Data Mining and Analysis: Fundamental Concepts and Algorithms

datamininingbook.info

Mohammed J. Zaki[1] Wagner Meira Jr.[2]

[1]Department of Computer Science
Rensselaer Polytechnic Institute, Troy, NY, USA

[2]Department of Computer Science
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Chapter 17: Clustering Validation

Selection of material from Chapter 17

Clustering Validation and Evaluation

Cluster validation and assessment encompasses three main tasks: *clustering evaluation* seeks to assess the goodness or quality of the clustering, *clustering stability* seeks to understand the sensitivity of the clustering result to various algorithmic parameters, for example, the number of clusters, and *clustering tendency* assesses the suitability of applying clustering in the first place, that is, whether the data has any inherent grouping structure.

Validity measures can be divided into three main types:

- **External:** External validation measures employ criteria that are not inherent to the dataset, e.g., class labels.
- **Internal:** Internal validation measures employ criteria that are derived from the data itself, e.g., intracluster and intercluster distances.
- **Relative:** Relative validation measures aim to directly compare different clusterings, usually those obtained via different parameter settings for the same algorithm.

External Measures

External measures assume that the correct or ground-truth clustering is known *a priori*, which is used to evaluate a given clustering.

Let $\mathbf{D} = \{\mathbf{x}_i\}_{i=1}^{n}$ be a dataset consisting of n points in a d-dimensional space, partitioned into k clusters. Let $y_i \in \{1, 2, \ldots, k\}$ denote the ground-truth cluster membership or label information for each point.

The ground-truth clustering is given as $\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$, where the cluster T_j consists of all the points with label j, i.e., $T_j = \{\mathbf{x}_i \in \mathbf{D} | y_i = j\}$. We refer to $\mathcal{T}$ as the ground-truth *partitioning*, and to each T_i as a *partition*.

Let $\mathcal{C} = \{C_1, \ldots, C_r\}$ denote a clustering of the same dataset into r clusters, obtained via some clustering algorithm, and let $\hat{y}_i \in \{1, 2, \ldots, r\}$ denote the cluster label for $\mathbf{x}_i$.

External Measures

External evaluation measures try capture the extent to which points from the same partition appear in the same cluster, and the extent to which points from different partitions are grouped in different clusters.

All of the external measures rely on the $r \times k$ *contingency table* **N** that is induced by a clustering $\mathcal{C}$ and the ground-truth partitioning $\mathcal{T}$, defined as follows

$$\mathbf{N}(i,j) = n_{ij} = |C_i \cap T_j|$$

The count n_{ij} denotes the number of points that are common to cluster C_i and ground-truth partition T_j.

Let $n_i = |C_i|$ denote the number of points in cluster C_i, and let $m_j = |T_j|$ denote the number of points in partition T_j.

The contingency table can be computed from $\mathcal{T}$ and $\mathcal{C}$ in $O(n)$ time by examining the partition and cluster labels, y_i and $\hat{y}_i$, for each point $\mathbf{x}_i \in \mathbf{D}$ and incrementing the corresponding count $n_{y_i \hat{y}_i}$.

Matching Based Measures: Purity

Purity quantifies the extent to which a cluster C_i contains entities from only one partition:

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

The purity of clustering $\mathcal{C}$ is defined as the weighted sum of the clusterwise purity values:

$$purity = \sum_{i=1}^{r} \frac{n_i}{n} purity_i = \frac{1}{n} \sum_{i=1}^{r} \max_{j=1}^{k} \{n_{ij}\}$$

where the ratio $\frac{n_i}{n}$ denotes the fraction of points in cluster C_i.

Matching Based Measures: Maximum Matching

The maximum matching measure selects the mapping between clusters and partitions, such that the sum of the number of common points (n_{ij}) is maximized, provided that only one cluster can match with a given partition.

Let G be a bipartite graph over the vertex set $V = \mathcal{C} \cup \mathcal{T}$, and let the edge set be $E = \{(C_i, T_j)\}$ with edge weights $w(C_i, T_j) = n_{ij}$. A *matching* M in G is a subset of E, such that the edges in M are pairwise nonadjacent, that is, they do not have a common vertex.

The *maximum weight matching* in G is given as:

$$match = \arg\max_M \left\{ \frac{w(M)}{n} \right\}$$

where $w(M)$ is the sum of the sum of all the edge weights in matching M, given as $w(M) = \sum_{e \in M} w(e)$

Matching Based Measures: F-measure

Given cluster C_i, let j_i denote the partition that contains the maximum number of points from C_i, that is, $j_i = \max_{j=1}^{k}\{n_{ij}\}$.

The *precision* of a cluster C_i is the same as its purity:

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

The *recall* of cluster C_i is defined as

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

where $m_{j_i} = |T_{j_i}|$.

