Package ‘stabm’

November 9, 2020

Title Stability Measures for Feature Selection

Version 1.1.3

Description An implementation of many measures for the assessment of the stability of feature selection. Both simple measures and measures which take into account the similarities between features are available, see Bommert et al. (2017) <doi:10.1155/2017/7907163>.

License LGPL-3

URL https://bommert.github.io/stabm/, https://github.com/bommert/stabm

BugReports https://github.com/bommert/stabm/issues

Depends R (>= 3.5.0)

Imports checkmate (>= 1.8.5), Matrix (>= 1.2-14), methods, stats, utils

Suggests cowplot (>= 0.9.2), ggdendro (>= 0.1-20), ggplot2 (>= 3.0.0), igraph (>= 1.2.1), knitr, rmarkdown, testthat (>= 2.0.0)

VignetteBuilder knitr

Encoding UTF-8

RoxygenNote 7.1.1

NeedsCompilation no

Author Andrea Bommert [aut, cre] (<https://orcid.org/0000-0002-1005-9351>), Michel Lang [aut] (<https://orcid.org/0000-0001-9754-0393>)

Maintainer Andrea Bommert <bommert@statistik.tu-dortmund.de>

Repository CRAN

Date/Publication 2020-11-09 13:10:06 UTC

R topics documented:

`stabm-package` *stabm: Stability Measures for Feature Selection*

Description

An implementation of many measures for the assessment of the stability of feature selection. Both simple measures and measures which take into account the similarities between features are available, see Bommert et al. (2017) <doi:10.1155/2017/7907163>.

Author(s)

Maintainer: Andrea Bommert `<bommert@statistik.tu-dortmund.de>` (ORCID)

Authors:

- Michel Lang `<michellang@gmail.com>` (ORCID)

See Also

Useful links:

- `https://bommert.github.io/stabm/`
- `https://github.com/bommert/stabm`
- Report bugs at `https://github.com/bommert/stabm/issues`

`listStabilityMeasures` *List All Available Stability Measures*

Description

Lists all stability measures of package *stabm* and provides information about them.

Usage

```
listStabilityMeasures()
```

Value

`data.frame`
For each stability measure, its name, the information, whether it is corrected for chance by definition, the information, whether it is adjusted for similar features, its minimal value and its maximal value are displayed.

Note

The given minimal values might only be reachable in some scenarios, e.g. if the feature sets have a certain size. The measures which are not corrected for chance by definition can be corrected for chance with `correction.for.chance`. This however changes the minimal value. For the adjusted stability measures, the minimal value depends on the similarity structure.

`plotFeatures` *Plot Selected Features*

Description

Creates a heatmap of the features which are selected in at least one feature set. The sets are ordered according to average linkage hierarchical clustering based on the Manhattan distance. If `sim.mat` is given, the features are ordered according to average linkage hierarchical clustering based on `1 -sim.mat`. Otherwise, the features are ordered in the same way as the feature sets.

Note that this function needs the packages **ggplot2**, **cowplot** and **ggdendro** installed.

Usage

```
plotFeatures(features, sim.mat = NULL)
```

Arguments

`features` — `list (length >= 2)`
Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).

`sim.mat` — `numeric matrix`
Similarity matrix which contains the similarity structure of all features based on all datasets. The similarity values must be in the range of [0, 1] where 0 indicates very low similarity and 1 indicates very high similarity. If the list elements of `features` are integerish vectors, then the feature numbering must correspond to the ordering of `sim.mat`. If the list elements of `features` are character vectors, then `sim.mat` must be named and the names of `sim.mat` must correspond to the entries in `features`.

Value

Object of class `ggplot`.

Examples

```
feats = list(1:3, 1:4, 1:5)
mat = 0.92 ^ abs(outer(1:10, 1:10, "-"))
plotFeatures(features = feats)
plotFeatures(features = feats, sim.mat = mat)
```

---

stabilityDavis — *Stability Measure Davis*

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilityDavis(
  features,
  p,
  correction.for.chance = "none",
  N = 10000,
  impute.na = NULL,
  penalty = 0
)
```

Arguments

`features` — `list (length >= 2)`
Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).

`p` — `numeric(1)`
Total number of features in the datasets. Required, if `correction.for.chance` is set to "estimate" or "exact".

`correction.for.chance` — `character(1)`
Should a correction for chance be applied? Correction for chance means that if features are chosen at random, the expected value must be independent of the number of chosen features. To correct for chance, the original score is transformed by $(score - expected)/(maximum - expected)$. For stability measures whose score is the average value of pairwise scores, this transformation is done for all components individually. Options are "none", "estimate" and "exact". For "none", no correction is performed, i.e. the original score is used. For "estimate", `N` random feature sets of the same sizes as the input feature sets (`features`) are generated. For "exact", all possible combinations of feature sets of the same sizes as the input feature sets are used. Computation is only feasible for very small numbers of features (`p`) and numbers of considered datasets (`length(features)`).

`N` — `numeric(1)`
Number of random feature sets to consider. Only relevant if `correction.for.chance` is set to "estimate".

`impute.na` — `numeric(1)`
In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? `NULL` means no imputation.

`penalty` — `numeric(1)`
Penalty parameter, see Details.

Details

The stability measure is defined as (see Notation)

$$\max\left\{0, \frac{1}{|V|}\sum_{j=1}^{p}\frac{h_j}{m} - \frac{penalty}{p}\cdot median\{|V_1|,\ldots,|V_m|\}\right\}.$$

Value

`numeric(1)` Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1, \ldots, V_m$ denote the sets of chosen features for the m datasets, i.e. `features` has length m and V_i is a set which contains the i-th entry of `features`. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^{p} h_j = \sum_{i=1}^{m} |V_i|$ and $V = \bigcup_{i=1}^{m} V_i$.

References

Bommert A, Rahnenführer J, Lang M (2017). "A Multicriteria Approach to Find Predictive and Sparse Models with Stable Feature Selection for High-Dimensional Data." *Computational and Mathematical Methods in Medicine*, **2017**, 1–18. doi: 10.1155/2017/7907163. Davis CA, Gerick F, Hintermair V, Friedel CC, Fundel K, Kuffner R, Zimmer R (2006). "Reliable gene signatures for microarray classification: assessment of stability and performance." *Bioinformatics*, **22**(19), 2356–2363. doi: 10.1093/bioinformatics/btl400.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
stabilityDavis(features = feats, p = 10)
```

