Understanding external validity indices

Matthijs J Warrens
&
Hanneke van der Hoef

IFCS Tokyo
August 10, 2017
Comparing partitions

- two partitions $U$ and $U'$ of same set of objects
- matching table $N = \{n_{ij}\}$ of size $I \times J$ where $n_{ij}$ is number of objects in $U_i \in U$ and in $U'_j \in U'$

Example of $N$ of size $3 \times 3$:

<table>
<thead>
<tr>
<th></th>
<th>$U'_1$</th>
<th>$U'_2$</th>
<th>$U'_3$</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_1$</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>$n_{13}$</td>
<td>$n_{1+}$</td>
</tr>
<tr>
<td>$U_2$</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
<td>$n_{23}$</td>
<td>$n_{2+}$</td>
</tr>
<tr>
<td>$U_3$</td>
<td>$n_{31}$</td>
<td>$n_{32}$</td>
<td>$n_{33}$</td>
<td>$n_{3+}$</td>
</tr>
<tr>
<td>total</td>
<td>$n_{+1}$</td>
<td>$n_{+2}$</td>
<td>$n_{+3}$</td>
<td>$n$</td>
</tr>
</tbody>
</table>
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

This talk
- analysis of indices based on **counting pairs**
- what information do these indices reflect?
This talk

Analysis of indices
- express indices in terms of cluster information
- weighted averages of cluster indices

Three index families (prototypical examples)
- Dice (1945), Wallace (1983)
- adjusted Rand index (Hubert and Arabie 1985)
- Rand (1971) index

What information do these indices reflect?
- agreement on large clusters
- indices have a very limited usefulness
Total number of object pairs (with $n$ objects):

$$N = \binom{n}{2}$$

Number of objects pairs in $U_i$ and $U$:

$$P_i = \binom{n_i+}{2} \quad \text{and} \quad P = \sum_{i=1}^{I} P_i$$

Number of objects pairs in $U'_j$ and $U'$:

$$P'_j = \binom{n+j}{2} \quad \text{and} \quad P' = \sum_{j=1}^{J} P'_j$$
Proportion of object pairs in $U_i$ also joined in $U'$:

$$w_i = \sum_{j=1}^{J} \left( \frac{n_{ij}}{2} \right)$$

For example:

$$w_1 = \left[ \left( \frac{102}{2} \right) + \left( \frac{0}{2} \right) + \left( \frac{0}{2} \right) \right] \div \left( \frac{102}{2} \right) = 1$$

$$w_2 = w_3 = \left[ \left( \frac{0}{2} \right) + \left( \frac{15}{2} \right) + \left( \frac{10}{2} \right) \right] \div \left( \frac{25}{2} \right) = .50$$

<table>
<thead>
<tr>
<th></th>
<th>$U'_1$</th>
<th>$U'_2$</th>
<th>$U'_3$</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_1$</td>
<td>102</td>
<td>0</td>
<td>0</td>
<td>102</td>
</tr>
<tr>
<td>$U_2$</td>
<td>0</td>
<td>15</td>
<td>10</td>
<td>25</td>
</tr>
<tr>
<td>$U_3$</td>
<td>0</td>
<td>10</td>
<td>15</td>
<td>25</td>
</tr>
<tr>
<td>total</td>
<td>102</td>
<td>25</td>
<td>25</td>
<td>152</td>
</tr>
</tbody>
</table>
Building blocks

Proportion of object pairs in $U_i$ also joined in $U'$:

$$w_i = \frac{\sum_{j=1}^{J} \binom{n_{ij}}{2}}{P_i}$$

Proportion of object pairs in $U'_j$ also joined in $U$:

$$w'_j = \frac{\sum_{i=1}^{I} \binom{n_{ij}}{2}}{P'_j}$$

adjusted (Warrens 2008a,b):

$$Aw_i = \frac{N \sum_{j=1}^{J} \binom{n_{ij}}{2} - P_i P'}{P_i (N - P')}$$

adjusted:

$$Aw'_j = \frac{N \sum_{i=1}^{I} \binom{n_{ij}}{2} - P'_j P}{P'_j (N - P)}$$
Dice and Wallace indices

Dice (1945) index

\[
D = \frac{\sum_{i=1}^{I} w_i P_i + \sum_{j=1}^{J} w_j' P'_j}{\sum_{i=1}^{I} P_i + \sum_{j=1}^{J} P'_j}
\]

Basic building blocks are Wallace (1983) indices

\[
W = \frac{\sum_{i=1}^{I} w_i P_i}{\sum_{i=1}^{I} P_i}
\]

and

\[
W' = \frac{\sum_{j=1}^{J} w_j' P'_j}{\sum_{j=1}^{J} P'_j}
\]

