Cluster stability for class discovery: when and how to use it

Martina Sundqvist, Julien Chiquet - Inra,
Thierry Dubois - Institut Curie, Guillem Rigaill - Inra

15/10/2019

Model selection in unsupervised classification

Selecting the number of clusters (k)

- An *open question in statistics*
- Many methods exists
- Google answer:
  "The best k is, which works best for your particular task".

Model selection in unsupervised classification

Selecting the number of clusters (k)

- An *open question in statistics*
- Many methods exists
- Google answer:
  "The best k is, which works best for your particular task".

Using stability?

- Principle:
  A stable clustering reveals the true structure of the data
- Commonly used method for cluster determination in oncology. . .
- Several variants: Consensus clustering [Monti et al., 2003], **Cluster Stability**

Outline

1. **Introduction to cluster stability**
2. When does it work?
3. How can we improve it?

Cluster Stability Algorithm

In the vein of Von Luxburg 2010:

Algorithm Clustering Stability

1: **Generate perturbed versions of the dataset** $(subsampling)$
2: **Cluster each perturbed dataset** $(clustering\ algorithm)$
3: **Compare obtained clusters** $Score$: $Sc()$
4: **Compute instability index** $\hat{I}_k()$

Choose the parameter k that gives the best stability (lowest instability):

$$\widehat{k} = \underset{k=1,\ldots,K}{Argmin}\ \hat{I}_k$$

Cluster Stability Algorithm

In the vein of Von Luxburg 2010:

Algorithm Clustering Stability

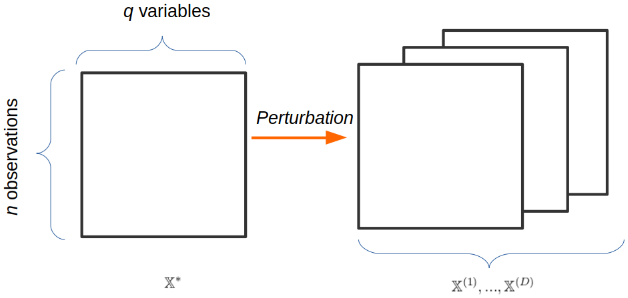
1: **Generate perturbed versions of the dataset** $(subsampling)$
2: **Cluster each perturbed dataset** $(clustering\ algorithm)$
3: **Compare obtained clusters** $Score$: $Sc()$
4: **Compute instability index** $\hat{I}_k()$

Choose the parameter k that gives the best stability (lowest instability):

$$\widehat{k} = \underset{k=1,\ldots,K}{Argmin}\ \hat{I}_k$$

1. Generate perturbed versions of the dataset

- Perturbed versions of the dataset can be obtained by:
  - Subsample variables or observations of the dataset
  - Adding noise
  - Random projecting of data in a smaller space



- *Bias linked to the parameters of perturbation? eg. percentage of subsampling?*

Cluster Stability Algorithm

In the vein of Von Luxburg 2010:

Algorithm Clustering Stability

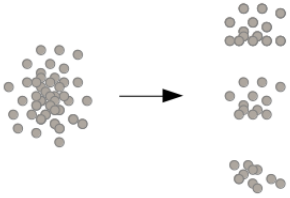
1: **Generate perturbed versions of the dataset** $(subsampling)$
2: **Cluster each perturbed dataset** $(clustering\ algorithm)$
3: **Compare obtained clusters** $Score$: $Sc()$
4: **Compute instability index** $\hat{I}_k()$

Choose the parameter k that gives the best stability (lowest instability):

$$\widehat{k} = \underset{k=1,\ldots,K}{Argmin}\ \hat{I}_k$$

2. Cluster each perturbed dataset

- Cluster Algorithms:
  - Probabilistic: Gaussian mixture model
  - Model free: Hierarchical ascendant clustering, K-means, Spectral clustering, etc.