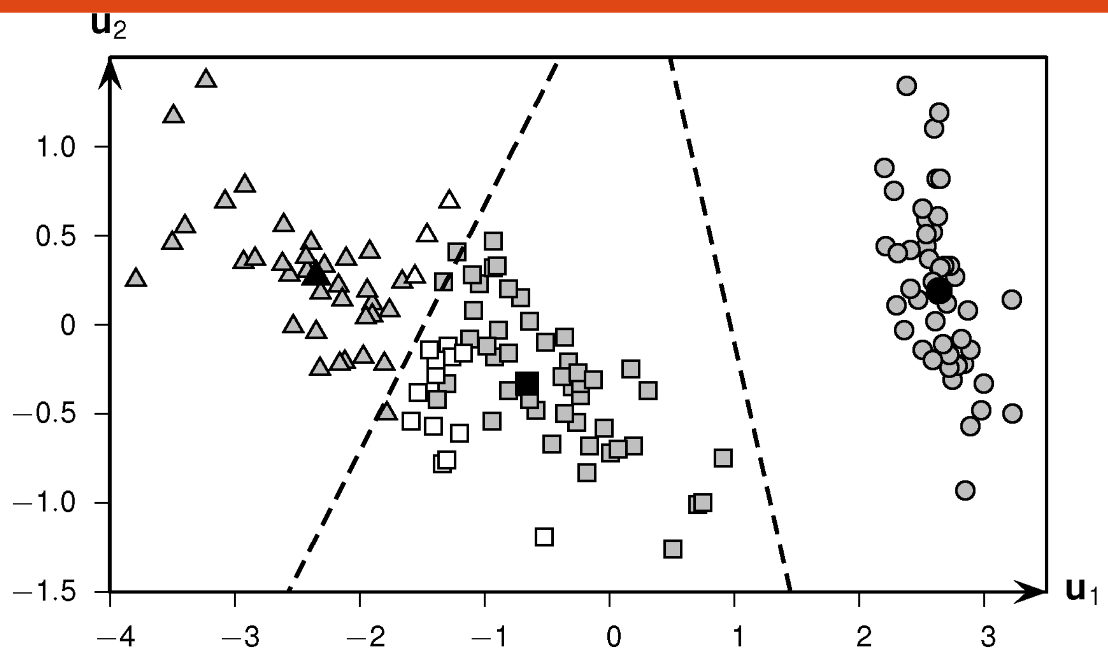
The F-measure is the harmonic mean of the precision and recall values for each cluster C_i

$$F_i = \frac{2}{\frac{1}{prec_i} + \frac{1}{recall_i}} = \frac{2 \cdot prec_i \cdot recall_i}{prec_i + recall_i} = \frac{2\ n_{ij_i}}{n_i + m_{j_i}}$$

The F-measure for the clustering $\mathcal{C}$ is the mean of clusterwise F-measure values:

$$F = \frac{1}{r}\sum_{i=1}^{r} F_i$$

K-means: Iris Principal Components Data (Good Case)

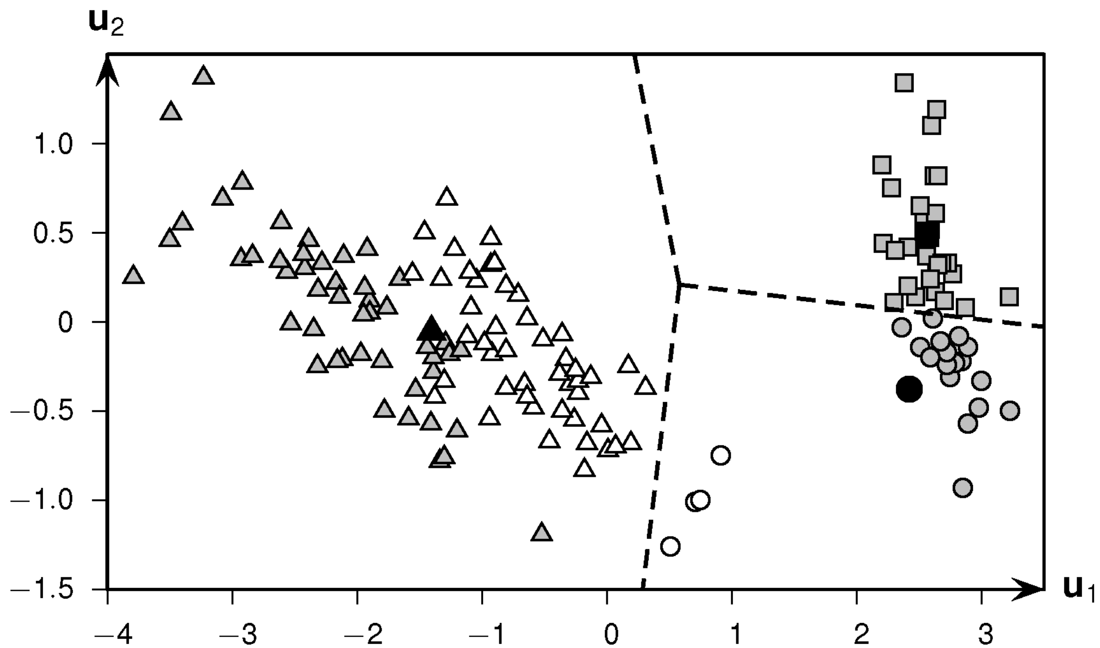


Contingency table:

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1(squares)	0	47	14	61
C_2(circles)	50	0	0	50
C_3(triangles)	0	3	36	39
m_j	50	50	50	$n = 100$

purity $= 0.887$, *match* $= 0.887$, $F = 0.885$

K-means: Iris Principal Components Data (Bad Case)



Contingency table:

	iris-setosa T_1	iris-versicolor T_2	iris-virginica T_3	n_i
C_1(*squares*)	30	0	0	30
C_2(*circles*)	20	4	0	24
C_3(*triangles*)	0	46	50	96
m_j	50	50	50	$n = 150$

$purity = 0.667$, $match = 0.560$, $F = 0.658$

Entropy-based Measures: Conditional Entropy

The entropy of a clustering $\mathcal{C}$ and partitioning $\mathcal{T}$ is given as

$$H(\mathcal{C}) = -\sum_{i=1}^{r} p_{C_i} \log p_{C_i} \qquad\qquad H(\mathcal{T}) = -\sum_{j=1}^{k} p_{T_j} \log p_{T_j}$$

where $p_{C_i} = \frac{n_i}{n}$ and $p_{T_j} = \frac{m_j}{n}$ are the probabilities of cluster C_i and partition T_j.

The cluster-specific entropy of $\mathcal{T}$, that is, the conditional entropy of $\mathcal{T}$ with respect to cluster C_i is defined as

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

Entropy-based Measures: Conditional Entropy

The conditional entropy of $\mathcal{T}$ given clustering $\mathcal{C}$ is defined as the weighted sum:

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

where $p_{ij} = \frac{n_{ij}}{n}$ is the probability that a point in cluster i also belongs to partition and where $H(\mathcal{C},\mathcal{T}) = -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} \log p_{ij}$ is the joint entropy of $\mathcal{C}$ and $\mathcal{T}$.

$H(\mathcal{T}|\mathcal{C}) = 0$ if and only if $\mathcal{T}$ is completely determined by $\mathcal{C}$, corresponding to the ideal clustering. If $\mathcal{C}$ and $\mathcal{T}$ are independent of each other, then $H(\mathcal{T}|\mathcal{C}) = H(\mathcal{T})$.

Entropy-based Measures: Normalized Mutual Information

The *mutual information* tries to quantify the amount of shared information between the clustering $\mathcal{C}$ and partitioning $\mathcal{T}$, and it is defined as

$$I(\mathcal{C}, \mathcal{T}) = \sum_{i=1}^{r} \sum_{j=1}^{k} p_{ij} \log\left(\frac{p_{ij}}{p_{C_i} \cdot p_{T_j}}\right)$$

When $\mathcal{C}$ and $\mathcal{T}$ are independent then $p_{ij} = p_{C_i} \cdot p_{T_j}$, and thus $I(\mathcal{C}, \mathcal{T}) = 0$. However, there is no upper bound on the mutual information.

The *normalized mutual information* (NMI) is defined as the geometric mean:

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

The NMI value lies in the range $[0, 1]$. Values close to 1 indicate a good clustering.

Entropy-based Measures: Variation of Information

This criterion is based on the mutual information between the clustering $\mathcal{C}$ and the ground-truth partitioning $\mathcal{T}$, and their entropy; it is defined as

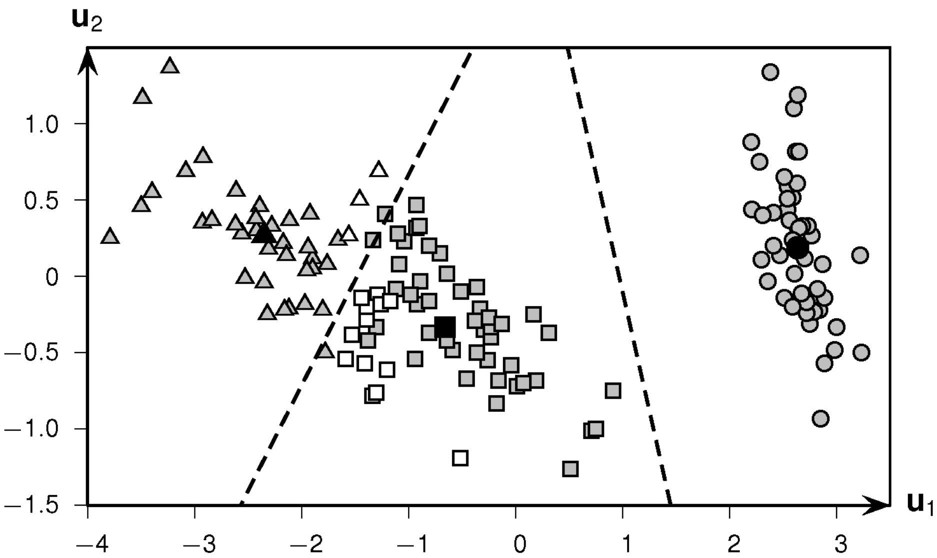
$$\begin{aligned} VI(\mathcal{C},\mathcal{T}) &= (H(\mathcal{T}) - I(\mathcal{C},\mathcal{T})) + (H(\mathcal{C}) - I(\mathcal{C},\mathcal{T})) \\ &= H(\mathcal{T}) + H(\mathcal{C}) - 2I(\mathcal{C},\mathcal{T}) \end{aligned}$$

Variation of information (VI) is zero only when $\mathcal{C}$ and $\mathcal{T}$ are identical. Thus, the lower the VI value the better the clustering $\mathcal{C}$.