\(W\) is proportion of object pairs in \(U\) also joined in \(U'\)

Family 1: indices that are functions of \(W\) and \(W'\)
adjusted Rand index

Hubert and Arabie (1985)

\[ AR = \frac{\sum_{i=1}^{I} Aw_i P_i + \sum_{j=1}^{J} Aw'_j P'_j}{\sum_{i=1}^{I} P_i + \sum_{j=1}^{J} P'_j} \]

Basic blocks are adjusted Wallace indices (Severiano et al. 2011)

\[ AW = \frac{\sum_{i=1}^{I} Aw_i P_i}{\sum_{i=1}^{I} P_i} \quad \text{and} \quad AW' = \frac{\sum_{j=1}^{J} Aw'_j P'_j}{\sum_{j=1}^{J} P'_j} \]

\( AW \) is adjusted version of \( W \)

Family 2: indices that are functions of \( AW \) and \( AW' \)
Rand (1971) index

$$R = \frac{\sum_{i=1}^{I} w_i P_i + \sum_{j=1}^{J} w'_j P'_j + V(N - P) + V'(N - P')}{\sum_{i=1}^{I} P_i + \sum_{j=1}^{J} P'_j + N - P + N - P'}$$

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

$$V = \frac{N + T - P - P'}{N - P}$$ and $$V' = \frac{N + T - P - P'}{N - P'}$$

$$V$$ is proportion object pairs not together in $$U$$ also not in $$U'$$

Family 3: indices that are functions of $$W$$, $$W'$$, $$V$$ and $$V'$$
## Two examples

<table>
<thead>
<tr>
<th></th>
<th>$U'_1$</th>
<th>$U'_2$</th>
<th>$U'_3$</th>
<th>indices</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_1$</td>
<td>102</td>
<td>0</td>
<td>0</td>
<td>$D = .95$</td>
</tr>
<tr>
<td>$U_2$</td>
<td>0</td>
<td>15</td>
<td>10</td>
<td>$AR = .90$</td>
</tr>
<tr>
<td>$U_3$</td>
<td>0</td>
<td>10</td>
<td>15</td>
<td>$R = .95$</td>
</tr>
</tbody>
</table>

indices high ($\geq .90$): high agreement?
perfect agreement on large cluster, low agreement on small clusters

<table>
<thead>
<tr>
<th></th>
<th>$U'_1$</th>
<th>$U'_2$</th>
<th>$U'_3$</th>
<th>indices</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_1$</td>
<td>52</td>
<td>48</td>
<td>0</td>
<td>$D = .50$</td>
</tr>
<tr>
<td>$U_2$</td>
<td>46</td>
<td>54</td>
<td>0</td>
<td>$AR = .08$</td>
</tr>
<tr>
<td>$U_3$</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>$R = .55$</td>
</tr>
</tbody>
</table>

indices low ($0.08 - .55$): low agreement?
perfect agreement on small cluster, low agreement on large clusters
Limited usefulness

Pair-counting indices
- functions of cluster indices
- reflect agreement on large clusters
- do not reflect agreement on small clusters

To be useful

Requirement: cluster sizes must be equal
- otherwise, indices reflect agreement on large clusters only
- Romano et al. (2016) have same conclusion for AR
Cluster size sensitivity

Unexplored idea for removing sensitivity to cluster size imbalance

Instead of weighted averages

\[
D = \frac{\sum_{i=1}^{I} w_i P_i + \sum_{j=1}^{J} w_j' P_j'}{\sum_{i=1}^{I} P_i + \sum_{j=1}^{J} P_j'} \quad \text{and} \quad AR = \frac{\sum_{i=1}^{I} A w_i P_i + \sum_{j=1}^{J} A w_j' P_j'}{\sum_{i=1}^{I} P_i + \sum_{j=1}^{J} P_j'}
\]

use normal averages of cluster indices

\[
D^* = \frac{1}{2I} \sum_{i=1}^{I} w_i + \frac{1}{2J} \sum_{j=1}^{J} w_j' \quad \text{and} \quad AR^* = \frac{1}{2I} \sum_{i=1}^{I} A w_i + \frac{1}{2J} \sum_{j=1}^{J} A w_j'
\]
Discussion

Pair-counting indices
- functions of cluster indices
- limited usefulness

Alternative agreement analysis
- normal instead of weighted averages
- information-theoretic indices (Hanneke van der Hoef)
- set matching indices
- analyze agreement on the cluster level


