Package ‘morphonode’

May 31, 2022

Title The Morphonode Predictive Model for ultrasound signatures detection and malignancy prediction in vulvar cancer

Version 1.0.0

Date 2022-05-23

Maintainer Fernando Palluzzi <fernando.palluzzi@gmail.com>

Description The R package morphonode is an ensemble predictor for malignancy prediction and patient stratification in four signatures. The package uses 14 ultrasound features, processed by four modules: (i) malignancy prediction through random forest classifiers (Morphonode-RFC), (ii) malignancy risk estimation via robust binomial modeling (Morphonode-RBM), (iii) signature detection using decision trees (Morphonode-DT), and (iv) search for similar ultrasound profiles (Morphonode-SP). The methodology, referred to as Morphonode Predictive Model, is described in Fragomeni et al. <https://doi.org/0000/0000>.

URL https://github.com/Morphonodepredictivemodel/morphonode

Depends R (>= 4.0)

Imports Biobase (>= 2.52.0), BiocGenerics (>= 0.38.0), boot (>= 1.3-25), CMA (>= 1.50.0), dplyr (>= 1.0.7), ggformula (>= 0.10.1), ggplot2 (>= 3.3.5), ggstance (>= 0.3.5), ggridges (>= 0.5.3), imputeR (>= 2.2), lattice (>= 0.20-45), lsa (>= 0.73.2), MASS (>= 7.3-57), Matrix (>= 1.4-1), mosaic (>= 1.8.3), mosaicData (>= 0.20.2), randomForest (>= 4.6-14)

License GPL-3

LazyData true

Encoding UTF-8

NeedsCompilation no

RoxygenNote 7.1.2

R topics documented:

boot.se *Compute bootstrap standard errors*

Description

Compute bootstrap standard errors (SE) for a given (generalized) linear model.

Usage

```
boot.se(fit, boot, probs = c(0.025, 0.975), z0 = 1.96, b0 = 0, ...)
```

Arguments

`fit` An object of class `glm` or `lm`.

`boot` A bootstrapped model object (see the examples section).

`probs` Bootstrap intervals (default = c(0.025, 0.975)).

`z0` z-score under null hypothesis (default = 1.96).

`b0` Effect size under null hypothesis (default = 0).

`...` Currently ignored.

Value

A data.frame of 9 columns:

1. "Variable", variable name;
2. "Estimate", parameter (effect size) estimate;
3. "se.boot", bootstrap standard error;
4. "lower", confidence interval lower bound;
5. "upper", confidence interval upper bound;
6. "conf.level", confidence level;
7. "method", estimation method;
8. "z", z-score = (estimate - b0)/SE;
9. "P", 2-sided p-value; i.e., 2*pnorm(-abs(z)).

Author(s)

Fernando Palluzzi <`fernando.palluzzi@gmail.com`>

References

Pruim R, Kaplan DT, Horton NJ (2017). The mosaic Package: Helping Students to ’Think with Data’ Using R. The R Journal, 9(1), 77–102. <https://journal.r-project.org/archive/2017/RJ-2017-024/index.html>

See Also

See `p.boot` for performance indices bootstrap confidence intervals. See also `do` for resampling.

Examples

```
# Dichotomized dataset creation
x <- dichotomize(mpm.us, asFactor = TRUE)

# Model specification
model <- formula(paste0(c("y ~ shortAxis + cortical + hilum + ",
                          "inflammatoryStroma + extracapsularSpread + ",
                          "ecostructure + FID + VFL + corticalThickening + ",
                          "vascularPattern + CMID + shape + grouping + ",
                             "colorScore"), collapse = ""))

# Binomial model fitting (MLE)
fit <- glm(model, data = x, family = "binomial")

# Binomial model fitting
n.reps <- 100
boot <- mosaic::do(n.reps) * coef(glm(model, data = mosaic::resample(x),
                                      family = "binomial"))

# Bootstrap SE calculation
SE <- boot.se(fit, boot)
print(SE)
```

brier *Compute Brier scores*

Description

Compute Brier scores for a given dataset.

