests like the t-test or the z-test even if the data are normal

https://epilab.ich.ucl.ac.uk/coursematerial/statistics/non_parametric/wilcox_mann_whitney.html

# Wilcoxon Mann Whitney Test(2)

- For small samples, use direct method:
  - Choose the sample for which the ranks seem to be smaller. Call this "sample 1" and call the other sample "sample 2"
  - Taking each observation in sample 1, count the number of observations in sample 2 that are smaller than it (count a half for any that are equal to it)
  - The total of these counts is **U**

# Wilcoxon Mann Whitney Test(3)

- For large samples,
  - Arrange all the observations into a single ranked series. That is, ranks all the observations without regard to which sample they are in.
  - Add up the ranks for the observations which come from sample 1. The sum of ranks in sample 2 follows by calculation, since the sum of all the ranks equal *N(N+1)/2* where *N* is the total number of observations.
  - U is given by

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

    where $n_1$ is the sample size for sample 1, and $R_1$ is the sume of the ranks in sample1. Alternatively,

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

  - The smaller value of $U_1$ and $U_2$ is the one used when consulting the significance table.
  - Null hypothesis: probability that member of the 1st population drawn at random will exceed a member of the 2nd population drawn at random=0.5

# Table A5.07: Critical Values for the Wilcoxon/Mann-Whitney Test (U)

**Nondirectional α=.05 (Directional α=.025)**

| $n_1$ \ $n_2$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 2 | - | - | - | - | - | - | - | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| 3 | - | - | - | - | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 | 5 | 6 | 6 | 7 | 7 | 8 |
| 4 | - | - | - | 0 | 1 | 2 | 3 | 4 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 11 | 12 | 13 | 13 |
| 5 | - | - | 0 | 1 | 2 | 3 | 5 | 6 | 7 | 8 | 9 | 11 | 12 | 13 | 14 | 15 | 17 | 18 | 19 | 20 |
| 6 | - | - | 1 | 2 | 3 | 5 | 6 | 8 | 10 | 11 | 13 | 14 | 16 | 17 | 19 | 21 | 22 | 24 | 25 | 27 |
| 7 | - | - | 1 | 3 | 5 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 | 22 | 24 | 26 | 28 | 30 | 32 | 34 |
| 8 | - | 0 | 2 | 4 | 6 | 8 | 10 | 13 | 15 | 17 | 19 | 22 | 24 | 26 | 29 | 31 | 34 | 36 | 38 | 41 |
| 9 | - | 0 | 2 | 4 | 7 | 10 | 12 | 15 | 17 | 21 | 23 | 26 | 28 | 31 | 34 | 37 | 39 | 42 | 45 | 48 |
| 10 | - | 0 | 3 | 5 | 8 | 11 | 14 | 17 | 20 | 23 | 26 | 29 | 33 | 36 | 39 | 42 | 45 | 48 | 52 | 55 |
| 11 | - | 0 | 3 | 6 | 9 | 13 | 16 | 19 | 23 | 26 | 30 | 33 | 37 | 40 | 44 | 47 | 51 | 55 | 58 | 62 |
| 12 | - | 1 | 4 | 7 | 11 | 14 | 18 | 22 | 26 | 29 | 33 | 37 | 41 | 45 | 49 | 53 | 57 | 61 | 65 | 69 |
| 13 | - | 1 | 4 | 8 | 12 | 16 | 20 | 24 | 28 | 33 | 37 | 41 | 45 | 50 | 54 | 59 | 63 | 67 | 72 | 76 |
| 14 | - | 1 | 5 | 9 | 13 | 17 | 22 | 26 | 31 | 36 | 40 | 45 | 50 | 55 | 59 | 64 | 67 | 74 | 78 | 83 |
| 15 | - | 1 | 5 | 10 | 14 | 19 | 24 | 29 | 34 | 39 | 44 | 49 | 54 | 59 | 64 | 70 | 75 | 80 | 85 | 90 |
| 16 | - | 1 | 6 | 11 | 15 | 21 | 26 | 31 | 37 | 42 | 47 | 53 | 59 | 64 | 70 | 75 | 81 | 86 | 92 | 98 |
| 17 | - | 2 | 6 | 11 | 17 | 22 | 28 | 34 | 39 | 45 | 51 | 57 | 63 | 67 | 75 | 81 | 87 | 93 | 99 | 105 |
| 18 | - | 2 | 7 | 12 | 18 | 24 | 30 | 36 | 42 | 48 | 55 | 61 | 67 | 74 | 80 | 86 | 93 | 99 | 106 | 112 |
| 19 | - | 2 | 7 | 13 | 19 | 25 | 32 | 38 | 45 | 52 | 58 | 65 | 72 | 78 | 85 | 92 | 99 | 106 | 113 | 119 |
| 20 | - | 2 | 8 | 14 | 20 | 27 | 34 | 41 | 48 | 55 | 62 | 69 | 76 | 83 | 90 | 98 | 105 | 112 | 119 | 127 |

