Clustering - VI

Prof. Dan A. Simovici

UMB

The quality of clusterings can be evaluated using criteria that are unrelated to the data set that is subjected to clustering (external criteria), or criteria that are derived from the data set (internal criteria).

- Typical internal criteria of clustering quality formalize the goal of attaining high intra-cluster similarity and low inter-cluster similarity. Good scores on an internal criterion do not necessarily translate into good effectiveness in an application.
- External validation criteria are useful when a "ground truth" is known (as it is typically the case for classification problems) and we seek to evaluate the appropriateness of a clustering algorithm for separating objects into clusters that conform more or less to the existing classification.

The ground truth is captured by a *reference partition* $\tau = \{T_1, \ldots, T_r\}$ of data set D (also known as the *ground-truth* partition). We discuss modalities of comparing a clustering $\kappa = (C_1, \ldots, C_m)$ with the ground truth partition.

Definition

Let τ and κ be two partitions of a data set D.
The *contingency matrix* of τ and κ is the matrix $G(\tau, \kappa) \in \mathbb{R}^{r \times m}$, where τ contains r reference blocks, κ contains m clusters, and $g_{ij} = |T_i \cap G_j|$ for $1 \leqslant i \leqslant r$ and $1 \leqslant j \leqslant m$.

Suppose that the classes of objects of a data set D relative to the partitions τ and κ are described respectively by the **R** -vectors `t` and `k` whose length is $n = |D|$.
Then, the contingency matrix $G(\tau, \kappa)$ of partitions τ and κ can be obtained by using `table(t,k)`.

Example

Let D be a data set with $|D| = 12$ and let τ and κ be two partitions of D:

$$\begin{aligned} \tau &= \{\{d_1, d_6, d_{10}\}, \{d_3, d_4, d_7, d_8, d_{12}\}, \{d_2, d_5, d_9, d_{11}\}\}, \\ \kappa &= \{\{d_4, d_6, d_{10}, d_{12}\}, \{d_1, d_3, d_8\}, \{d_2, d_5, d_7, d_9, d_{11}\}\}. \end{aligned}$$

The **R** -vectors that describe these partitions are:

```
t <- c(1,3,2,2,3,1,2,2,3,1,3,2)
k <- c(2,3,2,1,3,1,3,2,3,1,3,1).
```

A call to the function `table` returns the contingency table of partitions:

```
> table(t,k)
   k
t   1 2 3
  1 2 1 0
  2 2 2 1
  3 0 0 4
```

A cluster C_i is *τ-pure* if it is included in a block T_j of the reference partition τ.
We denote by T_{i_j} the largest block of the reference partition τ that has the largest intersection with the cluster C_j.
Next we introduce the notions of precision and recall for a cluster.

Definition

The *precision* of a cluster C_j is defined as

$$\text{precision}_\tau(C_j) = \frac{1}{|C_j|} \cdot \max\{|T_i \cap C_j| \mid 1 \leqslant j \leqslant r\},$$

and it measures the largest fraction of the cluster in a block of the reference partition.

Definition

The *precision* of clustering κ is the average precision of the clusters $C_1, \ldots, C_m$, that is,

$$\begin{aligned} \text{precision}_\tau(\kappa) &= \sum_{j=1}^{m} \frac{|C_j|}{|D|} \text{precision}_\tau(C_j) \\ &= \frac{1}{|D|} \sum_{j=1}^{m} \max\{|T_i \cap C_j| \mid 1 \leqslant i \leqslant r\}. \end{aligned}$$

If all clusters of κ are pure, then $\text{precision}_\tau(\kappa) = 1$.

Definition

The *recall* of cluster C_j is defined as

$$\text{recall}_\tau(C_j) = \frac{1}{|T_{i_j}|}|T_{i_j} \cap C_j|$$

and measures the fraction of the largest reference block that has the largest intersection with C_j which is shared with C_i.
The *F-measure* of cluster C_j is the harmonic average of its precision and recall:

$$F(C_j) = \frac{2}{\frac{1}{\text{precision}_\tau(C_j)} + \frac{1}{\text{recall}_\tau(C_j)}} = 2\frac{n_{i_j j}}{n_j + |T_{i_j}|}$$

The F-measure $F(\kappa)$ for the clustering κ is the mean of the F-measures for the clusters:

$$F(\kappa) = \frac{1}{m} \sum_{j=1}^{m} F(C_j).$$

Higher values for the F-measure indicate a better fit between the reference partition τ and the clustering κ.

