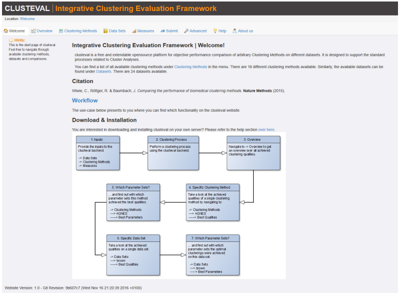
Comparing the Performance of Biomedical Clustering Methods



Hints:
This is the start page of clusteval. Feel free to navigate through available clustering methods, datasets and comparisons.

Integrative Clustering Evaluation Framework | Welcome!

clusteval is a free and extendable opensource platform for objective performance comparison of arbitrary Clustering Methods on different datasets. It is designed to support the standard processes related to Cluster Analyses.

You can find a list of all available clustering methods under Clustering Methods in the menu. There are 18 different clustering methods available. Similarly, the available datasets can be found under Datasets. There are 24 datasets available.

Citation

Wiwie, C., Röttger, R. & Baumbach, J. *Comparing the performance of biomedical clustering methods.* **Nature Methods** (2015).

Workflow

The use-case below presents to you where you can find which functionality on the clusteval website.

Download & Installation

You are interested in downloading and installing clusteval on your own server? Please refer to the help section over here.

1. Inputs
Provide the inputs to the clusteval backend.
-> Data Sets
-> Clustering Methods
-> Measures

2. Clustering Process
Perform a clustering process using the clusteval backend.

3. Overview
Navigate to -> Overview to get an overview over all achieved clustering qualities.

4. Specific Clustering Method
Take a look at the achieved qualities of a single clustering method by navigating to:
-> Clustering Methods
---> AGNES
-----> Best Qualities

5. Which Parameter Sets?
... and find out with which parameter sets this method achieved the best qualities:
-> Clustering Methods
---> AGNES
-----> Best Parameters

6. Specific Data Set
Take a look at the achieved qualities on a single data set.
-> Data Sets
---> brown
-----> Best Qualities

7. Which Parameter Sets?
... and find out with which parameter sets the optimal clusterings were achieved on this data set.
-> Data Sets
---> brown
-----> Best Parameters

Website Version: 1.0 - Git Revision: 9b607c7 (Wed Nov 16 21:20:39 2016 +0100)

Barbara Ikica

Faculty of Mathematics and Physics, UL & Institute of Mathematics, Physics and Mechanics

9 January 2018

ClustEval - Integrative Clustering Evaluation Framework

C. Wiwie, R. Röttger, J. Baumbach, Comparing the performance of biomedical clustering methods, *Nat. Methods* **12** (2015), 1033–1038.

ClustEval - Integrative Clustering Evaluation Framework

C. Wiwie, R. Röttger, J. Baumbach, Comparing the performance of biomedical clustering methods, *Nat. Methods* **12** (2015), 1033–1038.

- a clustering analysis platform *ClustEval*, available at `https://clusteval.sdu.dk/1/mains`

ClustEval - Integrative Clustering Evaluation Framework

C. Wiwie, R. Röttger, J. Baumbach, Comparing the performance of biomedical clustering methods, *Nat. Methods* **12** (2015), 1033–1038.

- a clustering analysis platform *ClustEval*, available at `https://clusteval.sdu.dk/1/mains`
- 18 different clustering methods, 24 datasets, 18 common cluster validity indices

ClustEval - Integrative Clustering Evaluation Framework

C. Wiwie, R. Röttger, J. Baumbach, Comparing the performance of biomedical clustering methods, *Nat. Methods* **12** (2015), 1033–1038.

- a clustering analysis platform *ClustEval*, available at `https://clusteval.sdu.dk/1/mains`
- 18 different clustering methods, 24 datasets, 18 common cluster validity indices
- robustness analysis (density reduction, noise addition)

ClustEval - Integrative Clustering Evaluation Framework

C. Wiwie, R. Röttger, J. Baumbach, Comparing the performance of biomedical clustering methods, *Nat. Methods* **12** (2015), 1033–1038.