Usage

```
brier(data, status, k = 5, method = "CV", ...)
```

Arguments

`data`	An (n, m) matrix or data.frame with no outcome attribute.
`status`	A vector of length n, containing the outcome.
`k`	number of cross-validation iterations (default = 5).
`method`	One of the `GenerateLearningsets` methods, including: "LOOCV", "CV", "MCCV", and "bootstrap" (default = "CV").
`...`	Currently ignored.

Value

A vector of Brier scores of length n (one value per subject).

Author(s)

Fernando Palluzzi <`fernando.palluzzi@gmail.com`>

References

Slawski M, Daumer M, Boulesteix AL. CMA - a comprehensive Bioconductor package for supervised classification with high dimensional data. BMC Bioinformatics 9, 439 (2008). <https://doi.org/10.1186/1471-2105-9-439>

Brier GW. Verification of forecasts expressed in terms of probability. Monthly Weather Review. 1950;78(1):1-3. <https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2>

See Also

`loss`

Examples

```
# Extract a subset of 300 subjects and an outcome vector of length 30
# from the default simulated dataset

x <- mosaic::sample(mpm.us, 300, replace = FALSE, prob = NULL)
y <- x$y
x <- x[, 2:15]
print(dim(x))
print(length(y))

# Compute brier scores
E <- brier(x, y)
print(quantile(E))
```

`buildPredictor` *Build a randomForest object*

Description

Build a Random Forest Classifier (RFC) of class `randomForest`.

Usage

```
buildPredictor(model, data, n = 10000, m = 3, vset = NULL, ...)
```

Arguments

`model` An R formula of shape y ~ x1 + x2 + ... + xn. You can use `model <- formula("y ~ .")` to include all the variables of the input dataset.

`data` A matrix or data.frame with subjects as rows and ultrasound features as columns. The outcome should be named "y".

`n` Number of trees to be generated by bootstrap.

`m` Number of randomly sampled variables per tree branching.

`vset` An optional data.frame of the same shape as the input data. This will be used as external validation set. A validation set can be also extracted from the input dataset, using the `vpart` function.

`...` Currently ignored.

Details

This is the class of RFC used in the us.predict and rfc.predict fuctions. The default randomForest object in the morphonode package is an ensemble of 5 RFCs trained over a simulated dataset of 948 subjects (508 non-malignant and 440 malignant profiles), using a nested 5-fold cross-validation scheme. The training/validation details and performances of this dataset are stored in the mpm.rfc object.

Value

A list of 2 objects:

1. "RFC", an object of class randomForest;
2. "performance", a list containing RFC performances.

Author(s)

Fernando Palluzzi <fernando.palluzzi@gmail.com>

References

Fragomeni SM, Moro F, Palluzzi F, Mascilini F, Rufini V, Collarino A, Inzani F, Giacò L, Scambia G, Testa AC, Garganese G (2022). Evaluating the risk of inguinal lymph node metastases before surgery using the Morphonode Predictive Model: a prospective diagnostic study. Ultrasound xx Xxxxxxxxxx xxx Xxxxxxxxxxx. 00(0):000-000. <https://doi.org/00.0000/00000000000000000000>

Liaw A, Wiener M. Classification and Regression by randomForest (2002). R News, 2(3):18-22. <https://doi.org/10.1023/A:1010933404324>

See Also

us.predict, rfc.predict, vpart

Examples

```
# Extract a subset of 500 subjects and an outcome vector of length 30
# from the default simulated dataset

x <- mosaic::sample(mpm.us, 500, replace = FALSE, prob = NULL)
x <- x[, 2:16]
model <- formula("y ~ .")

# Data partitioning (75% training set, 25% validation set)

x <- vpart(x, p = 0.75)

# RFC building (1000 bootstrapped trees, 3 random variables per split)

rfc <- buildPredictor(model, x$training.set, n = 1000,
                      vset = x$validation.set)
print(rfc$performance)
```

```
# RFC building (10000 bootstrapped trees, 3 random variables per split)

rfc <- buildPredictor(model, x$training.set, vset = x$validation.set)
print(rfc$performance)
```

`check.rfcdata` *Check RFC input data*

Description

Assign ultasound data types for RFC building and validation.

