External validity indices for individual clusters

Matthijs J Warrens
and
Hanneke van der Hoef

Dutch-Flemish Classification Society
Leiden May 19, 2017

Cluster validation

Goal?

- understand characteristics of clustering methods
- what type (shape) of clusters are typically found?

Why?

- there are hundreds of different clustering methods
- choose the 'best' clustering method for a particular data set

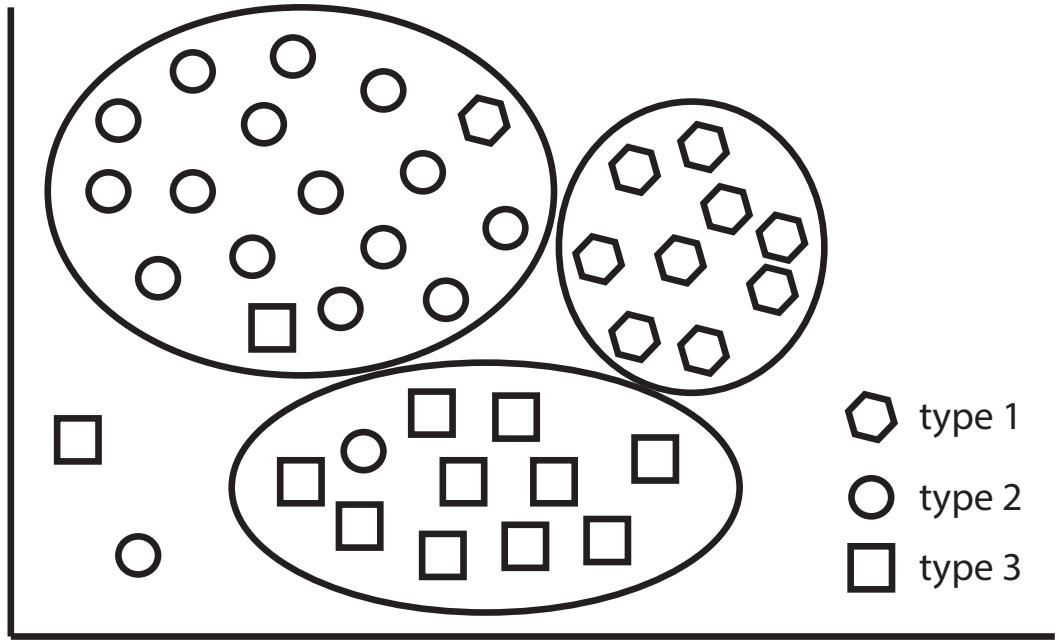
How? compare

- reference standard partition (with 'true' structure of objects) and trial partition found with a clustering method

Reference and trial partition

How?

- reference standard partition (three types of objects)
- trial partition found with a clustering method (ellipses)
- summarize agreement in matching table



Notation

- reference standard partition $\mathcal{C} = \{C_1, C_2, \ldots, C_I\}$
- trial partition $\mathcal{B} = \{B_1, B_2, \ldots, B_J\}$
- matching table $\mathbf{M} = \{m_{ij}\}$ of size $I \times J$ where m_{ij} is number of objects in C_i (reference) and in B_j (trial)

example of **M**:

reference	trial partition			
partition	B_1	B_2	B_3	Totals
C_1	102	0	0	102
C_2	0	15	10	25
C_3	0	10	15	25
Totals	102	25	25	152

indices			
R	.95	AR	.90

External validity indices

Summarize matching table

- information in matching table may be complex
- convenient to summarize information with numbers

Three approaches

- counting pairs of objects
- information theory (mutual information, entropy)
- matching sets (sensitivity, precision)

Counting object pairs

- Rand (1971) index (R)
- adjusted Rand index (AR; Hubert & Arabie 1985)

More notation: counting pairs

- # objects: n
- # of object pairs: $N = \binom{n}{2} = \frac{n(n-1)}{2}$
- # object pairs in same cluster in both partitions:

$$T = \sum_{i=1}^{I} \sum_{j=1}^{J} \binom{m_{ij}}{2}$$

- # object pairs in same cluster in reference/trial partition:

$$P = \sum_{i=1}^{I} \binom{m_{i+}}{2} \qquad \text{and} \qquad Q = \sum_{j=1}^{J} \binom{m_{+j}}{2}$$

Commonly used validity indices

- Rand index: $$R = \frac{N + 2T - P - Q}{N}$$
- adjusted Rand index: $AR = \frac{2(NT - PQ)}{N(P + Q) - 2PQ}$
- asymmetric Wallace indices: $W = \frac{T}{P}$ and $V = \frac{T}{Q}$