Example

Let τ and κ be the partitions introduced on Slide 7, where $\tau = \{T_1, T_2, T_3\}$, $\kappa = \{C_1, C_2, C_3\}$ and

$$\begin{aligned} T_1 &= \{d_1, d_6, d_{10}\} & C_1 &= \{d_4, d_6, d_{10}, d_{12}\}, \\ T_2 &= \{d_3, d_4, d_7, d_8, d_{12}\} & C_2 &= \{d_1, d_3, d_8\}, \\ T_3 &= \{d_2, d_5, d_9, d_{11}\} & C_3 &= \{d_2, d_5, d_7, d_9, d_{11}\}. \end{aligned}$$

Note that contingency matrix $G(\tau, \sigma)$ can be written as

$$G = \begin{pmatrix} |T_1 \cap C_1| & |T_1 \cap C_2| & |T_1 \cap C_3| \\ |T_2 \cap C_1| & |T_2 \cap C_2| & |T_2 \cap C_3| \\ |T_3 \cap C_1| & |T_3 \cap C_2| & |T_3 \cap C_3| \end{pmatrix} = \begin{pmatrix} 2 & 1 & 0 \\ 2 & 2 & 1 \\ 0 & 0 & 4 \end{pmatrix}$$

Thus, the blocks of the reference partitions that have the largest intersection with the clusters C_1, C_2 and C_3 are T_2, again T_2 and T_3, respectively.

The precision of the clusters of κ relative to τ are

$$\text{precision}_\tau(C_1) = \frac{2}{4}, \text{precision}_\tau(C_2) = \frac{2}{3}, \text{precision}_\tau(C_3) = \frac{4}{5},$$

so C_3 has the largest precision.
The precision of κ is

$$\begin{aligned} \text{precision}_\tau(\kappa) &= \sum_{j=1}^{m} \frac{|C_j|}{|D|} \text{precision}_\tau(C_j) \\ &= \frac{4}{12}\frac{2}{4} + \frac{3}{12}\frac{2}{3} + \frac{5}{12}\frac{4}{5} = \frac{2}{3}. \end{aligned}$$

The recalls of the clusters are

$$\text{recall}_\tau(C_1) = \frac{2}{5}, \text{recall}_\tau(C_2) = \frac{2}{5}, \text{recall}_\tau(C_3) = \frac{4}{4}.$$

The F-score of C_1 is

$$F(C_1) = \frac{2\text{precision}(C_1) \cdot \text{recall}(C_1)}{\text{precision}(C_1) + \text{recall}(C_1)} = \frac{4}{9}.$$

Similarly, $F(C_2) = \frac{1}{2}$ and $F(C_3) = \frac{8}{9}$. The F-score for the cluster κ is the average of these scores, that is, $\frac{1}{3}(\frac{4}{9} + \frac{1}{2} + \frac{8}{9}) = \frac{11}{18}$.
A good score is usually close to 1.

For a partition $\pi \in \mathrm{PART}(D)$ we write $x \equiv_\pi y$ if there is a block $B \in \pi$ such that $\{x, y\} \subseteq B$. It is immediate that "$\equiv_\pi$ is an equivalence relation.

Let $\tau = \{T_1, \ldots, T_r\}$ be a *reference partition* of data set D (also known as the *ground-truth* partition) and let $\kappa = (C_1, \ldots, C_m)$ be a clustering. The pairs of elements of D can be classified into four classes relative to the partitions τ and σ. Namely, a pair (x, y) with $x \neq y$ is

i. a true positive pair if $x \equiv_\tau y$ and $x \equiv_\kappa y$;
ii. a true negative pair if $x \not\equiv_\tau y$ and $x \not\equiv_\kappa y$;
iii. a false positive pair if $x \not\equiv_\tau y$ and $x \equiv_\kappa y$;
iv. a false negative pair if $x \equiv_\tau y$ and $x \not\equiv_\kappa y$.

The number of true positive pairs is denoted by $\mathtt{TP}(\tau, \kappa)$, that of true negative pairs is $\mathtt{TN}(\tau, \kappa)$, the number of false positive pairs is $\mathtt{FP}(\tau, \kappa)$, and the number of false negative pairs is $\mathtt{FN}(\tau, \kappa)$. All there values can be computed in $O(rm)$ time.