Usage

```
check.rfcdata(x, levels = NULL, ...)
```

Arguments

`x`	An (n, 14) ultrasound features data.frame, where n is the number of subjects.
`levels`	A list of length 15, corresponding to the levels of each ultrasound variable plus the phenotype (y = 0, 1). Needed for categorical variables (factors) checking; for continuous variables, it should assume the nominal value of 0. If NULL (default), the `mpm.levels` object will be used.
`...`	Currently ignored.

Value

A data.frame of ultrasound features with checked data type.

Author(s)

Fernando Palluzzi <`fernando.palluzzi@gmail.com`>

See Also

`buildPredictor`.

Examples

```
# generate a dataset of 500 subjects
x <- mosaic::sample(mpm.us, 500, replace = FALSE, prob = NULL)
x <- x[, 2:16]

# Assign the correct data types for RFC building and validation
x <- check.rfcdata(x)
summary(x)
```

dichotomize *Generates a dichotomous ultrasound feature data.frame*

Description

Convert an ultrasound feature data.frame (each row is an ultrasound vector) into a dichotomous data.frame. Defalut dichotomization cutoffs are computed as described in Fragomeni et al. (2022).

Usage

```
dichotomize(
  x,
  short = 8,
  cortical = 2,
  ist = 1,
  ecs = 1,
  hab = 0,
  eco = 1,
  vp = c(1, 2, 3),
  vfl = c(2, 3, 4),
  ct = c(2),
  fid = c(1, 2, 3),
  cmid = c(2, 3, 4),
  shape = c(3),
  cs = 3,
  grouping = c(2, 3),
  asFactor = FALSE,
  ...
)
```

Arguments

`x` An (n, 14) ultrasound features data.frame, where n is the number of subjects.

`short` Numeric value corresponding to the short axis cutoff in millimeters (default = 8).

`cortical` Numeric value corresponding to the cortical thickness cutoff in millimeters (default = 2).

`ist` Dichotomous value 0, 1 for the presence of inflammatory stroma (perinodal hyperechogenic ring; default = 1).

`ecs` Dichotomous value 0, 1 for the presence of extracapsular spread (cortical interruption; default = 1).

`hab` Dichotomous value 0, 1 for the absence of the hilum (nodal core sign; default = 0).

`eco` Dichotomous value 0, 1 for heterogeneous echogenicity (echostructure; default = 1).

`vp` Categorical value (integers from 0 to 4) associated to a high-risk vascular flow architecture pattern (default = c(1, 2, 3)).

`vfl` Categorical value (integers from 0 to 4) associated to a high-risk vascular flow localization (default = c(2, 3, 4)).

`ct` Categorical value (integers from 0 to 4) associated to a high-risk cortical thickening (default = 2).

`fid` Categorical value (integers from 0 to 3) associated to a high-risk focal intranodal deposit (default = c(1, 2, 3)).

`cmid` Categorical value (integers from 0 to 4) associated to a high-risk cortical-medullar interface distortion (default = c(2, 3, 4)).

`shape` Categorical value (integers from 1 to 3) associated to a high-risk shape (default = 3).

`cs` Ordinal value (integers from 1 to 5) associated to a high-risk color score (default = 3).

`grouping` Categorical value (integers from 1 to 3) associated to a high-risk grouping (default = c(2, 3)).

`asFactor` Logical value. If TRUE, data.frame columns are converted to factors (default = FALSE).