`stabilityDice` *Stability Measure Dice*

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilityDice(
  features,
  p = NULL,
  correction.for.chance = "none",
  N = 10000,
  impute.na = NULL
)
```

Arguments

features `list (length >= 2)`
Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).

p `numeric(1)`
Total number of features in the datasets. Required, if `correction.for.chance` is set to "estimate" or "exact".

correction.for.chance
`character(1)`
Should a correction for chance be applied? Correction for chance means that if features are chosen at random, the expected value must be independent of the number of chosen features. To correct for chance, the original score is transformed by $(score - expected)/(maximum - expected)$. For stability measures whose score is the average value of pairwise scores, this transformation is done for all components individually. Options are "none", "estimate" and "exact". For "none", no correction is performed, i.e. the original score is used. For "estimate", `N` random feature sets of the same sizes as the input feature sets (`features`) are generated. For "exact", all possible combinations of feature sets of the same sizes as the input feature sets are used. Computation is only feasible for very small numbers of features (`p`) and numbers of considered datasets (`length(features)`).

N `numeric(1)`
Number of random feature sets to consider. Only relevant if `correction.for.chance` is set to "estimate".

impute.na `numeric(1)`
In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? `NULL` means no imputation.

Details

The stability measure is defined as (see Notation)

$$\frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{2|V_i \cap V_j|}{|V_i| + |V_j|}.$$

Value

`numeric(1)` Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1, \ldots, V_m$ denote the sets of chosen features for the m datasets, i.e. `features` has length m and V_i is a set which contains the i-th entry of `features`. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is

chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^{p} h_j = \sum_{i=1}^{m} |V_i|$ and $V = \bigcup_{i=1}^{m} V_i$.

References

Bommert A, Rahnenführer J, Lang M (2017). "A Multicriteria Approach to Find Predictive and Sparse Models with Stable Feature Selection for High-Dimensional Data." *Computational and Mathematical Methods in Medicine*, **2017**, 1–18. doi: 10.1155/2017/7907163. Dice LR (1945). "Measures of the Amount of Ecologic Association Between Species." *Ecology*, **26**(3), 297–302. doi: 10.2307/1932409.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
stabilityDice(features = feats)
```

---

stabilityHamming *Stability Measure Hamming*

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilityHamming(
  features,
  p,
  correction.for.chance = "none",
  N = 10000,
  impute.na = NULL
)
```

Arguments

features
: `list` (length >= 2)
Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).

p
: `numeric(1)`
Total number of features in the datasets. Required, if `correction.for.chance` is set to "estimate" or "exact".

correction.for.chance
: `character(1)`
Should a correction for chance be applied? Correction for chance means that if features are chosen at random, the expected value must be independent of the number of chosen features. To correct for chance, the original score is transformed by $(score - expected)/(maximum - expected)$. For stability measures whose score is the average value of pairwise scores, this transformation is done for all components individually. Options are "none", "estimate" and "exact". For "none", no correction is performed, i.e. the original score is used. For "estimate", `N` random feature sets of the same sizes as the input feature sets (`features`) are generated. For "exact", all possible combinations of feature sets of the same sizes as the input feature sets are used. Computation is only feasible for very small numbers of features (`p`) and numbers of considered datasets (`length(features)`).

N
: `numeric(1)`
Number of random feature sets to consider. Only relevant if `correction.for.chance` is set to "estimate".

impute.na
: `numeric(1)`
In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? `NULL` means no imputation.

Details

The stability measure is defined as (see Notation)

$$\frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{|V_i \cap V_j| + |V_i^c \cap V_j^c|}{p}.$$

Value

`numeric(1)` Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1, \ldots, V_m$ denote the sets of chosen features for the m datasets, i.e. `features` has length m and V_i is a set which contains the i-th entry of `features`. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is

chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^{p} h_j = \sum_{i=1}^{m} |V_i|$ and $V = \bigcup_{i=1}^{m} V_i$.

References

Dunne, Kevin, Cunningham, Padraig, Azuaje, Francisco (2002). “Solutions to instability problems with sequential wrapper-based approaches to feature selection.” Machine Learning Group, Department of Computer Science, Trinity College, Dublin.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
stabilityHamming(features = feats, p = 10)
```

`stabilityIntersectionCount`

Stability Measure Adjusted Intersection Count

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilityIntersectionCount(
  features,
  sim.mat,
  threshold = 0.9,
  correction.for.chance = "estimate",
  N = 10000,
  impute.na = NULL
)
```

Arguments

features
: list (length >= 2)
Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (character) or indices (integerish).

sim.mat
: numeric matrix
Similarity matrix which contains the similarity structure of all features based on all datasets. The similarity values must be in the range of [0, 1] where 0 indicates very low similarity and 1 indicates very high similarity. If the list elements of features are integerish vectors, then the feature numbering must correspond to the ordering of sim.mat. If the list elements of features are character vectors, then sim.mat must be named and the names of sim.mat must correspond to the entries in features.

threshold
: numeric(1)
Threshold for indicating which features are similar and which are not. Two features are considered as similar, if and only if the corresponding entry of sim.mat is greater than or equal to threshold.

correction.for.chance
: character(1)
How should the expected value of the stability score (see Details) be assessed? Options are "estimate", "exact" and "none". For "estimate", N random feature sets of the same sizes as the input feature sets (features) are generated. For "exact", all possible combinations of feature sets of the same sizes as the input feature sets are used. Computation is only feasible for very small numbers of features and numbers of considered datasets (length(features)). For "none", the transformation $(score - expected)/(maximum - expected)$ is not conducted, i.e. only $score$ is used. This is not recommended.

N
: numeric(1)
Number of random feature sets to consider. Only relevant if correction.for.chance is set to "estimate".

impute.na
: numeric(1)
In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? NULL means no imputation.

Details

The stability measure is defined as (see Notation)

$$\frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{I(V_i, V_j) - E(I(V_i, V_j))}{\sqrt{|V_i| \cdot |V_j|} - E(I(V_i, V_j))}$$

with

$$I(V_i, V_j) = |V_i \cap V_j| + \min(C(V_i, V_j), C(V_j, V_i))$$

and

$$C(V_k, V_l) = |\{x \in V_k \backslash V_l : \exists y \in V_l \backslash V_k \; with \; Similarity(x, y) \geq threshold\}|.$$

Value

`numeric(1)` Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1, \dots, V_m$ denote the sets of chosen features for the m datasets, i.e. `features` has length m and V_i is a set which contains the i-th entry of `features`. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^p h_j = \sum_{i=1}^m |V_i|$ and $V = \bigcup_{i=1}^m V_i$.

References

Bommert A, Rahnenführer J (2020). "Adjusted Measures for Feature Selection Stability for Data Sets with Similar Features." In *International Conference on Machine Learning, Optimization, and Data Science - LOD 2020, in print*. https://arxiv.org/abs/2009.12075.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
mat = 0.92 ^ abs(outer(1:10, 1:10, "-"))
stabilityIntersectionCount(features = feats, sim.mat = mat, N = 1000)
```