For $|D| = n$ there are $\binom{n}{2}$ distinct pairs, hence

$$\binom{n}{2} = \mathtt{TP}(\tau, \kappa) + \mathtt{TN}(\tau, \kappa) + \mathtt{FP}(\tau, \kappa) + \mathtt{FN}(\tau, \kappa).$$

Let $G(\tau, \kappa) = (g_{ij}) \in \mathbb{R}^{r \times m}$ be the contingency matrix for the reference partition $\tau = \{T_1, \ldots, T_r\}$ and clustering $\kappa = \{C_1, \ldots, C_m\}$. We introduce the partial sums:

$$\begin{aligned} g_{i\cdot} &= \sum_{j=1}^{m} g_{ij} = |T_i|, \\ g_{\cdot j} &= \sum_{i=1}^{r} g_{ij} = |C_j|, \\ g_{\cdot\cdot} &= \sum_{i=1}^{r} \sum_{j=1}^{m} g_{ij} = |D|, \end{aligned}$$

for $1 \leqslant i \leqslant r$ and $1 \leqslant j \leqslant m$.

These notations are summarized by the following table:

		part. κ				
	class	C_1	C_2	$\cdots$	C_m	sums
part. τ	T_1	g_{11}	g_{12}	$\cdots$	g_{1m}	$g_{1\cdot}$
	T_2	g_{21}	g_{22}	$\cdots$	g_{2m}	$g_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
	T_r	g_{r1}	g_{r2}	$\cdots$	g_{rm}	$g_{r\cdot}$
	sums	$g_{\cdot 1}$	$g_{\cdot 2}$	$\cdots$	$g_{\cdot m}$	$g_{\cdot\cdot} = \lvert D \rvert$

Since the contingency matrix $G(\tau, \kappa)$ can be computed in linear time it is possible to compute efficiently the measures introduced above.

We have

$$\begin{aligned}
\mathtt{TP}(\tau,\kappa) &= \sum_{i=1}^{r}\sum_{j=1}^{m}\binom{g_{ij}}{2} \\
&= \frac{1}{2}\left(\sum_{i=1}^{r}\sum_{j=1}^{m} g_{ij}^2 - \sum_{i=1}^{r}\sum_{j=1}^{m} g_{ij}\right) \\
&= \frac{1}{2}\left(\sum_{i=1}^{r}\sum_{j=1}^{m} g_{ij}^2 - n\right).
\end{aligned}$$

The number of pairs that belong to the same block of the reference partition is $\sum_{i=1}^{r}\binom{g_{i\cdot}}{2}$. If we eliminate from these pairs the true positive pairs we obtain the number of false negative pairs:

$$\begin{aligned}
\mathtt{FN}(\tau,\kappa) &= \sum_{i=1}^{r}\binom{g_{i\cdot}}{2} - \mathtt{TP}(\tau,\kappa) \\
&= \frac{1}{2}\sum_{i=1}^{r} g_{i\cdot}^2 - \frac{1}{2}\sum_{i=1}^{r} g_{i\cdot} - \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{m} g_{ij}^2 + \frac{n}{2}
\end{aligned}$$

The number of false positive pairs is obtained by subtracting from the number of pairs that belong to the same cluster the number of true positive pairs:

$$\begin{aligned}
\mathrm{FP}(\tau, \kappa) &= \sum_{j=1}^{m} \binom{g_{\cdot j}}{2} - \mathrm{TP}(\tau, \kappa) \\
&= \frac{1}{2} \sum_{j=1}^{m} g_{\cdot j}^2 - \frac{1}{2} \sum_{j=1}^{m} g_{\cdot j} - \frac{1}{2} \sum_{i=1}^{r} \sum_{j=1}^{m} g_{ij}^2 + \frac{n}{2} \\
&= \frac{1}{2} \left(\sum_{j=1}^{m} g_{\cdot j}^2 - \sum_{i=1}^{r} \sum_{j=1}^{m} g_{ij}^2 \right)
\end{aligned}$$

because $\sum_{j=1}^{m} g_{\cdot j} = n$.