- a clustering analysis platform *ClustEval*, available at `https://clusteval.sdu.dk/1/mains`
- 18 different clustering methods, 24 datasets, 18 common cluster validity indices
- robustness analysis (density reduction, noise addition)
- parameter optimisation (Naïve Grid Optimisation, Adaptive Grid Optimisation)

Clustering in Biomedicine

Applicable to:

- cancer subtyping on the basis of gene expression levels,

Clustering in Biomedicine

Applicable to:

- cancer subtyping on the basis of gene expression levels,
- protein homology detection from amino acid sequences of structures,

Clustering in Biomedicine

Applicable to:

- cancer subtyping on the basis of gene expression levels,
- protein homology detection from amino acid sequences of structures,
- the identification of protein complexes using protein-protein interactions

Key Problems

Key Problems

1. **Tool picking**

Key Problems

1. **Tool picking** (partitioning (k-means), hierarchical, density-based and graph-based approaches)

Key Problems

1. **Tool picking** (partitioning (k-means), hierarchical, density-based and graph-based approaches)
2. **Parameter optimisation**

Key Problems

1. **Tool picking** (partitioning (k-means), hierarchical, density-based and graph-based approaches)
2. **Parameter optimisation** (parameters influence the number and size of resulting clusters)

Key Problems

1. **Tool picking** (partitioning (k-means), hierarchical, density-based and graph-based approaches)
2. **Parameter optimisation** (parameters influence the number and size of resulting clusters)
3. **Quality measures/Cluster validity indices**

Key Problems

1. **Tool picking** (partitioning (k-means), hierarchical, density-based and graph-based approaches)
2. **Parameter optimisation** (parameters influence the number and size of resulting clusters)
3. **Quality measures/Cluster validity indices** (internal, external)

Key Problems

1. **Tool picking** (partitioning (k-means), hierarchical, density-based and graph-based approaches)
2. **Parameter optimisation** (parameters influence the number and size of resulting clusters)
3. **Quality measures/Cluster validity indices** (internal, external)
4. **Standardised evaluation**

Key Problems

1. **Tool picking** (partitioning (k-means), hierarchical, density-based and graph-based approaches)
2. **Parameter optimisation** (parameters influence the number and size of resulting clusters)
3. **Quality measures/Cluster validity indices** (internal, external)
4. **Standardised evaluation** (poor comparability of clustering results)

Clustering Methods

Clustering Methods

Partitioning

The objects are assigned to clusters and iteratively change clusters based on their dissimilarity in order to optimise a given target function.

Clustering Methods

Partitioning

The objects are assigned to clusters and iteratively change clusters based on their dissimilarity in order to optimise a given target function.

Fanny, k-means, Partitioning Around Medoids

Clustering Methods

Partitioning

The objects are assigned to clusters and iteratively change clusters based on their dissimilarity in order to optimise a given target function.

Fanny, k-means, Partitioning Around Medoids

Hierarchical

The dataset is transformed into a tree-like structure where the leaves represent the objects and the inner nodes the hierarchical relationship between them.

Clustering Methods

Partitioning

The objects are assigned to clusters and iteratively change clusters based on their dissimilarity in order to optimise a given target function.

Fanny, k-means, Partitioning Around Medoids

Hierarchical

The dataset is transformed into a tree-like structure where the leaves represent the objects and the inner nodes the hierarchical relationship between them.

Hierarchical Clustering, Spectral Clustering

Clustering Methods

Density-based

Identifying regions with a locally similar object density.

Clustering Methods

Density-based

Identifying regions with a locally similar object density.

Clusterdp, DBSCAN

Clustering Methods

Density-based

Identifying regions with a locally similar object density.

Clusterdp, DBSCAN

Graph-based

The input is considered as a graph with the objects being the nodes connected by weighted or unweighted edges. The clustering problem is solved by solving an analogous graph theoretical problem (e.g. clique finding, simulating random walks).

Clustering Methods

Density-based

Identifying regions with a locally similar object density.

Clusterdp, DBSCAN

Graph-based

The input is considered as a graph with the objects being the nodes connected by weighted or unweighted edges. The clustering problem is solved by solving an analogous graph theoretical problem (e.g. clique finding, simulating random walks).

Affinity Propagation, clusterONE, Markov Clustering, MCODE, Transitivity Clustering

Datasets

Datasets

- Gene expression levels,

Datasets

- Gene expression levels,
- Protein-protein interaction,

Datasets

- Gene expression levels,
- Protein-protein interaction,
- Protein-sequence similarity,

Datasets

- Gene expression levels,
- Protein-protein interaction,
- Protein-sequence similarity,
- Social network,

Datasets

- Gene expression levels,
- Protein-protein interaction,
- Protein-sequence similarity,
- Social network,
- Synthetic (easy/medium/hard),

Datasets

- Gene expression levels,
- Protein-protein interaction,
- Protein-sequence similarity,
- Social network,
- Synthetic (easy/medium/hard),
- Language-processing datasets used for word-sense disambiguation

Clustering Validity Indices

Internal Indices

The clustering is judged on the basis of certain intrinsic statistical properties of the clustering itself.