`stabilityIntersectionGreedy`

Stability Measure Adjusted Intersection Greedy

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilityIntersectionGreedy(
  features,
  sim.mat,
  threshold = 0.9,
  correction.for.chance = "estimate",
  N = 10000,
  impute.na = NULL
)
```

Arguments

`features` — `list` (length >= 2)
Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).

`sim.mat` — `numeric matrix`
Similarity matrix which contains the similarity structure of all features based on all datasets. The similarity values must be in the range of [0, 1] where 0 indicates very low similarity and 1 indicates very high similarity. If the list elements of `features` are integerish vectors, then the feature numbering must correspond to the ordering of `sim.mat`. If the list elements of `features` are character vectors, then `sim.mat` must be named and the names of `sim.mat` must correspond to the entries in `features`.

`threshold` — `numeric(1)`
Threshold for indicating which features are similar and which are not. Two features are considered as similar, if and only if the corresponding entry of `sim.mat` is greater than or equal to `threshold`.

`correction.for.chance` — `character(1)`
How should the expected value of the stability score (see Details) be assessed? Options are "estimate", "exact" and "none". For "estimate", `N` random feature sets of the same sizes as the input feature sets (`features`) are generated. For "exact", all possible combinations of feature sets of the same sizes as the input feature sets are used. Computation is only feasible for very small numbers of features and numbers of considered datasets (`length(features)`). For "none", the transformation $(score - expected)/(maximum - expected)$ is not conducted, i.e. only $score$ is used. This is not recommended.

`N` — `numeric(1)`
Number of random feature sets to consider. Only relevant if `correction.for.chance` is set to "estimate".

`impute.na` — `numeric(1)`
In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? `NULL` means no imputation.

Details

The stability measure is defined as (see Notation)

$$\frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{I(V_i, V_j) - E(I(V_i, V_j))}{\sqrt{|V_i| \cdot |V_j|} - E(I(V_i, V_j))}$$

with

$$I(V_i, V_j) = |V_i \cap V_j| + GMBM(V_i \backslash V_j, V_j \backslash V_i).$$

$GMBM(V_i \backslash V_j, V_j \backslash V_i)$ denotes a greedy approximation of $MBM(V_i \backslash V_j, V_j \backslash V_i)$, see stabilityIntersectionMBM.

Value

`numeric(1)` Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1, \ldots, V_m$ denote the sets of chosen features for the m datasets, i.e. `features` has length m and V_i is a set which contains the i-th entry of `features`. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^{p} h_j = \sum_{i=1}^{m} |V_i|$ and $V = \bigcup_{i=1}^{m} V_i$.

References

Bommert A, Rahnenführer J (2020). "Adjusted Measures for Feature Selection Stability for Data Sets with Similar Features." In *International Conference on Machine Learning, Optimization, and Data Science - LOD 2020, in print*. https://arxiv.org/abs/2009.12075.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
mat = 0.92 ^ abs(outer(1:10, 1:10, "-"))
stabilityIntersectionGreedy(features = feats, sim.mat = mat, N = 1000)
```

`stabilityIntersectionMBM` *Stability Measure Adjusted Intersection MBM*

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilityIntersectionMBM(
  features,
  sim.mat,
  threshold = 0.9,
  correction.for.chance = "estimate",
  N = 10000,
  impute.na = NULL
)
```

Arguments

`features` `list` (`length` >= 2)
Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).

`sim.mat` `numeric` `matrix`
Similarity matrix which contains the similarity structure of all features based on all datasets. The similarity values must be in the range of [0, 1] where 0 indicates very low similarity and 1 indicates very high similarity. If the list elements of `features` are integerish vectors, then the feature numbering must correspond to the ordering of `sim.mat`. If the list elements of `features` are character vectors, then `sim.mat` must be named and the names of `sim.mat` must correspond to the entries in `features`.

`threshold` `numeric(1)`
Threshold for indicating which features are similar and which are not. Two features are considered as similar, if and only if the corresponding entry of `sim.mat` is greater than or equal to `threshold`.

`correction.for.chance`
`character(1)`
How should the expected value of the stability score (see Details) be assessed?

Options are "estimate", "exact" and "none". For "estimate", `N` random feature sets of the same sizes as the input feature sets (`features`) are generated. For "exact", all possible combinations of feature sets of the same sizes as the input feature sets are used. Computation is only feasible for very small numbers of features and numbers of considered datasets (`length(features)`). For "none", the transformation $(score - expected)/(maximum - expected)$ is not conducted, i.e. only $score$ is used. This is not recommended.

`N` `numeric(1)`
Number of random feature sets to consider. Only relevant if `correction.for.chance` is set to "estimate".

`impute.na` `numeric(1)`
In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? `NULL` means no imputation.

Details

The stability measure is defined as (see Notation)

$$\frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{I(V_i, V_j) - E(I(V_i, V_j))}{\sqrt{|V_i| \cdot |V_j|} - E(I(V_i, V_j))}$$

with

$$I(V_i, V_j) = |V_i \cap V_j| + MBM(V_i \backslash V_j, V_j \backslash V_i).$$

$MBM(V_i \backslash V_j, V_j \backslash V_i)$ denotes the size of the maximum bipartite matching based on the graph whose vertices are the features of $V_i \backslash V_j$ on the one side and the features of $V_j \backslash V_i$ on the other side. Vertices x and y are connected if and only if $Similarity(x, y) \geq threshold$. Requires the package **igraph**.

Value

`numeric(1)` Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1, \ldots, V_m$ denote the sets of chosen features for the m datasets, i.e. `features` has length m and V_i is a set which contains the i-th entry of `features`. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^{p} h_j = \sum_{i=1}^{m} |V_i|$ and $V = \bigcup_{i=1}^{m} V_i$.

References

Bommert A, Rahnenführer J (2020). “Adjusted Measures for Feature Selection Stability for Data Sets with Similar Features.” In *International Conference on Machine Learning, Optimization, and Data Science - LOD 2020, in print*. https://arxiv.org/abs/2009.12075.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
mat = 0.92 ^ abs(outer(1:10, 1:10, "-"))
stabilityIntersectionMBM(features = feats, sim.mat = mat, N = 1000)
```