- *Bias linked to each clustering algorithm?*

Cluster Stability Algorithm

In the vein of Von Luxburg 2010:

Algorithm Clustering Stability

1: **Generate perturbed versions of the dataset** $(subsampling)$
2: **Cluster each perturbed dataset** $(clustering\ algorithm)$
3: **Compare obtained clusters** $Score$: $Sc()$
4: **Compute instability index** $\hat{I}_k()$

Choose the parameter k that gives the best stability (lowest instability):

$$\widehat{k} = \underset{k=1,\dots,K}{Argmin}\ \hat{I}_k$$

3. Compare obtained clusters

- Type of score: [Vinh et al., 2010]
  - **Adjusted Rand Index (ARI)**:
    - Corrected for chance
    - Not a real distance
  - **Normalized Information Distance (NID)**:
    - Not corrected for chance
    - A real distance

3. Compare obtained clusters

- Type of score: [Vinh et al., 2010]
  - **Adjusted Rand Index (ARI)**:
    - ⇒ Corrected for chance
    - ⇒ Not a real distance
  - **Normalized Information Distance (NID)**:
    - ⇒ Not corrected for chance
    - ⇒ A real distance
- Type of clustering comparison:
  - Compare all pairs of obtained clusterings? Some of them?
  - Compare each obtained clustering to the initial classification?

Cluster Stability Algorithm

In the vein of Von Luxburg 2010:

Algorithm Clustering Stability

1: **Generate perturbed versions of the dataset** $(subsampling)$
2: **Cluster each perturbed dataset** $(clustering\ algorithm)$
3: **Compare obtained clusters** $Score$: $Sc()$
4: **Compute instability index** $\hat{I}_k()$

Choose the parameter k that gives the best stability (lowest instability):

$$\widehat{k} = \underset{k=1,\ldots,K}{Argmin}\ \hat{I}_k$$

Cluster Stability Algorithm

In the vein of Von Luxburg 2010:

Algorithm Clustering Stability

1: **Generate perturbed versions of the dataset** $(subsampling)$
2: **Cluster each perturbed dataset** $(clustering\ algorithm)$
3: **Compare obtained clusters** $Score$: $Sc()$
4: **Compute instability index** $\hat{I}_k()$

Choose the parameter k that gives the best stability (lowest instability):

$$\widehat{k} = \underset{k=1,\ldots,K}{Argmin}\ \hat{I}_k$$

Outline

1. Introduction to cluster stability
2. **When does it work?**
3. How can we improve it?

Cluster stability for class discovery, when does it work ?

1. Q1: Does the most stable cluster structure correspond to the real underlying structure of the data?
2. Q2: Is it possible to estimate the "true" cluster stability?
   - If yes, when is it the case?

Simulation: Experimental setting

- **Idealize model (IM)** Generate D datasets with n observations coming from k^* distinct Gaussian populations (distributions) with different population mean.

$$\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_D$$

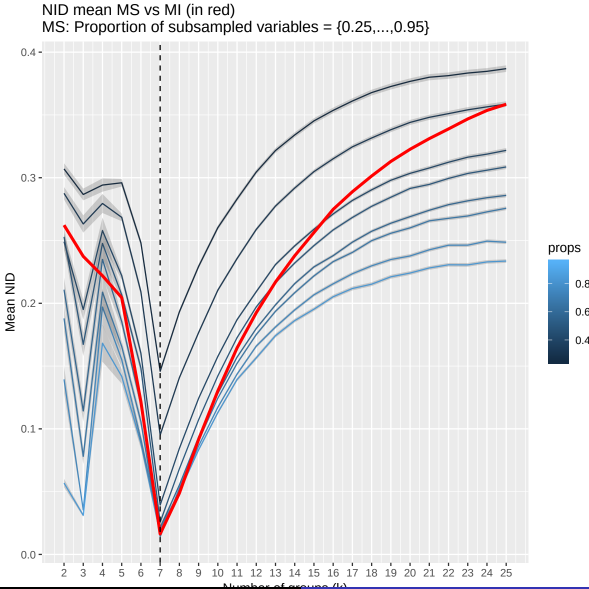
- **Sampled model (SM)** Generate **one** dataset as above, from which D datasets are subsampled.

$$\mathbb{X} \rightarrow \mathbb{X}^{(1)}, \mathbb{X}^{(2)}, ..., \mathbb{X}^{(D)}$$

Simulation parameters (simple setting)