The number of true negative pairs is

$$\mathtt{TN}(\tau, \kappa) \quad = \quad \frac{1}{2}\left(g_{\cdot\cdot}^2 - \sum_{i=1}^{r} g_{i\cdot}^2 - \sum_{j=1}^{m} g_{\cdot j}^2 + \sum_{i=1}^{r}\sum_{j=1}^{m} g_{ij}^2 \right) .$$

These numbers can be used, in turn, to compute efficiently several numerical characteristics of the pair (τ, κ).
Let ρ_τ and ρ_κ the equivalences that correspond to the partitions τ and κ. These equivalences are sets of pairs in $D \times D$. Therefore, it makes sense to consider their Jaccard coefficient:

$$J(\rho_\tau, \rho_\kappa) = \frac{|\rho_\tau \cap \rho_\kappa|}{|\rho_\tau \cup \rho_\kappa|},$$

which evaluates the similarity between the reference partition τ and the clustering κ. It is clear that

$$J(\rho_\tau, \rho_\kappa) = \frac{|\mathtt{TP}(\tau, \kappa)|}{|\mathtt{TP}(\tau, \kappa)| + |\mathtt{FN}(\tau, \kappa)| + |\mathtt{FP}(\tau, \kappa)|}.$$

The *Rand coefficient* is

$$R(\tau, \kappa) = \frac{|\mathtt{TP}(\tau, \kappa)| + |\mathtt{TN}(\tau, \kappa)|}{\binom{n}{2}},$$

and represents the fraction of objects where the reference partition and the clustering agree. When $R(\tau, \kappa) = 1$ the two partitions are identical.

The notions of precision and recall previously introduced are reformulated for pairs of objects.

The *precision* for τ and κ is

$$\text{precision}(\tau, \kappa) = \frac{\mathtt{TP}(\tau, \kappa)}{\mathtt{TP}(\tau, \kappa) + \mathtt{FP}(\tau, \kappa)}$$

and reflects the size of the set of correctly classified pairs of objects vs. the size of the sets of pairs of objects that reside in the same cluster. We have $\text{precision}(\tau, \kappa) = 1$ if and only if no false positive pairs.

The *recall* for τ and κ is

$$\text{recall}(\tau, \kappa) = \frac{\mathtt{TP}(\tau, \kappa)}{\mathtt{TP}(\tau, \kappa) + \mathtt{FN}(\tau, \kappa)}$$

Recall evaluates the fraction of correctly classiffied pairs of objects compared to all pairs of objects that inhabit the same block of reference partition.

We have $\text{recall}(\tau, \kappa) = 1$ if $\mathtt{FN}(\tau, \kappa) = 0$, that is, if there are no pairs in ρ_τ whose components belong to two distinct clusters.

The *Fowlkes-Mallows* coefficient $\mathtt{FM}(\tau, \kappa)$ is the geometric average of recall and precision, that is,

$$\begin{aligned} \mathtt{FM}(\tau, \kappa) &= \sqrt{\mathsf{precision}(\tau, \kappa) \cdot \mathsf{recall}(\tau, \kappa)} \\ &= \frac{\mathtt{TP}(\tau, \kappa)}{\sqrt{(\mathtt{TP}(\tau, \kappa) + \mathtt{FP}(\tau, \kappa))(\mathtt{TP}(\tau, \kappa) + \mathtt{FN}(\tau, \kappa))}}. \end{aligned}$$

Definition

Let (S, d) be a metric space. A *dispersion measure* on (S, d) is a function $s : \mathcal{P}(S) \longrightarrow \mathbb{R}_{\geqslant 0}$ such that $s(C) = 0$ if and only if $|C| = 1$.

Example

The function sse is a dispersion measure.

Example

The function $\delta : \mathcal{P}(S) \longrightarrow \mathbb{R}_{\geqslant 0}$ defined by

$$\delta(C) = \frac{\sum\{d(x,y) \mid x,y \in C, x \neq y\}}{|C|(|C|-1)}$$

yields the mean distance between all pairs of objects in C. It is immediate to see that $\delta(C) = 0$ if and only if $|C| = 1$, so δ is a dispersion measure.

Example

The diameter $diam : \mathcal{P}(S) \longrightarrow \mathbb{R}_{\geqslant 0}$ is a dispersion function for obvious reasons.

Definition

For a clustering $\kappa = \{C_1, \ldots, C_k\}$ let

- s_i be the dispersion of C_i,
- m_{ij} be the distance between the representatives c_i and c_j of the clusters C_i and C_j (usually chosen as the centroids of the clusters C_i and C_j) for $1 \leqslant i, j \leqslant k$.