---

stabilityIntersectionMean

Stability Measure Adjusted Intersection Mean

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilityIntersectionMean(
  features,
  sim.mat,
  threshold = 0.9,
  correction.for.chance = "estimate",
  N = 10000,
  impute.na = NULL
)
```

Arguments

`features`	`list` (length >= 2) Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).
`sim.mat`	`numeric matrix` Similarity matrix which contains the similarity structure of all features based on all datasets. The similarity values must be in the range of [0, 1] where 0 indicates very low similarity and 1 indicates very high similarity. If the list elements of `features` are integerish vectors, then the feature numbering must correspond to the ordering of `sim.mat`. If the list elements of `features` are character vectors, then `sim.mat` must be named and the names of `sim.mat` must correspond to the entries in `features`.

threshold `numeric(1)`
Threshold for indicating which features are similar and which are not. Two features are considered as similar, if and only if the corresponding entry of `sim.mat` is greater than or equal to `threshold`.

correction.for.chance `character(1)`
How should the expected value of the stability score (see Details) be assessed? Options are "estimate", "exact" and "none". For "estimate", `N` random feature sets of the same sizes as the input feature sets (`features`) are generated. For "exact", all possible combinations of feature sets of the same sizes as the input feature sets are used. Computation is only feasible for very small numbers of features and numbers of considered datasets (`length(features)`). For "none", the transformation $(score - expected)/(maximum - expected)$ is not conducted, i.e. only $score$ is used. This is not recommended.

N `numeric(1)`
Number of random feature sets to consider. Only relevant if `correction.for.chance` is set to "estimate".

impute.na `numeric(1)`
In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? `NULL` means no imputation.

Details

The stability measure is defined as (see Notation)

$$\frac{2}{m(m-1)}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\frac{I(V_i,V_j)-E(I(V_i,V_j))}{\sqrt{|V_i|\cdot|V_j|}-E(I(V_i,V_j))}$$

with

$$I(V_i,V_j)=|V_i\cap V_j|+\min(C(V_i,V_j),C(V_j,V_i)),$$

$$C(V_k,V_l)=\sum_{x\in V_k\setminus V_l:|G_x^{kl}|>0}\frac{1}{|G_x^{kl}|}\sum_{y\in G_x^{kl}}Similarity(x,y)$$

and

$$G_x^{kl}=\{y\in V_l\setminus V_k:\ Similarity(x,y)\geq threshold\}.$$

Value

`numeric(1)` Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1,\ldots,V_m$ denote the sets of chosen features for the m datasets, i.e. `features` has length m and V_i is a set which contains the i-th entry of `features`. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q=\sum_{j=1}^{p}h_j=\sum_{i=1}^{m}|V_i|$ and $V=\bigcup_{i=1}^{m}V_i$.

References

Bommert A, Rahnenführer J (2020). "Adjusted Measures for Feature Selection Stability for Data Sets with Similar Features." In *International Conference on Machine Learning, Optimization, and Data Science - LOD 2020, in print*. https://arxiv.org/abs/2009.12075.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
mat = 0.92 ^ abs(outer(1:10, 1:10, "-"))
stabilityIntersectionMean(features = feats, sim.mat = mat, N = 1000)
```

---

stabilityJaccard *Stability Measure Jaccard*

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilityJaccard(
  features,
  p = NULL,
  correction.for.chance = "none",
  N = 10000,
  impute.na = NULL
)
```

Arguments

`features` `list (length >= 2)`
Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).

`p` `numeric(1)`
Total number of features in the datasets. Required, if `correction.for.chance` is set to "estimate" or "exact".

correction.for.chance
: character(1)
Should a correction for chance be applied? Correction for chance means that if features are chosen at random, the expected value must be independent of the number of chosen features. To correct for chance, the original score is transformed by $(score - expected)/(maximum - expected)$. For stability measures whose score is the average value of pairwise scores, this transformation is done for all components individually. Options are "none", "estimate" and "exact". For "none", no correction is performed, i.e. the original score is used. For "estimate", N random feature sets of the same sizes as the input feature sets (features) are generated. For "exact", all possible combinations of feature sets of the same sizes as the input feature sets are used. Computation is only feasible for very small numbers of features (p) and numbers of considered datasets (length(features)).

N
: numeric(1)
Number of random feature sets to consider. Only relevant if correction.for.chance is set to "estimate".

impute.na
: numeric(1)
In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? NULL means no imputation.

Details

The stability measure is defined as (see Notation)

$$\frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{|V_i \cap V_j|}{|V_i \cup V_j|}.$$

Value

numeric(1) Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1, \ldots, V_m$ denote the sets of chosen features for the m datasets, i.e. features has length m and V_i is a set which contains the i-th entry of features. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^{p} h_j = \sum_{i=1}^{m} |V_i|$ and $V = \bigcup_{i=1}^{m} V_i$.

References

Bommert A, Rahnenführer J, Lang M (2017). "A Multicriteria Approach to Find Predictive and Sparse Models with Stable Feature Selection for High-Dimensional Data." *Computational and Mathematical Methods in Medicine*, **2017**, 1–18. doi: 10.1155/2017/7907163. Jaccard, Paul (1901).

"Étude comparative de la distribution florale dans une portion des Alpes et du Jura." *Bulletin de la Société Vaudoise des Sciences Naturelles*, **37**, 547-579. doi: 10.5169/SEALS266450.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
stabilityJaccard(features = feats)
```