- Simulated data
  - $k^* = 7$
  - group size: 50
  - group means: $\mu = [-6, -4, -2, 0, 2, 4, 6]$
  - $\sigma = 1$
- Clustering
  - clustering algorithm: k-means
  - $k = \{1, \ldots, 25\}$
  - $score$: NID
- Varying parameter:
  - Proportion of subsampled variables (sampled model)

Simulated results (simple setting)



Observation (simple setting)

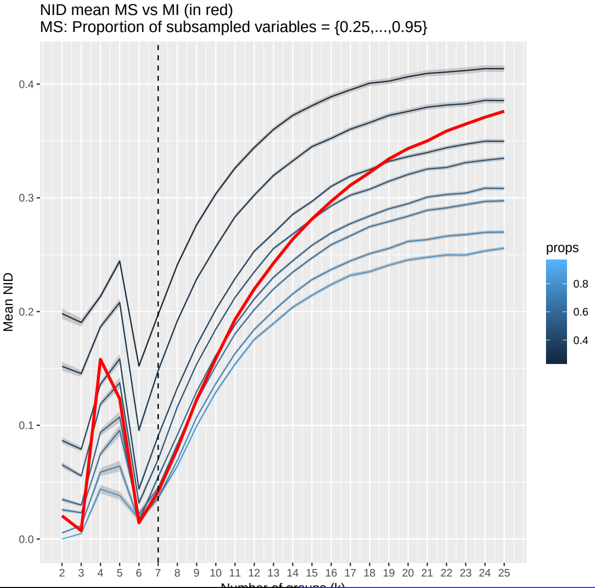
- **Idealized stability:**
  - I_k has its minimum at k^*.
- **Sampled model:**
  - $\hat{I}_k$ tends to have the same minimum as I_k, but unstable for some proportions of subsampling.

What happens if we **change the mean value** of one of the groups?

Simulation parameters (a bit more complex setting)

- Simulated data
  - $k^* = 7$
  - group size: 50
  - group means: $\mu = [-6, -4, -2, 0, 2, \mathbf{5}, 6]$
  - $\sigma = 1$
- Clustering
  - clustering algorithm: k-means
  - $k = \{1, \ldots, 25\}$
  - $score$: NID
- Varying parameter:
  - Proportion of subsampled variables (sampled model)

Simulated results (a bit more complex setting)



Observation (a bit more complex setting)

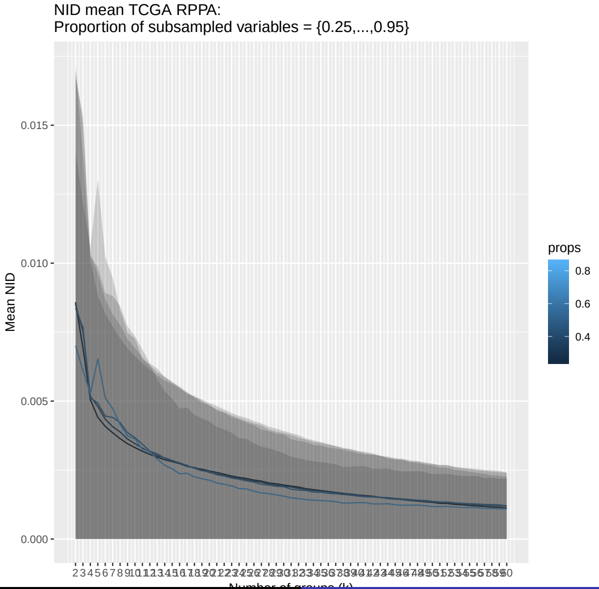
- **Idealized stability:**
  - I_k minimum is not at 7
  - I_k minimum is not at 6
  - but at 3
- **Sampled stability:**
  - Minimum of $\hat{I}_k$ depends on the proportion of subsampled variables.

What happens for **more complex** data?

Triple Negative Breast Cancers (TNBC) study

- Data cohort: TCGA (public) [TCGA, 2012]
- Tumor samples extracted from TNBC patients
- Type of data: protein expression (RPPA)
- $n = 350$, $p = 100$

Results: Cluster stability TCGA



Conclusions - when does it work?