A *cluster similarity measure* $r : \mathbb{R}^3_{\geqslant 0} \longrightarrow \hat{\mathbb{R}}$ satisfies the following conditions:

1. $r(s_i, s_j, m_{ij}) \geqslant 0$;
2. $r(s_i, s_j, m_{ij}) = r(s_j, s_i, m_{ij})$;
3. $r(s_i, s_j, m_{ij}) = 0$ if and only if $s_i = s_j$;
4. if $s_j = s_k$ and $m_{ij} < m_{ik}$, then $r(s_i, s_j, m_{ij}) > r(s_i, s_k, m_{ik})$;
5. if $m_{ik} = m_{ij}$ and $s_j > s_k$, then $r(s_i, s_j, m_{ij}) > r(s_i, s_k, m_{ik})$.

The previous definition means that:

- when the distance between cluster centers increases while their dispersions remain constant, the similarity of the clusters decreases;
- if the distances between cluster centroids remains constant while the dispersion increase, the similarity increases.

Example

Consider the function r given by

$$r(s, s', m) = \frac{s + s'}{m}$$

for $s, s', m \in \mathbb{R}_{\geqslant 0}$. It is immediate that r satifies the conditions imposed on similarity measures.

Definition

Let $\kappa = \{C_1, \ldots, C_k\}$ be a clustering in a metric space (S, d). The *Davies-Bouldin index* of κ is the clustering average similarity measure r_κ given by

$$r_\kappa = \frac{1}{k} \sum_{i=1}^{k} \max\{r_{ij} \mid 1 \leqslant j \leqslant k\}.$$

The "best" clustering is the one that minimizes the average similarity measure.

Example

Consider a data set in $\mathbb{R}^2$ that consists of four points,

$$\mathbf{v}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \mathbf{v}_3 = \begin{pmatrix} 8 \\ 1 \end{pmatrix}, \mathbf{v}_4 = \begin{pmatrix} 8 \\ 3 \end{pmatrix}$$

grouped into two clusterings:

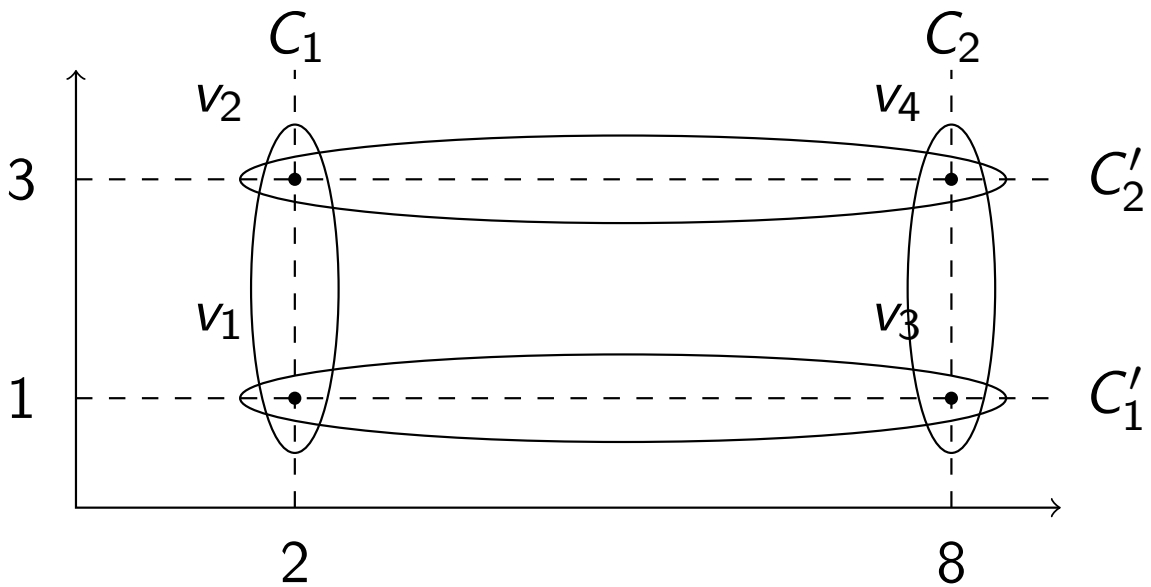
$$\kappa = \{C_1, C_2\}, \kappa' = \{C_1', C_2\},$$

where

$$C_1 = \{\mathbf{v}_1, \mathbf{v}_2\}, C_2 = \{\mathbf{v}_3, \mathbf{v}_4\},$$

and

$$C_1' = \{\{\mathbf{v}_1, \mathbf{v}_3\}, C_2' = \{\mathbf{v}_2, \mathbf{v}_4\}.$$


C_1
C_2
v_2
v_4
3
C_2'
v_1
v_3
1
C_1'
2
8

The centroids of the clusters are:

cluster	C_1	C_2	C_1'	C_2'
centroid	$\begin{pmatrix}2\\2\end{pmatrix}$	$\begin{pmatrix}8\\2\end{pmatrix}$	$\begin{pmatrix}5\\1\end{pmatrix}$	$\begin{pmatrix}5\\3\end{pmatrix}$

We choose the dispersion measure as the sum of the square errors.