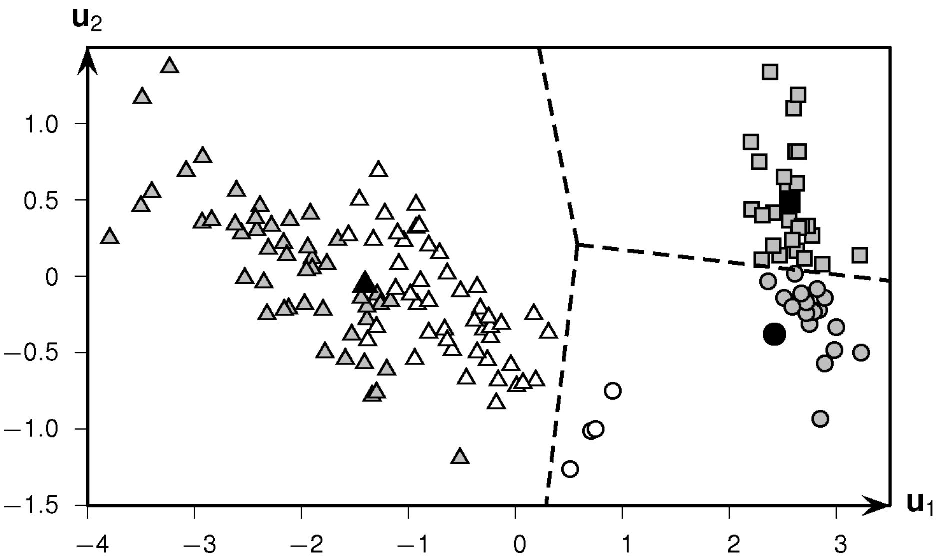
VI can also be expressed as:

$$VI(\mathcal{C},\mathcal{T}) = H(\mathcal{T}|\mathcal{C}) + H(\mathcal{C}|\mathcal{T})$$
$$VI(\mathcal{C},\mathcal{T}) = 2H(\mathcal{T},\mathcal{C}) - H(\mathcal{T}) - H(\mathcal{C})$$

K-means: Iris Principal Components Data (Good Case)



(a) K-means: good



(b) K-means: bad

	purity	*match*	*F*	$H(\mathcal{T}\|\mathcal{C})$	*NMI*	*VI*
(a) Good	0.887	0.887	0.885	0.418	0.742	0.812
(b) Bad	0.667	0.560	0.658	0.743	0.587	1.200

Pairwise Measures

Given clustering $\mathcal{C}$ and ground-truth partitioning $\mathcal{T}$, let $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{D}$ be any two points, with $i \neq j$. Let y_i denote the true partition label and let $\hat{y}_i$ denote the cluster label for point $\mathbf{x}_i$.

If both $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same cluster, that is, $\hat{y}_i = \hat{y}_j$, we call it a *positive* event, and if they do not belong to the same cluster, that is, $\hat{y}_i \neq \hat{y}_j$, we call that a *negative* event. Depending on whether there is agreement between the cluster labels and partition labels, there are four possibilities to consider:

True Positives: $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same partition in $\mathcal{T}$, and they are also in the same cluster in $\mathcal{C}$. The number of true positive pairs is given as

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

False Negatives: $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same partition in $\mathcal{T}$, but they do not belong to the same cluster in $\mathcal{C}$. The number of all false negative pairs is given as

$$FN = \left|\{(\mathbf{x}_i, \mathbf{x}_j) : \; y_i = y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}\right|$$

False Positives: $\mathbf{x}_i$ and $\mathbf{x}_j$ do not belong to the same partition in $\mathcal{T}$, but they do belong to the same cluster in $\mathcal{C}$. The number of false positive pairs is given as

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

True Negatives: $\mathbf{x}_i$ and $\mathbf{x}_j$ neither belong to the same partition in $\mathcal{T}$, nor do they belong to the same cluster in $\mathcal{C}$. The number of such true negative pairs is given as

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

Because there are $N = \binom{n}{2} = \frac{n(n-1)}{2}$ pairs of points, we have the following identity:

$$N = TP + FN + FP + TN$$

Pairwise Measures: TP, TN, FP, FN

They can be computed efficiently using the contingency table $\mathbf{N} = \{n_{ij}\}$. The number of true positives is given as

$$TP = \frac{1}{2}\left(\left(\sum_{i=1}^{r}\sum_{j=1}^{k} n_{ij}^2\right) - n\right)$$