**Nondirectional α=.01 (Directional α=.005)**

| $n_1$ \ $n_2$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 0 | 0 |
| 3 | - | - | - | - | - | - | - | - | 0 | 0 | 0 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 3 | 3 |
| 4 | - | - | - |  |  | 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 5 | 5 | 6 | 6 | 7 | 8 |
| 5 | - | - | - |  | 0 | 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| 6 | - | - | - | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 9 | 10 | 11 | 12 | 13 | 15 | 16 | 17 | 18 |
| 7 | - | - | - | 0 | 1 | 3 | 4 | 6 | 7 | 9 | 10 | 12 | 13 | 15 | 16 | 18 | 19 | 21 | 22 | 24 |
| 8 | - | - | - | 1 | 2 | 4 | 6 | 7 | 9 | 11 | 13 | 15 | 17 | 18 | 20 | 22 | 24 | 26 | 28 | 30 |
| 9 | - | - | 0 | 1 | 3 | 5 | 7 | 9 | 11 | 13 | 16 | 18 | 20 | 22 | 24 | 27 | 29 | 31 | 33 | 36 |
| 10 | - | - | 0 | 2 | 4 | 6 | 9 | 11 | 13 | 16 | 18 | 21 | 24 | 26 | 29 | 31 | 34 | 37 | 39 | 42 |
| 11 | - | - | 0 | 2 | 5 | 7 | 10 | 13 | 16 | 18 | 21 | 24 | 27 | 30 | 33 | 36 | 39 | 42 | 45 | 46 |
| 12 | - | - | 1 | 3 | 6 | 9 | 12 | 15 | 18 | 21 | 24 | 27 | 31 | 34 | 37 | 41 | 44 | 47 | 51 | 54 |
| 13 | - | - | 1 | 3 | 7 | 10 | 13 | 17 | 20 | 24 | 27 | 31 | 34 | 38 | 42 | 45 | 49 | 53 | 56 | 60 |
| 14 | - | - | 1 | 4 | 7 | 11 | 15 | 18 | 22 | 26 | 30 | 34 | 38 | 42 | 46 | 50 | 54 | 58 | 63 | 67 |
| 15 | - | - | 2 | 5 | 8 | 12 | 16 | 20 | 24 | 29 | 33 | 37 | 42 | 46 | 51 | 55 | 60 | 64 | 69 | 73 |
| 16 | - | - | 2 | 5 | 9 | 13 | 18 | 22 | 27 | 31 | 36 | 41 | 45 | 50 | 55 | 60 | 65 | 70 | 74 | 79 |
| 17 | - | - | 2 | 6 | 10 | 15 | 19 | 24 | 29 | 34 | 39 | 44 | 49 | 54 | 60 | 65 | 70 | 75 | 81 | 86 |
| 18 | - | - | 2 | 6 | 11 | 16 | 21 | 26 | 31 | 37 | 42 | 47 | 53 | 58 | 64 | 70 | 75 | 81 | 87 | 92 |
| 19 | - | 0 | 3 | 7 | 12 | 17 | 22 | 28 | 33 | 39 | 45 | 51 | 56 | 63 | 69 | 74 | 81 | 87 | 93 | 99 |
| 20 | - | 0 | 3 | 8 | 13 | 18 | 24 | 30 | 36 | 42 | 46 | 54 | 60 | 67 | 73 | 79 | 86 | 92 | 99 | 105 |

$U_{obt}$ is the lesser of the two calculated test statistics ($U_1$ & $U_2$). If $U_{obt} \leq U_{crit}$, reject $H_0$.
Dashes (-) indicate that the sample size is too small to reject the Null Hypothesis at the chosen α level.

If n > 20 this table cannot be used. A **p** can be computed for $U_{obt}$, using the normal distribution approximation:

$$z_U = \frac{\left| U_{obt} - \left( n_1 n_2 / 2 \right) \right|}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

# Evidence Accumulation-based Clustering(1)

- Basic Idea:
  - Based on the idea of evidence accumulation for combining the results of multiple clusterings.
  - Initially, *n* *d*-dimensional data is decomposed into a large number of compact clusters using **K-means algorithm**, with several clusterings obtained by *N* random initialization of the K-means.
  - Take the co-occurences of pairs of patterns in the same cluster as votes for their association, the data partitions are mapped into a co-association matrix of patterns.
  - The *n* x *n* matrix represents a new similarity measure between patterns.
  - The final clusters are obtained by applying a **minimum spanning tree** (MST) based clustering algorithm on this matrix.

A. L. N. Fred and A. K. Jain. **Data clustering using evidence accumulation**. In Proc. ICPR, 2002.

# Evidence Accumulation-based Clustering(2)

- **Split**
  - Decompose multidimensional data into a large number of small, spherical clusters using K-means algorithm
- **Combine**
  - Take the co-occurences of pairs of patterns in the same cluster as votes for their association, the data partitions produced by multiple runs of K-means are mapped into a *n* x *n* co-association matrix:

$$co\_assoc(i,j) = \frac{votes_{ij}}{N}$$

where *N* is the number of clusterings and $votes_{ij}$ is the number of times the pattern pair (*i,j*) is assigned to the same cluster among the *N* clusterings

# Evidence Accumulation-based Clustering(3)

- **Merge**
  - Apply a minimum spanning tree (MST) algorithm, cutting weak links at a threshold of *t*, this is equivalent to cutting the dendrogram produced by the single link (SL) method over this similarity matrix at the threshold *t*, thus merging clusters produced in the splitting phase.

# Evidence Accumulation-based Clustering(4)

- Algorithm:

Data clustering using Evidence Accumulation.

*Input:*
$n$ $d$—dimensional patterns;
$k\_min$ - minimum initial number of clusters;
$k\_max$ - maximum initial number of clusters;
$N$ - number of clusterings.

*Output:* Data partitioning.

*Initialization:* Set $co\_assoc$ to a null $n \times n$ matrix.

**1.** Do $N$ times:
- **1.1.** Randomly select $k$ in the interval $[k\_min; k\_max]$.
- **1.2.** Randomly select $k$ cluster centers.
- **1.3.** Run the K-means algorithm with the above $k$ and initialization, and produce a partition $P$.
- **1.4.** Update the co-association matrix: for each pattern pair, $(i, j)$, in the same cluster in $P$, set $co\_assoc(i, j) = co\_assoc(i, j) + \frac{1}{N}$.

**2.** Detect consistent clusters in the co-association matrix using the SL technique: compute the SL dendrogram and identify the final clusters as the ones with the highest lifetime.