`stabilityKappa` *Stability Measure Kappa*

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilityKappa(features, p, impute.na = NULL)
```

Arguments

`features`	`list (length >= 2)` Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).
`p`	`numeric(1)` Total number of features in the datasets.
`impute.na`	`numeric(1)` In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? `NULL` means no imputation.

Details

The stability measure is defined as the average kappa coefficient between all pairs of feature sets. It can be rewritten as (see Notation)

$$\frac{2}{m(m-1)}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\frac{|V_i \cap V_j| - \frac{|V_i|\cdot|V_j|}{p}}{\frac{|V_i|+|V_j|}{2} - \frac{|V_i|\cdot|V_j|}{p}}.$$

Value

`numeric(1)` Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1, \ldots, V_m$ denote the sets of chosen features for the m datasets, i.e. `features` has length m and V_i is a set which contains the i-th entry of `features`. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^{p} h_j = \sum_{i=1}^{m} |V_i|$ and $V = \bigcup_{i=1}^{m} V_i$.

References

Cohen J (1960). "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement*, **20**(1), 37–46. doi: 10.1177/001316446002000104.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
stabilityKappa(features = feats, p = 10)
```

`stabilityLustgarten` *Stability Measure Lustgarten*

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilityLustgarten(features, p, impute.na = NULL)
```

Arguments

`features` `list (length >= 2)`
Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).

`p` `numeric(1)`
Total number of features in the datasets.

`impute.na` `numeric(1)`
In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? `NULL` means no imputation.

Details

The stability measure is defined as (see Notation)

$$\frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{|V_i \cap V_j| - \frac{|V_i| \cdot |V_j|}{p}}{\min\{|V_i|, |V_j|\} - \max\{0, |V_i| + |V_j| - p\}}.$$

Value

`numeric(1)` Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1, \ldots, V_m$ denote the sets of chosen features for the m datasets, i.e. `features` has length m and V_i is a set which contains the i-th entry of `features`. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^{p} h_j = \sum_{i=1}^{m} |V_i|$ and $V = \bigcup_{i=1}^{m} V_i$.

References

Bommert A, Rahnenführer J, Lang M (2017). "A Multicriteria Approach to Find Predictive and Sparse Models with Stable Feature Selection for High-Dimensional Data." *Computational and Mathematical Methods in Medicine*, **2017**, 1–18. doi: 10.1155/2017/7907163. Lustgarten, L J, Gopalakrishnan, Vanathi, Visweswaran, Shyam (2009). "Measuring stability of feature selection in biomedical datasets." In *AMIA annual symposium proceedings*, volume 2009, 406. American Medical Informatics Association.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
stabilityLustgarten(features = feats, p = 10)
```

`stabilityNogueira` *Stability Measure Nogueira*

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilityNogueira(features, p, impute.na = NULL)
```

Arguments

`features`	`list` (`length` >= 2) Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).
`p`	`numeric(1)` Total number of features in the datasets.
`impute.na`	`numeric(1)` In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? `NULL` means no imputation.

Details

The stability measure is defined as (see Notation)

$$1 - \frac{\frac{1}{p}\sum_{j=1}^{p}\frac{m}{m-1}\frac{h_j}{m}\left(1-\frac{h_j}{m}\right)}{\frac{q}{mp}(1-\frac{q}{mp})}.$$

Value

`numeric(1)` Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1, \ldots, V_m$ denote the sets of chosen features for the m datasets, i.e. `features` has length m and V_i is a set which contains the i-th entry of `features`. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^{p} h_j = \sum_{i=1}^{m} |V_i|$ and $V = \bigcup_{i=1}^{m} V_i$.

References

Nogueira S, Sechidis K, Brown G (2018). "On the Stability of Feature Selection Algorithms." *Journal of Machine Learning Research*, **18**(174), 1–54. https://jmlr.org/papers/v18/17-514.html.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
stabilityNogueira(features = feats, p = 10)
```

`stabilityNovovicova` *Stability Measure Novovičová*

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilityNovovicova(
  features,
  p = NULL,
  correction.for.chance = "none",
  N = 10000,
  impute.na = NULL
)
```

Arguments

features — list (length >= 2)
Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (character) or indices (integerish).

p — numeric(1)
Total number of features in the datasets. Required, if correction.for.chance is set to "estimate" or "exact".

correction.for.chance — character(1)
Should a correction for chance be applied? Correction for chance means that if features are chosen at random, the expected value must be independent of the number of chosen features. To correct for chance, the original score is transformed by $(score - expected)/(maximum - expected)$. For stability measures whose score is the average value of pairwise scores, this transformation is done for all components individually. Options are "none", "estimate" and "exact". For "none", no correction is performed, i.e. the original score is used. For "estimate", N random feature sets of the same sizes as the input feature sets (features) are generated. For "exact", all possible combinations of feature sets of the same sizes as the input feature sets are used. Computation is only feasible for very small numbers of features (p) and numbers of considered datasets (length(features)).

N — numeric(1)
Number of random feature sets to consider. Only relevant if correction.for.chance is set to "estimate".

impute.na — numeric(1)
In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? NULL means no imputation.

Details

The stability measure is defined as (see Notation)

$$\frac{1}{q\log_2(m)}\sum_{j:X_j\in V} h_j\log_2(h_j).$$

Value

numeric(1) Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1,\dots,V_m$ denote the sets of chosen features for the m datasets, i.e. features has length m and V_i is a set which contains the i-th entry of features. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is

chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^{p} h_j = \sum_{i=1}^{m} |V_i|$ and $V = \bigcup_{i=1}^{m} V_i$.

References

Bommert A, Rahnenführer J, Lang M (2017). "A Multicriteria Approach to Find Predictive and Sparse Models with Stable Feature Selection for High-Dimensional Data." *Computational and Mathematical Methods in Medicine*, **2017**, 1–18. doi: 10.1155/2017/7907163. Novovičová J, Somol P, Pudil P (2009). "A New Measure of Feature Selection Algorithms' Stability." In *2009 IEEE International Conference on Data Mining Workshops*. doi: 10.1109/icdmw.2009.32.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
stabilityNovovicova(features = feats)
```