The false negatives can be computed as

$$FN = \frac{1}{2}\left(\sum_{j=1}^{k} m_j^2 - \sum_{i=1}^{r}\sum_{j=1}^{k} n_{ij}^2\right)$$

The number of false positives are:

$$FP = \frac{1}{2}\left(\sum_{i=1}^{r} n_i^2 - \sum_{i=1}^{r}\sum_{j=1}^{k} n_{ij}^2\right)$$

Finally, the number of true negatives can be obtained via

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

Clustering Validation and Evaluation

Cluster validation and assessment encompasses three main tasks: *clustering evaluation* seeks to assess the goodness or quality of the clustering, *clustering stability* seeks to understand the sensitivity of the clustering result to various algorithmic parameters, for example, the number of clusters, and *clustering tendency* assesses the suitability of applying clustering in the first place, that is, whether the data has any inherent grouping structure.

Validity measures can be divided into three main types:

- **External:** External validation measures employ criteria that are not inherent to the dataset, e.g., class labels.
- **Internal:** Internal validation measures employ criteria that are derived from the data itself, e.g., intracluster and intercluster distances.
- **Relative:** Relative validation measures aim to directly compare different clusterings, usually those obtained via different parameter settings for the same algorithm.

Internal Measures

Internal evaluation measures do not have recourse to the ground-truth partitioning. To evaluate the quality of the clustering, internal measures therefore have to utilize notions of intracluster similarity or compactness, contrasted with notions of intercluster separation, with usually a trade-off in maximizing these two aims.

The internal measures are based on the $n \times n$ *distance matrix*, also called the *proximity matrix*, of all pairwise distances among the n points:

$$\mathbf{W} = \left\{ \delta(\mathbf{x}_i, \mathbf{x}_j) \right\}_{i,j=1}^{n}$$

where $\delta(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2$ is the Euclidean distance between $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{D}$.

The proximity matrix $\mathbf{W}$ is the adjacency matrix of the weighted complete graph G over the n points, that is, with nodes $V = \{\mathbf{x}_i \mid \mathbf{x}_i \in \mathbf{D}\}$, edges $E = \{(\mathbf{x}_i, \mathbf{x}_j) \mid \mathbf{x}_i, \mathbf{x}_j \in \mathbf{D}\}$, and edge weights $w_{ij} = \mathbf{W}(i, j)$ for all $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{D}$.

Internal Measures

The clustering $\mathcal{C}$ can be considered as a k-way cut in G. Given any subsets $S, R \subset V$, define $W(S, R)$ as the sum of the weights on all edges with one vertex in S and the other in R, given as

$$W(S,R) = \sum_{\mathbf{x}_i \in S} \sum_{\mathbf{x}_j \in R} w_{ij}$$

We denote by $\overline{S} = V - S$ the complementary set of vertices.

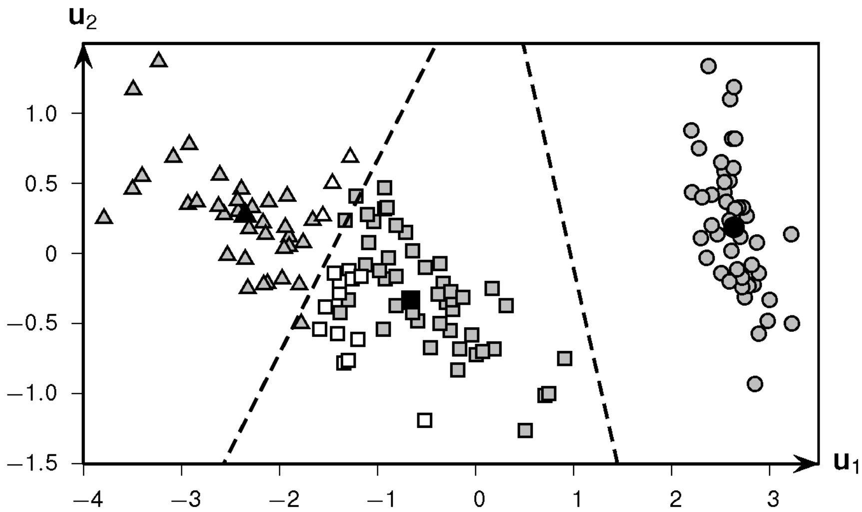
The sum of all the intracluster and intercluster weights are given as

