Variable Selection for Consistent Clustering

Ron Yurko    Rebecca Nugent    Sam Ventura

Department of Statistics & Data Science
Carnegie Mellon University

Symposium on Data Science and Statistics 2018
Variable choice → inconsistent clusters

Methods disagree using both variables, but agree on two consistent clusters with Variable 2
Variable Selection for Consistent Clustering

GOAL:

Search for the variables yielding consistent clusters based on the level of agreement between methods
Variable Selection for Consistent Clustering

GOAL:

Search for the variables yielding consistent clusters based on the level of agreement between methods

We are NOT optimizing for recovery of ”true cluster labels”

We ARE optimizing for agreement of obvious group structure
Measuring clustering agreement with ARI

\[ ARI(p_1, p_2) = \text{Adjusted Rand Index (ARI)}^{1}, \text{ similarity index between two partitions } p_1 \text{ and } p_2 \]

Corrected for chance agreement,

\[ \mathbb{E}[ARI(p_1, p_2)] = 0 \]

\[ ARI(p_1, p_2) < 0 \rightarrow \text{worse than random} \]

\[ ARI(p_1, p_2) = 1 \rightarrow \text{identical partitions} \]

\[^1\text{[Hubert and Arabie, 1985]}\]
Maximum Clustering Similarity (MCS)$^2$

An approach to determine $K$, number of clusters

Let $M =$ set of clustering methods

Choose $K$ with most frequent max similarity,

\[ e.g. ARI(p_{1,K}, p_{2,K}) \] from \[ \binom{|M|}{2} \] partition pairs

$^2$[Albatineh and Niewiadomska-Bugaj, 2011]
Greedy search algorithm for variable selection

Idea: **Greedily search for the most consistent subset of variables across clustering methods and number of clusters** $K$

Notation:
- $\mathbf{X} = N \times D$ data matrix, $d \in \{1, \ldots, D\}$
- $S$ = set of selected variables
- $U$ = set of unselected variables, where $S \cup U = \{1, \ldots, D\}$ and $S \cap U = \{\emptyset\}$
- $M = \{\text{complete, single, Ward, average, McQuitty, median, centroid, kmeans}\}$ (*just for illustrative purposes*)

**Step 0:** Initialize $S = \{\emptyset\}$, $U = \{1, \ldots, D\}$
Greedy search algorithm for variable selection

**Step 1:** For each variable \( d \in U \) and \( K \):

Create partitions \( p_{m_1}, K, S \cup \{d\}, \ldots, p_{m_{|M|}}, K, S \cup \{d\} \)

Compute \( ARI(p_{m_i}, K, S \cup \{d\}, p_{m_j}, K, S \cup \{d\}) \) for each of the \( \binom{|M|}{2} \) pairs of partitions
Greedy search algorithm for variable selection

**Step 1:** For each variable $d \in U$ and $K$:

Create partitions $p_{m_1,K,S \cup \{d\}}, \ldots, p_{m_{|M|},K,S \cup \{d\}}$

Compute $ARI(p_{m_i,K,S \cup \{d\}}, p_{m_j,K,S \cup \{d\}})$ for each of the $\binom{|M|}{2}$ pairs of partitions

**Step 2:** Select most consistent result:

$$d^*, K^* := \arg \max_{d \in U, K} ARI_{K,S \cup \{d\}}$$
Greedy search algorithm for variable selection

**Step 1:** For each variable $d \in U$ and $K$:

Create partitions $p_{m_1,K,S \cup \{d\}}, \ldots, p_{m_{|M|},K,S \cup \{d\}}$

Compute $\text{ARI}(p_{m_i,K,S \cup \{d\}}, p_{m_j,K,S \cup \{d\}})$ for each of the $\binom{|M|}{2}$ pairs of partitions

**Step 2:** Select most consistent result:

$$d^*, K^* := \arg \max_{d \in U, K} \text{ARI}_{K,S \cup \{d\}}$$

**Step 3:** Update $S = S \cup \{d^*\}$ and $U = U \setminus \{d^*\}$
Greedy search algorithm for variable selection

**Step 1:** For each variable \( d \in U \) and \( K \):

Create partitions \( p_{m_1, K, S \cup \{d\}}, \ldots, p_{m_{|M|}, K, S \cup \{d\}} \)

Compute \( ARI(p_{m_i, K, S \cup \{d\}}, p_{m_j, K, S \cup \{d\}}) \) for each of the \( \binom{|M|}{2} \) pairs of partitions

**Step 2:** Select most consistent result:

\[
d^*, K^* := \arg \max_{d \in U, K} ARI(K, S \cup \{d\})
\]

**Step 3:** Update \( S = S \cup \{d^*\} \) and \( U = U \setminus \{d^*\} \)

Repeat 1-3 until \( U = \{\emptyset\} \) or met stopping criteria
Demo data

4 true variables, 1 noise variable, and $K = 3$
Step 1 of demo search

Step 1: Select True4, ARI = 0.83, K = 2
Consensus matrices for full search

Step 1:
True4,
ARI = 0.83, K = 2

Step 2:
True1,
ARI = 1.00, K = 2

Step 3:
True2,
ARI = 1.00, K = 3

Step 4:
True3,
ARI = 1.00, K = 3

Step 5:
Noise,
ARI = 0.74, K = 4

Number of Methods in Agreement
8 6 4 2 0
Bootstrap consistency distributions to address limitations

We want to provide a measure of **confidence** in our decision:

- $f_{K,S} = \text{bootstrap distribution for } \text{ARI}_{K,S}$
- $f_{K,S \cup \{d\}} = \text{bootstrap distribution for } \text{ARI}_{K,S \cup \{d\}}$
- $\text{overlap}(f_{K,S}, f_{K,S \cup \{d\}}) = \text{area of overlap between the two}$
Bootstrap consistency distributions to address limitations

We want to provide a measure of **confidence** in our decision:

- \( f_{K,S} = \) bootstrap distribution for \( \overline{ARI}_{K,S} \)
- \( f_{K,S \cup \{d\}} = \) bootstrap distribution for \( \overline{ARI}_{K,S \cup \{d\}} \)
- \( overlap(f_{K,S}, f_{K,S \cup \{d\}}) = \) area of overlap between the two

Include variable \( d \) based on distribution **overlap**

**IF** \( \exists d \in U \) such that \( \overline{ARI}_{K,S \cup \{d\}} > \overline{ARI}_{K,S} \) (more consistent)

\[
d^*, K^* := \arg \min_{d \in U, K} overlap(f_{K,S}, f_{K,S \cup \{d\}}) \quad \text{(minimize overlap)}
\]
Bootstrap consistency distributions to address limitations

We want to provide a measure of **confidence** in our decision:

- $f_{K,S} = \text{bootstrap distribution for } \overline{ARI}_{K,S}$
- $f_{K,S∪\{d\}} = \text{bootstrap distribution for } \overline{ARI}_{K,S∪\{d\}}$
- $\text{overlap}(f_{K,S}, f_{K,S∪\{d\}}) = \text{area of overlap between the two}$

Include variable $d$ based on distribution **overlap**

**IF** $\exists \ d \in U$ such that $\overline{ARI}_{K,S∪\{d\}} > \overline{ARI}_{K,S}$ (more consistent)

$$d^*, K^* := \arg \min_{d ∈ U,K} \text{overlap}(f_{K,S}, f_{K,S∪\{d\}}) \ (\text{minimize overlap})$$

**ELSE** (less consistent)

$$d^*, K^* := \arg \max_{d ∈ U,K} \text{overlap}(f_{K,S}, f_{K,S∪\{d\}}) \ (\text{maximize overlap})$$
Bootstrap distributions for step 2 of demo search

Step 2: Given True4, Select True1, $K = 3$, $\overline{ARI} = 0.87$, Overlap = 0.31
Noise has minimal overlap and is not selected

<table>
<thead>
<tr>
<th>Step</th>
<th>Variable</th>
<th>(\overline{ARI})</th>
<th>K</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True4</td>
<td>0.8339</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>True1</td>
<td>0.8668</td>
<td>3</td>
<td>0.3067</td>
</tr>
<tr>
<td>3</td>
<td>True2</td>
<td>1.000</td>
<td>3</td>
<td>0.1338</td>
</tr>
<tr>
<td>4</td>
<td>True3</td>
<td>0.9979</td>
<td>3</td>
<td>0.3277</td>
</tr>
<tr>
<td>5</td>
<td>Noise</td>
<td>0.7444</td>
<td>4</td>
<td>0.0969</td>
</tr>
</tbody>
</table>

By measuring the overlap, we are confident that including Noise leads to inconsistent clustering results.
Swiss bank notes example

200 bills that are either counterfeit or real with 6 measurements

Summary of search reveals decrease in consistency:

<table>
<thead>
<tr>
<th>STEP</th>
<th>VARIABLE</th>
<th>$\text{ARI}$</th>
<th>K</th>
<th>OVERLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Diagonal</td>
<td>0.8755</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>Left</td>
<td>0.7500</td>
<td>2</td>
<td>0.8969</td>
</tr>
<tr>
<td>3</td>
<td>Right</td>
<td>0.6418</td>
<td>2</td>
<td>0.8789</td>
</tr>
<tr>
<td>4</td>
<td>Bottom</td>
<td>0.6112</td>
<td>3</td>
<td>0.6401</td>
</tr>
<tr>
<td>5</td>
<td>Top</td>
<td>0.7438</td>
<td>4</td>
<td>0.7262</td>
</tr>
<tr>
<td>6</td>
<td>Length</td>
<td>0.4113</td>
<td>4</td>
<td>0.7916</td>
</tr>
</tbody>
</table>
Swiss bank notes consensus matrices

Step 1: Diagonal, ARI = 0.88, K = 2
Step 2: Left, ARI = 0.75, K = 2
Step 3: Right, ARI = 0.64, K = 3
Step 4: Bottom, ARI = 0.61, K = 3
Step 5: Top, ARI = 0.74, K = 4
Step 6: Length, ARI = 0.41, K = 4

Number of Methods in Agreement:
- 8
- 6
- 4
- 2
- 0
Future Work

Simulation study, examine properties of ARI values

Explore different notions of stopping criteria
  - Only considered average, but distributions are multimodal and asymmetrical (e.g. mass above threshold?)

Inclusion of removal step

Consider sensitivity to different types of clustering methods
  - What about soft partitions? $^3$

$^3$[Flynt et al.,]
Thanks and References


Flynt, A., Dean, N., and Nugent, R. sari: An agreement measure for pairs of class assignments incorporating posterior probabilities.
How many clusters? which clustering method? answers via model-based cluster analysis.

*seriation: Infrastructure for Ordering Objects Using Seriation.*
R package version 1.2-2.

Comparing partitions.
*Journal of Classification, 2*:193–218.

*gclus: Clustering Graphics.*
R package version 1.3.1.
Thanks and References III