`stabilityOchiai` *Stability Measure Ochiai*

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilityOchiai(
  features,
  p = NULL,
  correction.for.chance = "none",
  N = 10000,
  impute.na = NULL
)
```

Arguments

features `list (length >= 2)`
Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).

p `numeric(1)`
Total number of features in the datasets. Required, if `correction.for.chance` is set to "estimate" or "exact".

correction.for.chance `character(1)`
Should a correction for chance be applied? Correction for chance means that if features are chosen at random, the expected value must be independent of the number of chosen features. To correct for chance, the original score is transformed by $(score - expected)/(maximum - expected)$. For stability measures whose score is the average value of pairwise scores, this transformation is done for all components individually. Options are "none", "estimate" and "exact". For "none", no correction is performed, i.e. the original score is used. For "estimate", `N` random feature sets of the same sizes as the input feature sets (`features`) are generated. For "exact", all possible combinations of feature sets of the same sizes as the input feature sets are used. Computation is only feasible for very small numbers of features (`p`) and numbers of considered datasets (`length(features)`).

N `numeric(1)`
Number of random feature sets to consider. Only relevant if `correction.for.chance` is set to "estimate".

impute.na `numeric(1)`
In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? `NULL` means no imputation.

Details

The stability measure is defined as (see Notation)

$$\frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{|V_i \cap V_j|}{\sqrt{|V_i| \cdot |V_j|}}.$$

Value

`numeric(1)` Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1, \ldots, V_m$ denote the sets of chosen features for the m datasets, i.e. `features` has length m and V_i is a set which contains the i-th entry of `features`. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is

chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^{p} h_j = \sum_{i=1}^{m} |V_i|$ and $V = \bigcup_{i=1}^{m} V_i$.

References

Bommert A, Rahnenführer J, Lang M (2017). "A Multicriteria Approach to Find Predictive and Sparse Models with Stable Feature Selection for High-Dimensional Data." *Computational and Mathematical Methods in Medicine*, **2017**, 1–18. doi: 10.1155/2017/7907163. Ochiai A (1957). "Zoogeographical Studies on the Soleoid Fishes Found in Japan and its Neighbouring Regions-III." *Nippon Suisan Gakkaishi*, **22**(9), 531-535. doi: 10.2331/suisan.22.531.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
stabilityOchiai(features = feats)
```

---

stabilityPhi *Stability Measure Phi*

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilityPhi(features, p, impute.na = NULL)
```

Arguments

`features`	`list (length >= 2)` Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).
`p`	`numeric(1)` Total number of features in the datasets.

impute.na numeric(1)
In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? `NULL` means no imputation.

Details

The stability measure is defined as the average phi coefficient between all pairs of feature sets. It can be rewritten as (see Notation)

$$\frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{|V_i \cap V_j| - \frac{|V_i| \cdot |V_j|}{p}}{\sqrt{|V_i|(1 - \frac{|V_i|}{p}) \cdot |V_j|(1 - \frac{|V_j|}{p})}}.$$

Value

`numeric(1)` Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1, \ldots, V_m$ denote the sets of chosen features for the m datasets, i.e. `features` has length m and V_i is a set which contains the i-th entry of `features`. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^{p} h_j = \sum_{i=1}^{m} |V_i|$ and $V = \bigcup_{i=1}^{m} V_i$.

References

Bommert A, Rahnenführer J, Lang M (2017). "A Multicriteria Approach to Find Predictive and Sparse Models with Stable Feature Selection for High-Dimensional Data." *Computational and Mathematical Methods in Medicine*, **2017**, 1–18. doi: 10.1155/2017/7907163. Nogueira S, Brown G (2016). "Measuring the Stability of Feature Selection." In *Machine Learning and Knowledge Discovery in Databases*, 442–457. Springer International Publishing. doi: 10.1007/978331946227-1_28.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
stabilityPhi(features = feats, p = 10)
```

`stabilitySechidis` *Stability Measure Sechidis*

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilitySechidis(features, sim.mat, threshold = 0.9, impute.na = NULL)
```

Arguments

`features` — `list` (`length` >= 2)
Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).

`sim.mat` — `numeric` `matrix`
Similarity matrix which contains the similarity structure of all features based on all datasets. The similarity values must be in the range of [0, 1] where 0 indicates very low similarity and 1 indicates very high similarity. If the list elements of `features` are integerish vectors, then the feature numbering must correspond to the ordering of `sim.mat`. If the list elements of `features` are character vectors, then `sim.mat` must be named and the names of `sim.mat` must correspond to the entries in `features`.

`threshold` — `numeric(1)`
Threshold for indicating which features are similar and which are not. Two features are considered as similar, if and only if the corresponding entry of `sim.mat` is greater than or equal to `threshold`.

`impute.na` — `numeric(1)`
In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? `NULL` means no imputation.

Details

The stability measure is defined as

$$1 - \frac{trace(CS)}{trace(C\Sigma)}$$

with $(p \times p)$-matrices

$$(S)_{ij} = \frac{m}{m-1}\left(\frac{h_{ij}}{m} - \frac{h_i}{m}\frac{h_j}{m}\right)$$

and

$$(\Sigma)_{ii} = \frac{q}{mp}\left(1 - \frac{q}{mp}\right),$$

$$(\Sigma)_{ij} = \frac{\frac{1}{m}\sum_{i=1}^{m}|V_i|^2 - \frac{q}{m}}{p^2 - p} - \frac{q^2}{m^2p^2}, i \neq j.$$

The matrix C is created from matrix `sim.mat` by setting all values of `sim.mat` that are smaller than `threshold` to 0. If you want to C to be equal to `sim.mat`, use `threshold = 0`.

Value

`numeric(1)` Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1, \ldots, V_m$ denote the sets of chosen features for the m datasets, i.e. `features` has length m and V_i is a set which contains the i-th entry of `features`. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^{p} h_j = \sum_{i=1}^{m} |V_i|$ and $V = \bigcup_{i=1}^{m} V_i$.

Note

This stability measure is not corrected for chance. Unlike for the other stability measures in this R package, that are not corrected for chance, for `stabilitySechidis`, no `correction.for.chance` can be applied. This is because for `stabilitySechidis`, no finite upper bound is known at the moment, see listStabilityMeasures.

References

Sechidis K, Papangelou K, Nogueira S, Weatherall J, Brown G (2020). “On the Stability of Feature Selection in the Presence of Feature Correlations.” In *Machine Learning and Knowledge Discovery in Databases*, 327–342. Springer International Publishing. doi: 10.1007/9783030461508_20.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
mat = 0.92 ^ abs(outer(1:10, 1:10, "-"))
stabilitySechidis(features = feats, sim.mat = mat)
```