- Q1: Does the most stable cluster structure correspond to the real underlying structure of the data?
  - Yes, in certain cases
  - No even in some simple settings
- Q2: Is it possible to estimate the "true" cluster stability?
  - Yes in certain cases
  - Parameter dependent

⇒ Cluster stability for class discovery should be **used with caution**

Outline

1. Introduction to cluster stability
2. When does it work?
3. **How can we improve it?**

Problematic

- Cluster stability is not a **"magical measure"** and needs to be used with caution
- Stable does not imply **biologically relevant**
- A clearer **separation between statistical analysis and biological interpretation** is needed

⇒ **Return to cluster comparison scores!**

The Rand Index

- **The Rand Index (RI)**, counts the number of consistent pairs in between two classifications (Rand, 1971)
- **The Adjusted Rand Index (ARI)**, (Hubert & Arabie, 1985):

$$ARI = \frac{RI - \mathbb{E}(RI)}{1 - \mathbb{E}(RI)}$$

+ Corrected by change
– Supposes that classifications are independent
– Difficult to interpret

A new ARI?

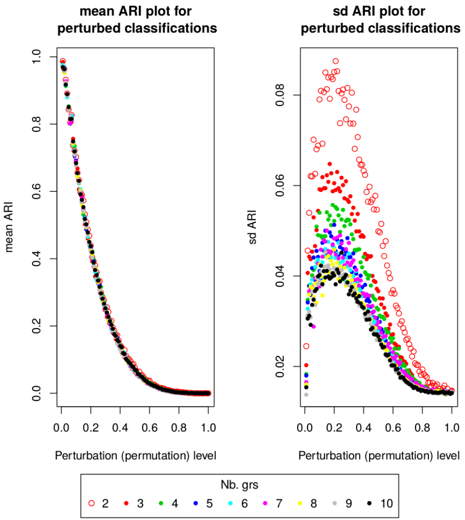
- **Idea:**
  - Introduce p the level of perturbation to the ARI
  - p being the probability of permutation

$$ARI_p = \frac{RI - \mathbb{E}(RI \mid p)}{\mathbb{V}(RI \mid p)}$$

- Which parameters might influence $\mathbb{E}(ARI)$ and $\mathbb{V}(ARI)$?

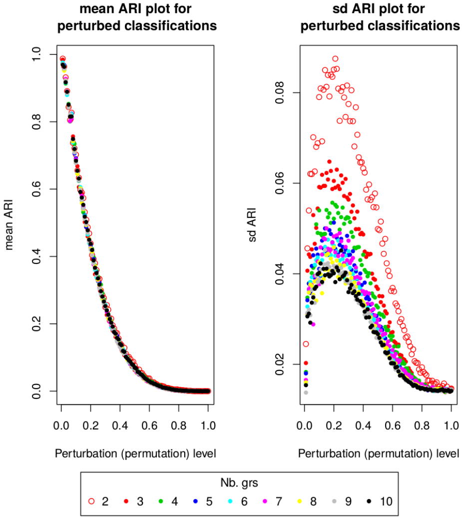
$\Rightarrow$ Simulations

$\mathbb{E}(ARI)$ & $\mathbb{V}(ARI)$ with varying p



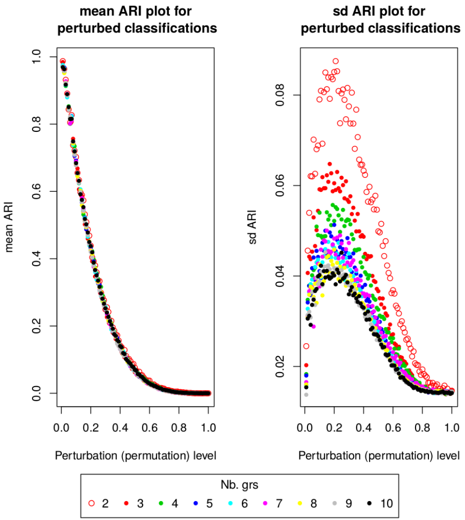
$\mathbb{E}(ARI)$ & $\mathbb{V}(ARI)$ with varying p

- $\mathbb{E}(ARI)$ depend on: p & <u>but not</u> K



$\mathbb{E}(ARI)$ & $\mathbb{V}(ARI)$ with varying p

- $\mathbb{E}(ARI)$ depend on: p & but not K



- $\mathbb{V}(ARI)$ depend on: p & K

A new ARI?