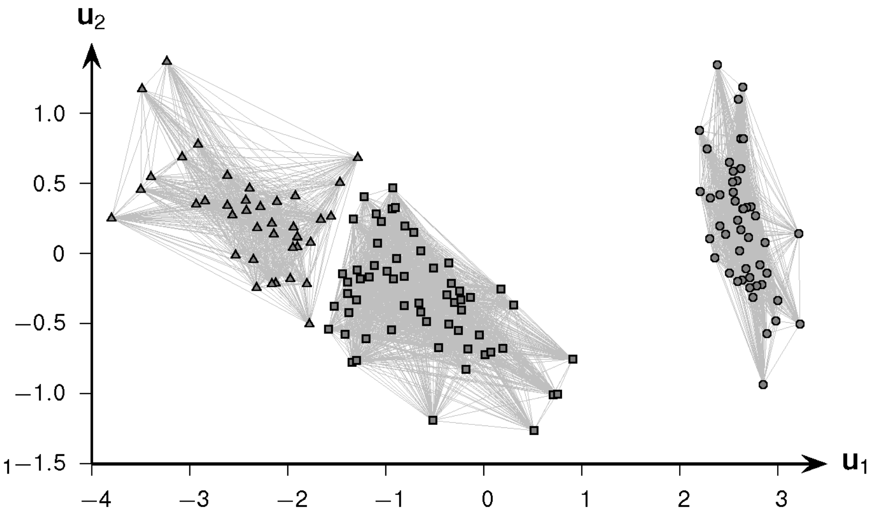
$$W_{in} = \frac{1}{2}\sum_{i=1}^{k} W(C_i, C_i) \qquad W_{out} = \frac{1}{2}\sum_{i=1}^{k} W(C_i, \overline{C_i}) = \sum_{i=1}^{k-1}\sum_{j>i} W(C_i, C_j)$$

The number of distinct intracluster and intracluster edges is given as

$$N_{in} = \sum_{i=1}^{k} \binom{n_i}{2} \qquad N_{out} = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} n_i \cdot n_j$$

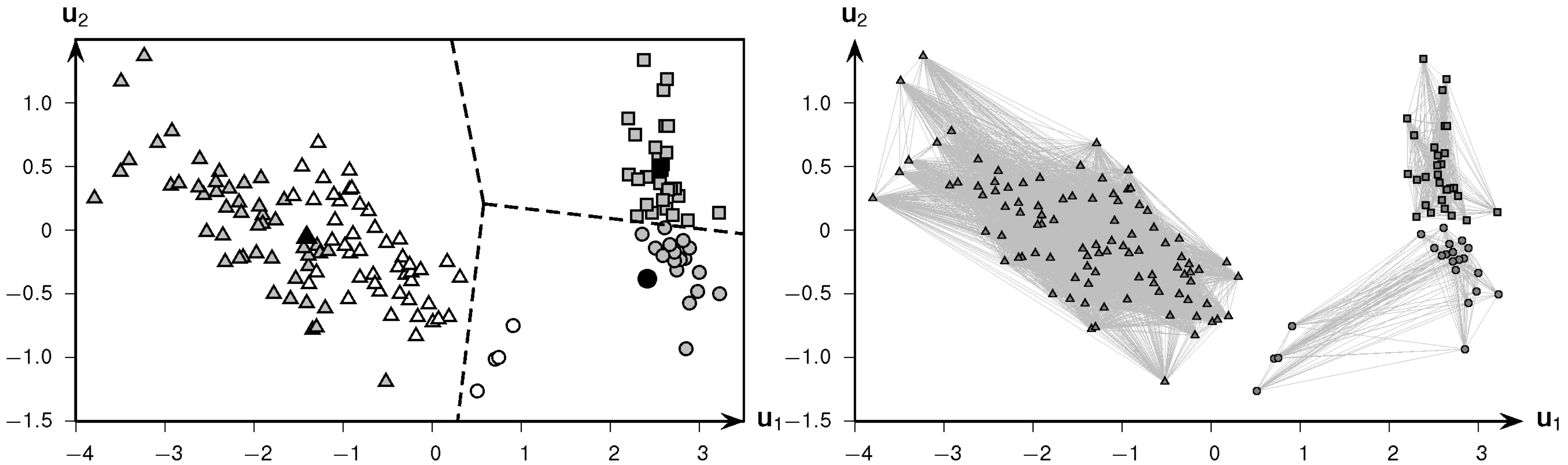
Clusterings as Graphs: Iris (Good Case)





Only intracluster edges shown.

Clusterings as Graphs: Iris (Bad Case)



Only intracluster edges shown.

Internal Measures: BetaCV and C-index

BetaCV Measure: The BetaCV measure is the ratio of the mean intracluster distance to the mean intercluster distance:

$$BetaCV = \frac{W_{in}/N_{in}}{W_{out}/N_{out}} = \frac{N_{out}}{N_{in}} \cdot \frac{W_{in}}{W_{out}} = \frac{N_{out}}{N_{in}} \frac{\sum_{i=1}^{k} W(C_i, C_i)}{\sum_{i=1}^{k} W(C_i, \overline{C_i})}$$

The smaller the BetaCV ratio, the better the clustering.