stabilitySomol *Stability Measure Somol*

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilitySomol(features, p, impute.na = NULL)
```

Arguments

`features`	`list` (length >= 2) Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).
`p`	`numeric(1)` Total number of features in the datasets.
`impute.na`	`numeric(1)` In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? `NULL` means no imputation.

Details

The stability measure is defined as (see Notation)

$$\frac{\left(\sum_{j=1}^{p} \frac{h_j}{q} \frac{h_j - 1}{m-1}\right) - c_{\min}}{c_{\max} - c_{\min}}$$

with

$$c_{\min} = \frac{q^2 - p(q - q \ mod \ p) - (q \ mod \ p)^2}{pq(m-1)},$$

$$c_{\max} = \frac{(q \ mod \ m)^2 + q(m-1) - (q \ mod \ m) \, m}{q(m-1)}.$$

Value

`numeric(1)` Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1, \ldots, V_m$ denote the sets of chosen features for the m datasets, i.e. `features` has length m and V_i is a set which contains the i-th entry of `features`. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^{p} h_j = \sum_{i=1}^{m} |V_i|$ and $V = \bigcup_{i=1}^{m} V_i$.

References

Bommert A, Rahnenführer J, Lang M (2017). "A Multicriteria Approach to Find Predictive and Sparse Models with Stable Feature Selection for High-Dimensional Data." *Computational and Mathematical Methods in Medicine*, **2017**, 1–18. doi: 10.1155/2017/7907163. Somol P, Novovičová J (2010). "Evaluating Stability and Comparing Output of Feature Selectors that Optimize Feature Subset Cardinality." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**(11), 1921–1939. doi: 10.1109/tpami.2010.34.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
stabilitySomol(features = feats, p = 10)
```

`stabilityUnadjusted` *Stability Measure Unadjusted*

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilityUnadjusted(features, p, impute.na = NULL)
```

Arguments

features `list (length >= 2)`
Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).

p `numeric(1)`
Total number of features in the datasets.

impute.na `numeric(1)`
In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? `NULL` means no imputation.

Details

The stability measure is defined as (see Notation)

$$\frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{|V_i \cap V_j| - \frac{|V_i| \cdot |V_j|}{p}}{\sqrt{|V_i| \cdot |V_j|} - \frac{|V_i| \cdot |V_j|}{p}}.$$

This is what stabilityIntersectionMBM, stabilityIntersectionGreedy, stabilityIntersectionCount and stabilityIntersectionMean become, when there are no similar features.

Value

`numeric(1)` Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1, \ldots, V_m$ denote the sets of chosen features for the m datasets, i.e. `features` has length m and V_i is a set which contains the i-th entry of `features`. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^{p} h_j = \sum_{i=1}^{m} |V_i|$ and $V = \bigcup_{i=1}^{m} V_i$.

References

Bommert A, Rahnenführer J (2020). "Adjusted Measures for Feature Selection Stability for Data Sets with Similar Features." In *International Conference on Machine Learning, Optimization, and Data Science - LOD 2020, in print*. `https://arxiv.org/abs/2009.12075`.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
stabilityUnadjusted(features = feats, p = 10)
```

---

stabilityWald *Stability Measure Wald*

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilityWald(features, p, impute.na = NULL)
```

Arguments

features	`list (length >= 2)` Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).
p	`numeric(1)` Total number of features in the datasets.
impute.na	`numeric(1)` In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? `NULL` means no imputation.

Details

The stability measure is defined as (see Notation)

$$\frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{|V_i \cap V_j| - \frac{|V_i| \cdot |V_j|}{p}}{\min\{|V_i|, |V_j|\} - \frac{|V_i| \cdot |V_j|}{p}}.$$

Value

`numeric(1)` Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1, \ldots, V_m$ denote the sets of chosen features for the m datasets, i.e. `features` has length m and V_i is a set which contains the i-th entry of `features`. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^{p} h_j = \sum_{i=1}^{m} |V_i|$ and $V = \bigcup_{i=1}^{m} V_i$.

References

Wald R, Khoshgoftaar TM, Napolitano A (2013). "Stability of Filter- and Wrapper-Based Feature Subset Selection." In *2013 IEEE 25th International Conference on Tools with Artificial Intelligence*. doi: 10.1109/ictai.2013.63.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
stabilityWald(features = feats, p = 10)
```