- **Conclusion**: simulations of ARI
  - $\mathbb{V}(ARI)$ and $\mathbb{E}(ARI)$ depend on p
  - $\mathbb{V}(ARI)$ and depends on K
- **Estimate** p: Analytically or Computationally

Estimate p analytically (ongoing work)

$$\mathbb{E}(RI \mid p) = (1-p) + p^2 \mathbb{E}_{\mathcal{H}0}(RI) - p(1-p)\sum_{k=1}^{K} \pi_k^2 + 2p(1-p)\sum_{k=1}^{K} \pi_k^3$$

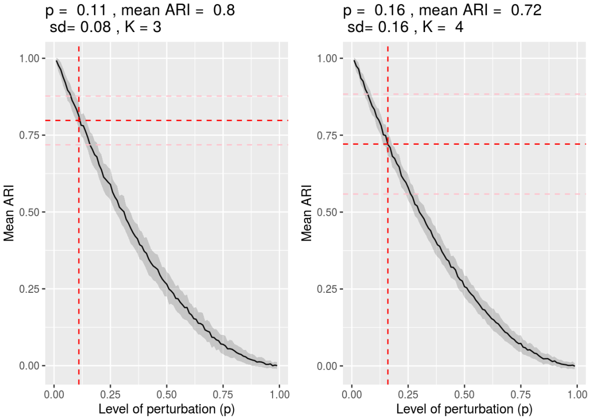
with π_k the probability for an observation to be in group k

- $p = 1$, $\mathbb{E}(RI \mid p) = \mathbb{E}_{\mathcal{H}0}(RI) \rightarrow$ Classifications are independent
- $p = 0$, $\mathbb{E}(RI \mid p) = 1 \rightarrow$ Classifications are identical

Estimate p computationally: Iris flower dataset

- Data
  - Fisher (1936), *The use of multiple measurements in taxonomic problems*
  - n = 150
  - 3 speices: Iris setosa, Iris virginica and Iris versicolor
  - 4 measured variables: the length and the width of the sepals and petals
  - **Debate:** 3 or 4 groups?
- Clustering
  - cluster algorithm: K-means
  - Proportion of subsampled variables: 0.5
  - $nsim = 100$

Estimate p for Iris flower dataset: Results



$$p : K = 3 < K = 4$$

Improvement: Conclusions

1. Estimate p from observed ARI
   - Computationally and analytically
   - Gives biological interpretation to cluster comparison score

2. Take into account p in $\mathbb{E}(RI)$ and $\mathbb{V}(RI)$ is needed to compute ARI_p

Conclusions and perspectives

How to use cluster stability as a class discovery criterion?

- Cluster stability as a class discovery criterion
  - Do not always work
  - Indicates for which K the classification is the most stable, but not to which extent it is biological pertinent

$\Rightarrow$ Introduce p as a measure of clustering perturbation

Conclusions and perspectives

How to use cluster stability as a class discovery criterion?

- Cluster stability as a class discovery criterion
  - Do not always work
  - Indicates for which K the classification is the most stable, but not to which extent it is biological pertinent

$\Rightarrow$ Introduce p as a measure of clustering perturbation

Perspectives

- Apply to classifications for Triple Negative Breast Cancers
- Implement in R package

Thanks for your attention!

✉ martina.sundqvist@agroparistech.fr

References

Hubert, L. and Arabie, P. (1985).
Comparing partitions.
Journal of classification, 2(1):193–218.

Rand, W. (1971).
Objective criteria for the evaluation of clustering methods.
Journal of the American Statistical Association, 66(336):846–850.

Vinh, N. X., Epps, J., and Bailey, J. (2010).
Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance.
Journal of Machine Learning Research, 11(Oct):2837–2854.

Von Luxburg, U. et al. (2010).
Clustering stability: an overview.
Foundations and Trends® in Machine Learning, 2(3):235–274.