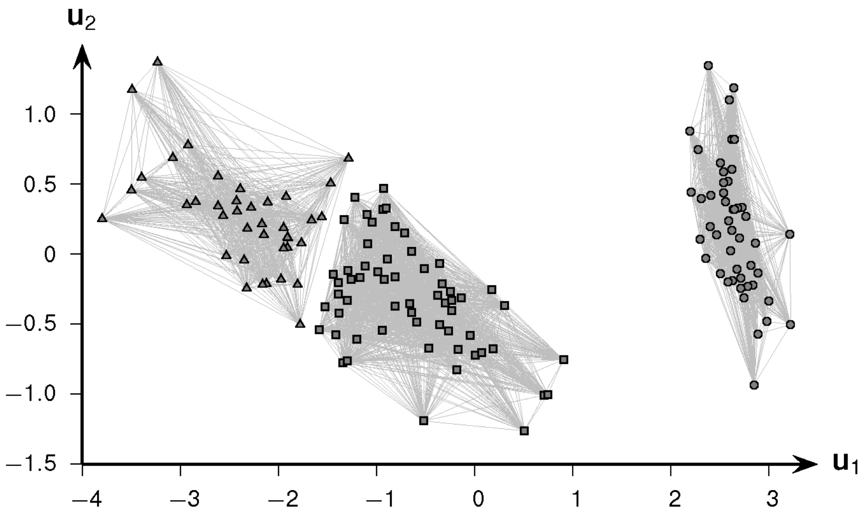
C-index: Let $W_{\min}(N_{in})$ be the sum of the smallest N_{in} distances in the proximity matrix $\mathbf{W}$, where N_{in} is the total number of intracluster edges, or point pairs. Let $W_{\max}(N_{in})$ be the sum of the largest N_{in} distances in $\mathbf{W}$.

The C-index measures to what extent the clustering puts together the N_{in} points that are the closest across the k clusters. It is defined as

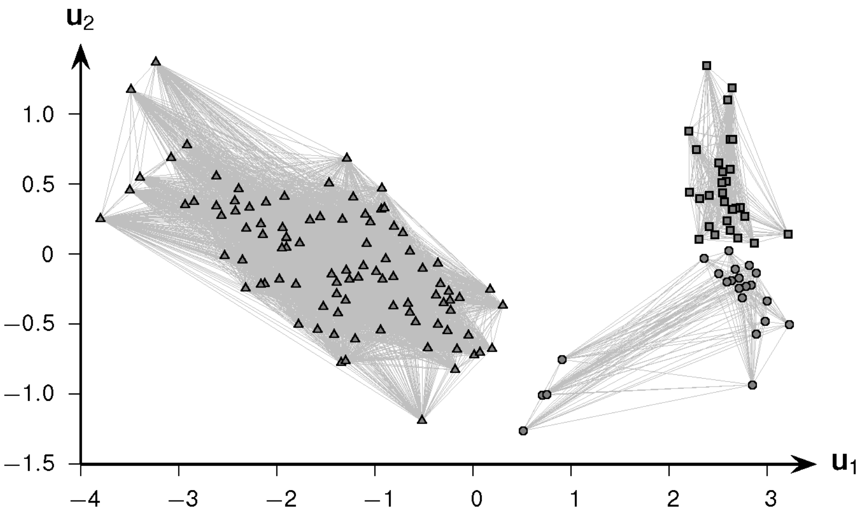
$$Cindex = \frac{W_{in} - W_{\min}(N_{in})}{W_{\max}(N_{in}) - W_{\min}(N_{in})}$$

The C-index lies in the range [0, 1]. The smaller the C-index, the better the clustering.

Iris Data: Good vs. Bad Clustering



(a) Good



(b) Bad

	BetaCV	Cindex
(a) Good	0.24	0.034
(b) Bad	0.33	0.08

Cluster Stability

The main idea behind cluster stability is that the clusterings obtained from several datasets sampled from the same underlying distribution as $\mathbf{D}$ should be similar or "stable."

Stability can be used to find a good value for k, the correct number of clusters.

We generate t samples of size n by sampling from $\mathbf{D}$ with replacement. Let $\mathcal{C}_k(\mathbf{D}_i)$ denote the clustering obtained from sample $\mathbf{D}_i$, for a given value of k.

Next, we compare the distance between all pairs of clusterings $\mathcal{C}_k(\mathbf{D}_i)$ and $\mathcal{C}_k(\mathbf{D}_j)$ using several of the external cluster evaluation measures. From these values we compute the expected pairwise distance for each value of k. Finally, the value k^* that exhibits the least deviation between the clusterings obtained from the resampled datasets is the best choice for k because it exhibits the most stability.