`...` Currently ignored.

Details

This function is internally used to estimate the malignancy risk through the robust binomial model (Morphonode-RBM). Dichotomization is performed to avoid the extremely low frequancy of some levels in categorical variables.

Value

An ultrasound profile with imputed missing values.

Author(s)

Fernando Palluzzi <`fernando.palluzzi@gmail.com`>

References

Fragomeni SM, Moro F, Palluzzi F, Mascilini F, Rufini V, Collarino A, Inzani F, Giacò L, Scambia G, Testa AC, Garganese G (2022). Evaluating the risk of inguinal lymph node metastases before surgery using the Morphonode Predictive Model: a prospective diagnostic study. Ultrasound xx Xxxxxxxxxx xxx Xxxxxxxxxx. 00(0):000-000. <https://doi.org/00.0000/00000000000000000000>

See Also

See `uss` for metastatic risk signature detection (Morphonode-DT module).

Examples

```
# Create a dichotomous version of subjects with metastatic signature,
# from the default simulated dataset.

M <- dichotomize(mpm.us[mpm.us$signature == "MET", 2:15])
print(head(M))
```

`euclidean` *Euclidean distance*

Description

Compute the euclidean distance between two vectors.

Usage

```
euclidean(x, y, ...)
```

Arguments

`x`	A numeric vector.
`y`	A numeric vector.
`...`	Currently ignored.

Value

A numeric value corresponding to the euclidean distrance.

Author(s)

Fernando Palluzzi <`fernando.palluzzi@gmail.com`>

See Also

`similarity`, `jaccard`

Examples

```
# Sample two random ultrasound profiles from the default dataset
x <- mosaic::sample(mpm.us, 1, replace = FALSE, prob = NULL)
x <- as.matrix(x[, 2:15])
y <- mosaic::sample(mpm.us, 1, replace = FALSE, prob = NULL)
y <- as.matrix(y[, 2:15])

# Compute the euclidean distance
d <- euclidean(x, y)
print(d)
```

---

imp.missing *Impute missing values*

Description

Wrapper for the imputeR function impute.

Usage

```
imp.missing(M, x = NULL, mode = NULL, ...)
```

Arguments

M	A matrix or data.frame containing missing values.
x	An optional vector that will be attached to M. This can be useful if data with missing values can be attached to a reference dataset.
mode	Either "cat" (cateorical variables) or "con" (continuous variables).
...	Currently ignored.

Value

A data.frame with imputed missing values.

Author(s)

Fernando Palluzzi <fernando.palluzzi@gmail.com>

Examples

```
# Sample 30 subjects from the morphonode simulated dataset
data <- mosaic::sample(mpm.us, 30, replace = FALSE, prob = NULL)[, 2:15]

# Entries with missing values
missing <- matrix(c(10.0, 6.3, 1, 0, 0, 0, 0, 1, NA, 2, NA, 2, 3, NA,
                     6.4, 2.1, 1, 0, 0, 0, 0, 1, NA, 2, NA, 1, 1, NA),
                  nrow = 2, byrow = TRUE)
```

```
colnames(missing) <- colnames(mpm.us[, 2:15])

# Defining categorical subset
data.cat <- data.frame(apply(mpm.us[, 2:15], 2, factor))

# Imputing missing values
data.cat <- imp.missing(data.cat, x = missing, mode = "cat")
print(tail(data.cat))
```

jaccard *Jaccard similarity*

Description

Compute Jaccard similarity between two vectors.

Usage

```
jaccard(x, y, ...)
```

Arguments

x	A dichotomous vector.
y	A dichotomous vector.
...	Currently ignored.

Value

A numeric value corresponding to the Jaccard similarity.