We choose the dispersion measure as the sum of the square errors. Its values for the clusters shown in Slide 34 are

$$\text{sse}(C_1) = 2, \text{sse}(C_2) = 2, \text{sse}(C_1') = 18, \text{sse}(C_2') = 18.$$

Thus, $r_{12} = 0.8$ and $r_{12}' = 18$, hence $r_\kappa = 0.8$ and $r_{\kappa'} = 18$, giving the edge to κ.

A related family of cluster quality indices is known as Dunn quality indices. For a clustering $\kappa = \{C_1, \ldots, C_k\}$ a Dunn index is a function

$$\Delta(\kappa) = \frac{\min_{1 \leqslant i < j \leqslant k} D(C_i, C_j)}{\max_{1 \leqslant j \leqslant k} s(C_j)},$$

where s is a dispersion measure, and $D(C_i, C_j)$ is an intercluster dissimilarity (which can be the least distance between two points in different clusters, the maximum distance between two such points, or the distance between the centroids of the clusters, etc.). Note that if a cluster has a high value of the dispersion this impacts negatively the value of the index due to the presence of max in the denominator.

The Silhouette Coefficient

Let $\kappa = \{C_1, \ldots, C_k\}$ be a clustering on a dissimilarity space (S, d), where $k > 1$. The *silhouette coefficient* of an object compares the similarity between an object and other objects located in the same cluster, and the similarity of the same object to objects located in other clusters.
Suppose that $x \in S$ is assigned to the cluster C_p and $\{x\} \subset C_p$. Define

$$a(x) = \frac{1}{|C_p|} \sum \{d(x, u) \mid u \in C_p - \{x\}\}.$$

For $r \neq p$ define $d(x, C_r) = \frac{1}{|C_r|} \sum \{d(x, y) \mid y \in C_r\}$ and

$$b(x) = \min\{d(x, C_r) \mid 1 \leqslant r \leqslant k \text{ and } r \neq p\}.$$

The cluster C_r that defines $b(x)$, that is, $b(x) = d(x, C_r)$ is the *neighbour* of x and represents the second-best choice for object x.

Definition

The *silhouette* of x is the number $s(x)$ defined as

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}} = \begin{cases} 1 - \frac{a(x)}{b(x)} & \text{if } a(x) < b(x), \\ 0 & \text{if } a(x) = b(x), \\ \frac{b(x)}{a(x)} - 1 & \text{if } a(x) > b(x). \end{cases}$$

If $C_p = \{x\}$ we define $s(x) = 0$.

- Note that $-1 \leqslant s(x) \leqslant 1$. When $s(x)$ is close to 1, the within dissimilarity $a(x)$ is much smaller than the smallest between dissimilarity $b(x)$. Therefore, x is well-classified; the second best-choice of a cluster for x is not nearly as closes as the actual choice.
- When $a(x)$ is close to 0, then $a(x)$ and $b(x)$ are about the same, hence it not clear whether x has been correctly assigned to C_p.
- When $a(x)$ is close to -1, then $a(x)$ is larger than $b(x)$, so x is closer to some cluster other than C_p; we say that x has been missassigned.

```
ir <- iris[,1:4]
```

Next, we apply the `pam` algorithm of the package `clust`:

```
pamc <- pam(ir,3)
```

The plot of the `pamc` object contains two subplots: the `clusplot`, which we discussed previously and the `sihouette` plot. These plots can be obtained by writing