Clustering Validity Indices

Internal Indices

The clustering is judged on the basis of certain intrinsic statistical properties of the clustering itself.

Davies-Bouldin Index, Dunn Index, Silhouette Value

Clustering Validity Indices

Internal Indices

The clustering is judged on the basis of certain intrinsic statistical properties of the clustering itself.

Davies-Bouldin Index, Dunn Index, Silhouette Value

External Indices

The clustering is compared to a user-given gold-standard clustering (using a pairwise/mapping approach).

Clustering Validity Indices

Internal Indices

The clustering is judged on the basis of certain intrinsic statistical properties of the clustering itself.

Davies-Bouldin Index, Dunn Index, Silhouette Value

External Indices

The clustering is compared to a user-given gold-standard clustering (using a pairwise/mapping approach).

F_β score, False Discovery rate, False Positive Rate, Fowles-Mallows Index, Jaccard Index, Rand Index, Sensitivity (Recall), Specificity, V-measure

Clustering Validity Indices

Davies-Bouldin Index

$$\boxed{DB = \frac{1}{n} \cdot \sum_{c_i \in C} \max_{c_i \neq c_j \in C} \left(\frac{\sigma_i + \sigma_j}{\|\overline{c_i} - \overline{c_j}\|} \right)},$$

$$\sigma_i = \sqrt{\frac{1}{|c_i|} \sum_{x_i \in c_i} \|x_i - \overline{c_i}\|}$$

Clustering Validity Indices

Davies-Bouldin Index

$$\boxed{DB = \frac{1}{n} \cdot \sum_{c_i \in C} \max_{c_i \neq c_j \in C} \left(\frac{\sigma_i + \sigma_j}{\|\overline{c_i} - \overline{c_j}\|} \right)},$$

$$\sigma_i = \sqrt{\frac{1}{|c_i|} \sum_{x_i \in c_i} \|x_i - \overline{c_i}\|}$$

Dunn Index

$$\boxed{D = \frac{\min\limits_{c_i \neq c_j \in C} \left(\min\limits_{x_i \in c_i, x_j \in c_j} d(x_i, x_j) \right)}{\max\limits_{c_k \in C} \left(\max\limits_{x_i, x_j \in c_k} d(x_i, x_j) \right)}}$$

Clustering Validity Indices

Silhouette Value

$c(x_i)$ – the cluster of object x_i
$o(x_i)$ – the closest cluster to x_i

$$\boxed{S = \frac{1}{n} \sum_i sv_i}$$

$$\begin{aligned} sv_i &= \frac{1}{n} \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}} \\ a(x_i) &= \frac{1}{|c(x_i)|} \sum_{x_j \in c(x_i)} d(x_i, x_j) \\ b(x_i) &= \frac{1}{|o(x_i)|} \sum_{x_j \in o(x_i)} d(x_i, x_j) \end{aligned}$$

Clustering Validity Indices

Silhouette Value

$c(x_i)$ – the cluster of object x_i

$o(x_i)$ – the closest cluster to x_i

$$\boxed{S = \frac{1}{n} \sum_i sv_i}$$

Correlation among internal and external indices

The silhouette value correlated best (0.71) with the F1 score, F2 score, FM index, Jaccard index and V-measure.

Guidelines (working with a new *biomedical* set without a gold standard)

Guidelines (working with a new *biomedical* set without a gold standard)

1. Use *Transitivity Clustering*, *Hierarchical Clustering* or *Partitioning Around Medoids*.

Guidelines (working with a new *biomedical* set without a gold standard)

1. Use *Transitivity Clustering*, *Hierarchical Clustering* or *Partitioning Around Medoids*.
2. Compute the silhouette values for clustering results using a broad range of parameter-set variations.

Guidelines (working with a new *biomedical* set without a gold standard)

1. Use *Transitivity Clustering*, *Hierarchical Clustering* or *Partitioning Around Medoids*.
2. Compute the silhouette values for clustering results using a broad range of parameter-set variations.
3. Pick the result for the parameter set yielding the highest silhouette value.

Guidelines (working with a new *biomedical* set without a gold standard)

1. Use *Transitivity Clustering*, *Hierarchical Clustering* or *Partitioning Around Medoids*.
2. Compute the silhouette values for clustering results using a broad range of parameter-set variations.
3. Pick the result for the parameter set yielding the highest silhouette value.

Remark. The silhouette value is a particularly poor measure for entangled and highly overlapping datasets.