Author(s)

Fernando Palluzzi <fernando.palluzzi@gmail.com>

See Also

similarity, euclidean

Examples

```
# Sample two random ultrasound profiles from the default dataset
x <- mosaic::sample(mpm.us, 1, replace = FALSE, prob = NULL)
x <- as.matrix(dichotomize(x[, 2:15]))
y <- mosaic::sample(mpm.us, 1, replace = FALSE, prob = NULL)
y <- as.matrix(dichotomize(y[, 2:15]))

# Compute the euclidean distance
r <- jaccard(x, y)
```

```
print(r)
```

loss — *Loss function*

Description

Compute the loss function needed for prediction error estimation.

Usage

```
loss(p, y, method = "log", ...)
```

Arguments

`p`	A vector of predicted outcome values.
`y`	A vector of observed outcome values.
`method`	Loss function definition. One between "log" (default) and "sqerror".
`...`	Currently ignored.

Details

The subject-level prediction error, calculated through the `topsim` function, is strongly correlated with the loss function values. However, while the former can be only calculated for the entire training set, the latter can be computed for the new input profile(s). Therefore, the prediction error (E) for the input is calculated as: E = b0 + b*L, where L is the loss function value. L can be currently defined as either "log": L = -1*(y*log(p) + (1 - y)*log(1 - p)) (default), or "sqerror": L = (p - y)^2. Additionally, the cost is calculated as either average loss sum(L)/n or root mean squared error sqrt(sum(L)/n), for "log" and "sqerror", respectively.

Value

A list of 2 objects:

1. "loss", loss function values;
2. "cost", cost function value.

Author(s)

Fernando Palluzzi <`fernando.palluzzi@gmail.com`>

See Also

`us.predict`, `brier`

Examples

```
# RBM prediction vs. reality over the simulated dataset

p <- predict(mpm.rbm$fit, dichotomize(mpm.us[2:15], asFactor = TRUE),
             type = "response")

L <- loss(p, mpm.us$y)

print(quantile(L$loss))

print(L$cost)

# Overall RBM performances

y.hat <- ifelse(p > 0.5, 1, 0)
P <- performance(obs = mpm.us$y, pred = y.hat)

print(P)
```

mpm.levels *Ultrasound features levels*

Description

A list of 14 elements containing the levels of each ultrasound feature.

Usage

```
mpm.levels
```

Format

"mpm.levels" is a list of 14 vectors, each containing the levels of an ultrasound feature. Continuous features (i.e., short axis and cortical thickness) have the nominal value of 0. Levels are used by predictive functions to switch from numeric to factor classes.

Examples

```
# Generate a simulated malignant ultrasound profile
x <- new.profile(us.simulate(y = 1))

# Prepare the profile for RFC malignancy prediction
u <- set.rfcdata(x, ref = mpm.us[, 2:15], levels = mpm.levels)
print(u)

# Predict subject's phenotype
P <- rfc.predict(u$ultrasound, rfc = mpm.rfc$rfc)
print(P)
```

`mpm.rbm` *Morphonode default Robust Binomial Model (RBM) endemble*

Description

Logistic model generated by fitting a simulated dataset of 948 ultrasound profiles (440 malignant and 508 non-malignant), with robust bootstrap standard errors estimation (5000 bootstrap iterations), through the internal morphonode function `boot.se`. The input ultrasound feature dataset was dichotomized before model fitting (see `dichotomize`) to avoid parameter estimation biases due to very low frequency of the levels of some categorical ultrasound features. The RBM is used to estimate the malignancy risk, (R) providing a continuous measure of malignancy (in contrast to the dichotomous prediction of the RFC ensemble). Considering the expected simulated phenotype (y) as the ground truth, two optimal malignancy risk cutoffs were estimated, defining three risk levels: low (R < 0.23), moderate (0.23 <= R <= 0.29), and high (R > 0.29).