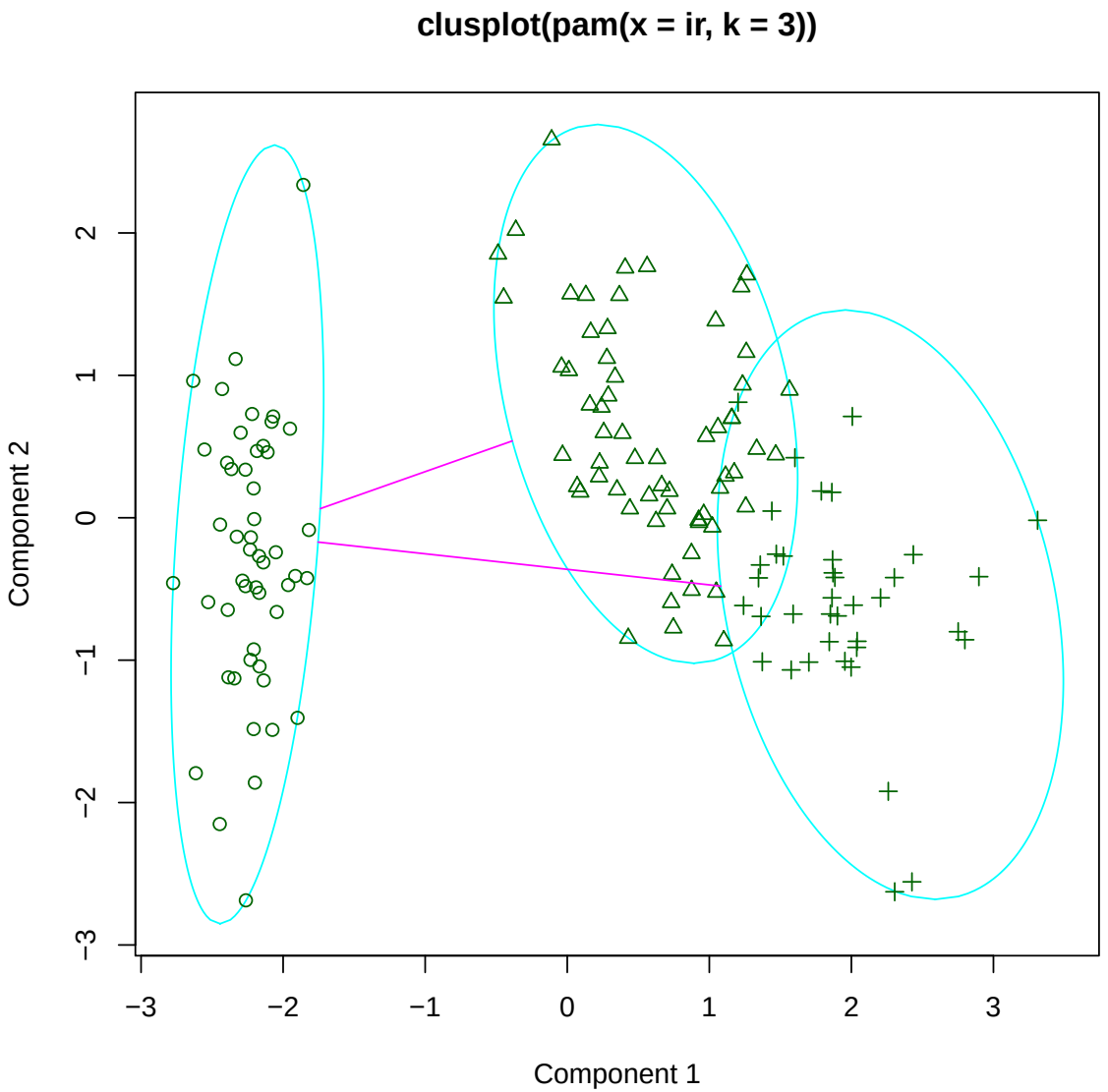
```
> pdf("pamc-clusplot.pdf")
> plot(pamc,which.plots=1)
> dev.off()
```