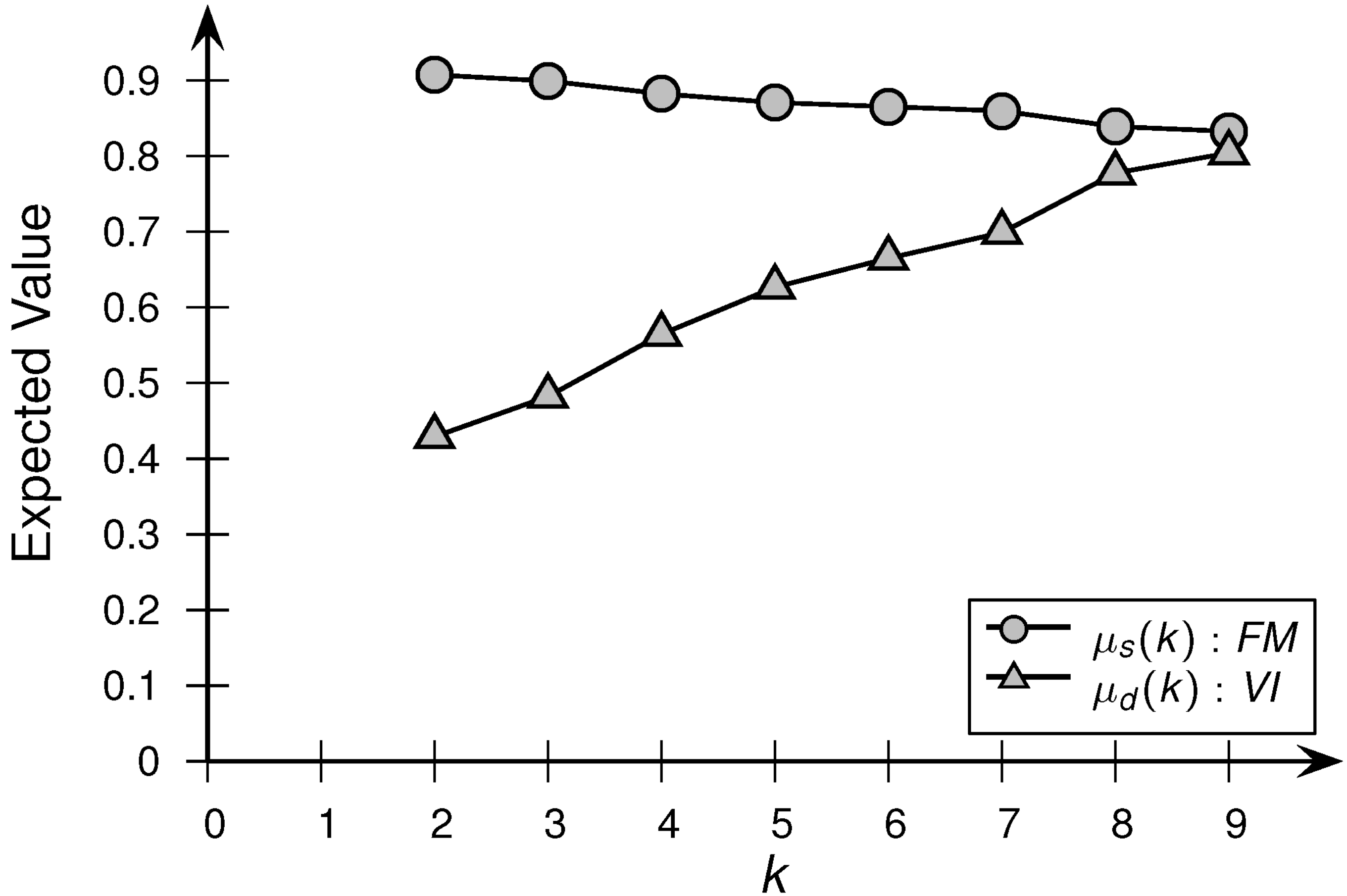
Clustering Stability Algorithm

ClusteringStability ($A, t, k^{\max}, \mathbf{D}$):

$n \leftarrow |\mathbf{D}|$
for $i = 1, 2, \ldots, t$ **do**
  $\mathbf{D}_i \leftarrow$ sample n points from $\mathbf{D}$ with replacement
for $i = 1, 2, \ldots, t$ **do**
  for $k = 2, 3, \ldots, k^{\max}$ **do**
    $\mathcal{C}_k(\mathbf{D}_i) \leftarrow$ cluster $\mathbf{D}_i$ into k clusters using algorithm A
foreach *pair* $\mathbf{D}_i, \mathbf{D}_j$ *with* $j > i$ **do**
  $\mathbf{D}_{ij} \leftarrow \mathbf{D}_i \cap \mathbf{D}_j$ `// create common dataset`
  for $k = 2, 3, \ldots, k^{\max}$ **do**
    $d_{ij}(k) \leftarrow d\big(\mathcal{C}_k(\mathbf{D}_i), \mathcal{C}_k(\mathbf{D}_j), \mathbf{D}_{ij}\big)$ `// distance between clusterings`
for $k = 2, 3, \ldots, k^{\max}$ **do**
  $\mu_d(k) \leftarrow \frac{2}{t(t-1)} \sum_{i=1}^{t} \sum_{j>i} d_{ij}(k)$
$k^* \leftarrow \arg\min_k \{\mu_d(k)\}$

Clustering Stability: Iris Data

$t = 500$ bootstrap samples; best K-means from 100 runs



The best choice is $k = 2$.