Usage

```
mpm.rbm
```

Format

"mpm.rbm" is a list of 3 objects:

1. "coef", a data.frame reporting bootstrap-based estimations: ultrasound feature (Variable), log(odds ratio) (Estimate), bootstrap standard errors (se.boot), confidence interval lower bound (lower), confidence interval upper bound (upper), confidence level (conf.level), bootstrap estimation method (method), z-score (z), 2-sided p-value (P);
2. "model", `R` formula representing the fitted model;
3. "fit", MLE-based fitted model object of class `glm`.

References

Fragomeni SM, Moro F, Palluzzi F, Mascilini F, Rufini V, Collarino A, Inzani F, Giacò L, Scambia G, Testa AC, Garganese G (2022). Evaluating the risk of inguinal lymph node metastases before surgery using the Morphonode Predictive Model: a prospective diagnostic study. Ultrasound xx Xxxxxxxxxx xxx Xxxxxxxxxx. 00(0):000-000. <https://doi.org/00.0000/00000000000000000000>

Examples

```
# Create a simulated malignant ultrasound profile
x <- new.profile(us.simulate(y = 1))

# Lauch the Morhonode Predictive Model
u <- us.predict(x)
```

`mpm.rfc` *Morphonode default Random Forest Classifier (RFC) endemble*

Description

RFC ensemble of 5 classifiers, based on 948 simulated ultrasound profiles (440 malignant and 508 non-malignant). The simulated dataset is divided into 5 random subsets and a nested 5-fold cross-validation (CV) is performed. For each CV cycle, 4/5 partitions are used as training set and the last one as validation set. Each RFC is trained over 10000 bootstrap trees, with 3/14 randomly chosen variables per tree branching. Bootstrapping enables independent prediction error estimation, using out-of-bag (OOB) samples. OOB error estimation allows the claculation of mean decrease accuracy (MDA) and mean decrease in Gini impurity (MDG). measuring ultrasound feature-level contribution in RFC predictive accuracy and classification entropy, respectively. These measures enable ultrasound feature ranking, based on the average of minmax-normalized MDA and MDG values (i.e., the most important feature scores 100, while the least important tends to 0). Each CV cycle provides a dichotomous phenotype classification in malignant (y = 1) and non-malignant (y = 0), OOB prediction error estimation, and subject-level estimation of the prediction uncertainty through Brier scores calculation. Given a new ultrasound profile, the resulting 5 RFCs yield independent predictions and the majority wins, whith higher priority to the RFCs with smaller OOB error.

Usage

```
mpm.rfc
```

Format

"mpm.rfc" is a list of 5 objects:

1. "training", a list of 5 data.frames (T1-5) corresponding to the 5 RFC training sets;
2. "validation", a list of 5 data.frames (V1-5) corresponding to the 5 RFC validation sets;
3. "rfc", a list of 5 `randomForest` objects, corresponding to the 5 classifiers of the ensemble;
4. "ranking", a data.frame reporting MDA and MDG values, as well as their minmax-normalized values (fA and fG, respectively), and the overall ranking score (f) being the average of fA and fG;
5. "performance", a list of 7 values summarizing the RFC ensemble performances, including: 2x2 contingency table, sensitivity, specificity, precision (PPV), negative predictive value (NPV), F1 score, and predictive accuracy.

References

Fragomeni SM, Moro F, Palluzzi F, Mascilini F, Rufini V, Collarino A, Inzani F, Giacò L, Scambia G, Testa AC, Garganese G (2022). Evaluating the risk of inguinal lymph node metastases before surgery using the Morphonode Predictive Model: a prospective diagnostic study. Ultrasound xx Xxxxxxxxxx xxx Xxxxxxxxxx. 00(0):000-000. <https://doi.org/00.0000/0000000000000000000>

Examples

```
# Create a simulated malignant ultrasound profile
x <- new.profile(us.simulate(y = 1))

# Lauch the Morhonode Predictive Model
u <- us.predict(x)
```

`mpm.us` *Morphonode simulated ultrasound feature dataset*

Description

A simulated dataset of 948 ultrasound profiles (440 malignant and 508 non-malignant) used to build the default ensemble random forest classifier (RFC) and the robust binomial model (RBM). This dataset was generated as a 4-fold expansion of the original ultrasound feature dataset of 237 groin samples (75 malignant and 162 non-malignant) from Fragomeni et al. (2022), using the morphonode function `us.simulate`.