and

```
> pdf("pamc-silh.pdf")
> plot(pamc,which.plots=2)
> dev.off()
```

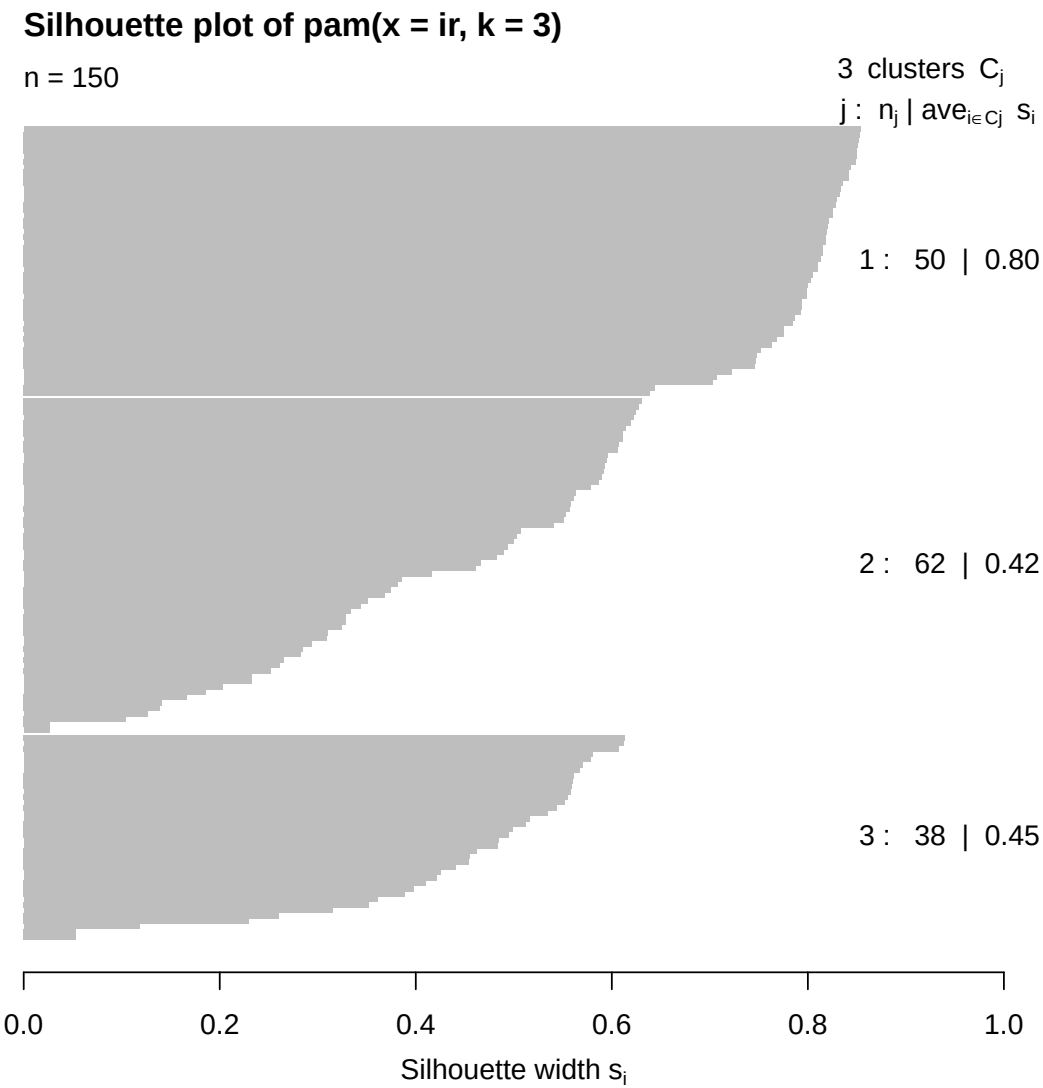
The plot which is generated is determined by the parameter `which.plots` (1 for `clusplot` and 2 for the `silhouette` plot.

The clusplot graph



These two components explain 95.81 % of the point variability.

The the silhouette plot



Example

The `silhouette` function can be used to determine the best number of clusters. Consider, the following example provided in the **R** documentation of the txpam function for a uni-dimensional set of objects defined by

```
x <- c(rnorm(50),rnorm(50,mean=5),rnorm(50,mean=15))
```

and define an array `w` as

```
w <- numeric(20)
```

The following fragment of **R** code generates a representation of the average silhouette for a various numbers of clusters between 2 and 20:

```
x <- c(rnorm(50),rnorm(50,mean=5),rnorm(30,mean=15))
w <- numeric(20)
for(k in 2:20)
   w[k] <- pam(x,k)$silinfo$avg.width
k.best <- which.max(w)

cat("silhouette-optimal number of clusters is: ",k.best,"\n")

plot(1:20,w,type="h",main="pam() clustering assessment",
    xlab="k (no of clusters)",ylab="avg. silhouette width")

axis(1,k.best,paste("best",k.best,sep="\n"),col="red",col.axis="red
```

The best value is $k = 3$, as it also follows from the next graph.