`stabilityYu` *Stability Measure Yu*

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilityYu(
  features,
  sim.mat,
  threshold = 0.9,
  correction.for.chance = "estimate",
  N = 10000,
  impute.na = NULL
)
```

Arguments

`features`
: `list (length >= 2)`
Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).

`sim.mat`
: `numeric matrix`
Similarity matrix which contains the similarity structure of all features based on all datasets. The similarity values must be in the range of [0, 1] where 0 indicates very low similarity and 1 indicates very high similarity. If the list elements of `features` are integerish vectors, then the feature numbering must correspond to the ordering of `sim.mat`. If the list elements of `features` are character vectors, then `sim.mat` must be named and the names of `sim.mat` must correspond to the entries in `features`.

`threshold`
: `numeric(1)`
Threshold for indicating which features are similar and which are not. Two features are considered as similar, if and only if the corresponding entry of `sim.mat` is greater than or equal to `threshold`.

`correction.for.chance`
: `character(1)`
How should the expected value of the stability score (see Details) be assessed? Options are "estimate", "exact" and "none". For "estimate", `N` random feature sets of the same sizes as the input feature sets (`features`) are generated. For "exact", all possible combinations of feature sets of the same sizes as the input feature sets are used. Computation is only feasible for very small numbers of features and numbers of considered datasets (`length(features)`). For "none", the transformation $(score - expected)/(maximum - expected)$ is not conducted, i.e. only $score$ is used. This is not recommended.

`N`
: `numeric(1)`
Number of random feature sets to consider. Only relevant if `correction.for.chance` is set to "estimate".

`impute.na`
: `numeric(1)`
In some scenarios, the stability cannot be assessed based on all feature sets. E.g. if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? `NULL` means no imputation.

Details

Let O_{ij} denote the number of features in V_i that are not shared with V_j but that have a highly simlar feature in V_j:

$$O_{ij} = |\{x \in (V_i \backslash V_j) : \exists y \in (V_j \backslash V_i) \ with \ Similarity(x, y) \geq threshold\}|.$$

Then the stability measure is defined as (see Notation)

$$\frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{I(V_i, V_j) - E(I(V_i, V_j))}{\frac{|V_i|+|V_j|}{2} - E(I(V_i, V_j))}$$

with
$$I(V_i, V_j) = |V_i \cap V_j| + \frac{O_{ij} + O_{ji}}{2}.$$
Note that this definition slightly differs from its original in order to make it suitable for arbitrary datasets and similarity measures and applicable in situations with $|V_i| \neq |V_j|$.

Value

`numeric(1)` Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1, \ldots, V_m$ denote the sets of chosen features for the m datasets, i.e. `features` has length m and V_i is a set which contains the i-th entry of `features`. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^{p} h_j = \sum_{i=1}^{m} |V_i|$ and $V = \bigcup_{i=1}^{m} V_i$.

References

Yu L, Han Y, Berens ME (2012). "Stable Gene Selection from Microarray Data via Sample Weighting." *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **9**(1), 262–272. doi: 10.1109/tcbb.2011.47. Zhang M, Zhang L, Zou J, Yao C, Xiao H, Liu Q, Wang J, Wang D, Wang C, Guo Z (2009). "Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes." *Bioinformatics*, **25**(13), 1662–1668. doi: 10.1093/bioinformatics/btp295.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
mat = 0.92 ^ abs(outer(1:10, 1:10, "-"))
stabilityYu(features = feats, sim.mat = mat, N = 1000)
```

`stabilityZucknick` *Stability Measure Zucknick*

Description

The stability of feature selection is defined as the robustness of the sets of selected features with respect to small variations in the data on which the feature selection is conducted. To quantify stability, several datasets from the same data generating process can be used. Alternatively, a single dataset can be split into parts by resampling. Either way, all datasets used for feature selection must contain exactly the same features. The feature selection method of interest is applied on all of the datasets and the sets of chosen features are recorded. The stability of the feature selection is assessed based on the sets of chosen features using stability measures.

Usage

```
stabilityZucknick(
  features,
  sim.mat,
  threshold = 0.9,
  correction.for.chance = "none",
  N = 10000,
  impute.na = NULL
)
```

Arguments

`features` — `list` (length >= 2)
Chosen features per dataset. Each element of the list contains the features for one dataset. The features must be given by their names (`character`) or indices (`integerish`).

`sim.mat` — `numeric matrix`
Similarity matrix which contains the similarity structure of all features based on all datasets. The similarity values must be in the range of [0, 1] where 0 indicates very low similarity and 1 indicates very high similarity. If the list elements of `features` are integerish vectors, then the feature numbering must correspond to the ordering of `sim.mat`. If the list elements of `features` are character vectors, then `sim.mat` must be named and the names of `sim.mat` must correspond to the entries in `features`.

`threshold` — `numeric(1)`
Threshold for indicating which features are similar and which are not. Two features are considered as similar, if and only if the corresponding entry of `sim.mat` is greater than or equal to `threshold`.

`correction.for.chance` — `character(1)`
Should a correction for chance be applied? Correction for chance means that if features are chosen at random, the expected value must be independent of the number of chosen features. To correct for chance, the original score is transformed by $(score - expected)/(maximum - expected)$. For stability measures whose score is the average value of pairwise scores, this transformation is done for all components individually. Options are "none", "estimate" and "exact". For "none", no correction is performed, i.e. the original score is used. For "estimate", `N` random feature sets of the same sizes as the input feature sets (`features`) are generated. For "exact", all possible combinations of feature sets of the same sizes as the input feature sets are used. Computation is only feasible for very small numbers of features (p) and numbers of considered datasets (`length(features)`).

`N` — `numeric(1)`
Number of random feature sets to consider. Only relevant if `correction.for.chance` is set to "estimate".

`impute.na` — `numeric(1)`
In some scenarios, the stability cannot be assessed based on all feature sets. E.g.

if some of the feature sets are empty, the respective pairwise comparisons yield NA as result. With which value should these missing values be imputed? `NULL` means no imputation.

Details

The stability measure is defined as

$$\frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{|V_i \cap V_j| + C(V_i, V_j) + C(V_j, V_i)}{|V_i \cup V_j|}$$

with

$$C(V_k, V_l) = \frac{1}{|V_l|} \sum_{(x,y) \in V_k \times (V_l \setminus V_k) \ with \ Similarity(x,y) \geq threshold} Similarity(x, y).$$

Note that this definition slightly differs from its original in order to make it suitable for arbitrary similarity measures.

Value

`numeric(1)` Stability value.

Notation

For the definition of all stability measures in this package, the following notation is used: Let $V_1, \ldots, V_m$ denote the sets of chosen features for the m datasets, i.e. `features` has length m and V_i is a set which contains the i-th entry of `features`. Furthermore, let h_j denote the number of sets that contain feature X_j so that h_j is the absolute frequency with which feature X_j is chosen. Analogously, let h_{ij} denote the number of sets that include both X_i and X_j. Also, let $q = \sum_{j=1}^{p} h_j = \sum_{i=1}^{m} |V_i|$ and $V = \bigcup_{i=1}^{m} V_i$.

References

Bommert A, Rahnenführer J, Lang M (2017). "A Multicriteria Approach to Find Predictive and Sparse Models with Stable Feature Selection for High-Dimensional Data." *Computational and Mathematical Methods in Medicine*, **2017**, 1–18. doi: 10.1155/2017/7907163. Zucknick M, Richardson S, Stronach EA (2008). "Comparing the Characteristics of Gene Expression Profiles Derived by Univariate and Multivariate Classification Methods." *Statistical Applications in Genetics and Molecular Biology*, **7**(1). doi: 10.2202/15446115.1307.

See Also

listStabilityMeasures

Examples

```
feats = list(1:3, 1:4, 1:5)
mat = 0.92 ^ abs(outer(1:10, 1:10, "-"))
stabilityZucknick(features = feats, sim.mat = mat)
```

Index