  $W = T/P$ is proportion of correct object pairs that are joined in the trial partition

  $V = T/Q$ is proportion of object pairs that are correctly joined in the trial partition

Problems

Commonly used indices are overall indices

- provide a general notion of what is going on
- little to no information on specific clusters

Available alternatives for individual clusters

- sensitivity (recall, classification rate)
- precision (positive predictive value)

Problem with alternatives

- require arbitrary matching of clusters (Steinley 2004)
- group assignment can be manipulated to get better sensitivity
- partitions can be compared only if same number of clusters

New measures for sensitivity

The proportion of object pairs in C_i joined in the trial partition is

$$w_i = \sum_{j=1}^{J} \binom{m_{ij}}{2} \Big/ \binom{m_{i+}}{2}$$

An adjusted version is

$$Aw_i = \frac{N \sum_{j=1}^{J} \binom{m_{ij}}{2} - \binom{m_{i+}}{2} Q}{\binom{m_{i+}}{2} (N - Q)}$$

Both value 1 if all objects in C_i in same cluster of trial partition

New measures for individual clusters

The proportion of object pairs correctly joined in cluster B_j is

$$v_j = \sum_{i=1}^{I} \binom{m_{ij}}{2} \Big/ \binom{m_{+j}}{2}$$

An adjusted version is

$$Av_j = \frac{N \sum_{i=1}^{I} \binom{m_{ij}}{2} - \binom{m_{+j}}{2} P}{\binom{m_{+j}}{2} (N - P)}$$

Both not perfect measures of precision (requires matching)

Relationship to overall indices

Overall index W (Wallace 1983) is weighted average of the w_i's:

$$W = \frac{\binom{m_{1+}}{2} w_1 + \binom{m_{2+}}{2} w_2 + \binom{m_{3+}}{2} w_3}{\binom{m_{1+}}{2} + \binom{m_{2+}}{2} + \binom{m_{3+}}{2}}$$

Weighted averages

- decomposition into blocks of cluster information
- overall indices V and AR have a similar decomposition
- large clusters contribute more to the overall value
- weight increases quadratically with cluster size

Example

reference partition	trial partition			
	B_1	B_2	B_3	Totals
C_1	102	0	0	102
C_2	0	15	10	25
C_3	0	10	15	25
Totals	102	25	25	152

overall indices high ($\geq .90$)
recovery satisfactory?

indices			
R	.95	AR	.90
W	.95	AW	.90
V	.95	AV	.90
w_1	1.0	Aw_1	1.0
w_2	.50	Aw_2	.00
w_3	.50	Aw_3	.00
v_1	1.0	Av_1	1.0
v_2	.50	Av_2	.00
v_3	.50	Av_3	.00

large cluster perfectly recovered ($w_1 = Aw_1 = 1.0$)
two smaller clusters not well recovered

Another example

reference partition	trial partition			
	B_1	B_2	B_3	Totals
C_1	52	48	0	100
C_2	46	54	0	100
C_3	0	0	10	10
Totals	98	102	10	210

overall indices low ($\approx .50, .08$)
recovery not satisfactory?

indices			
R	.55	AR	.08
W	.50	AW	.08
V	.50	AV	.08
w_1	.50	Aw_1	.08
w_2	.50	Aw_2	.08
w_3	1.0	Aw_3	1.0
v_1	.50	Av_1	.08
v_2	.50	Av_2	.08
v_3	1.0	Av_3	1.0

small cluster perfectly recovered ($w_3 = Aw_3 = 1.0$)
two large clusters not well recovered

Summary

New validity indices for individual clusters

- provide more information than overall measures
- do not require arbitrary matching of clusters

Overall indices are weighted averages of new cluster indices

- weights are # object pairs in each cluster
- decomposition of overall indices into cluster information
- reflect how well large clusters are recovered

(Indices from information theory, Hanneke this afternoon)

References

Hubert LJ, Arabie P (1985) Comparing partitions. Journal of Classification 2:193-218

Rand WM (1971) Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66:846-850

Steinley D (2004) Properties of the Hubert-Arabie adjusted Rand index. Psychological Methods 9:386-396

Wallace DL (1983) Comment on A method for comparing two hierarchical clusterings. Journal of the American Statistical Association 78:569-576

Warrens MJ (2008) On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index. Journal of Classification 25:177-183