Usage

```
mpm.us
```

Format

"mpm.us" is a data.frame of 948 rows (simulated ultrasound profiles) and 18 columns, including: a progressive number used as unique profile identifier (ID), 14 ultrasound features used for RFC and RBM building, expected simulation phenotype used as ground truth (y = 0: non-malignant, 1: malignant), metastatic risk signature associated to each simulated ultrasound profile (signature), subject-level prediction error calculated as Brier score (E).

References

Fragomeni SM, Moro F, Palluzzi F, Mascilini F, Rufini V, Collarino A, Inzani F, Giacò L, Scambia G, Testa AC, Garganese G (2022). Evaluating the risk of inguinal lymph node metastases before surgery using the Morphonode Predictive Model: a prospective diagnostic study. Ultrasound xx Xxxxxxxxxx xxx Xxxxxxxxxx. 00(0):000-000. <https://doi.org/00.0000/00000000000000000000>

Examples

```
# Default simulated dataset phenotype frequencies
print(table(mpm.us$y))

# Default simulated dataset signatures frequencies
print(table(mpm.us$signature))

# Default simulated dataset prediction error quartiles
print(quantile(mpm.us$E))
```

`new.profile` *Create a new ultrasound profile*

Description

Creates a new ultrasound profile, used as input for the `us.predict` function.

Usage

```
new.profile(u = NULL, ...)
```

Arguments

`u` A vector of length 14, corresponding to the input ultrasound profile. The order of the elements in u must be: short axis [numeric], cortical thickness [numeric], nodal core sign (hilum) [dichotomous], perinodal hyperechogenic ring [dichotomous], cortical interruption [dichotomous], echogenicity [dichotomous], focal intranodal deposit [categorical], vascular flow localization [categorical], cortical thickening [categorical], vascular flow architecture pattern [categorical], cortical-medullar interface distortion [categorical], shape [categorical], grouping [categorical], color score [ordinal]. Value -1 can be entered for missing values. If u = `NULL`, a guided interactive prompt will be launched. In the interactive mode, typing return without entering any value will introduce a missing value (this is not possible for short axis and cortical thickness, since they are necessary for a reliable prediction).

`...` Currently ignored.

Value

A list of 2 objects:

1. "ultrasound", ultrasound features vector;
2. "missing", missing values vector (empty, if no missing values are found).

Author(s)

Fernando Palluzzi `<fernando.palluzzi@gmail.com>`

See Also

See `show.mpmenv` for ultrasound variable description. See also `us.predict` to launch the full Morphonode Predictive Model suite.

Examples

```
### Profile with missing data

u <- new.profile(c(10.0, 6.3, 1, 0, 0, 0, -1, 1, 2, 2, 3, -1, -1, -1))
print(u)

### New profile from simulated data

# High metastatic risk profile
u.hmr <- new.profile(us.simulate(signature = "HMR"))
print(u.hmr)

# Low metastatic risk profile
u.lmr <- new.profile(us.simulate(signature = "LMR"))
print(u.lmr)

# Malignant profile
u1 <- new.profile(us.simulate(y = 1))
print(u1)

# Non-malignant profile
u0 <- new.profile(us.simulate(y = 0))
print(u0)
```

p.boot *Bootstrap confidence interfals of classifier performances*

Description

Compute bootstrap confidence intervals for various classifier performance indices.

Usage

```
p.boot(x, rep = 5000, ci.method = "bca", formula = "f1", ...)
```

Arguments

`x` An data.frame of n rows (subjects) and 2 columns: the first column contains the predicted values for an outcome y (i.e., y.hat), and the second column contains the observed ("true") values for y.

`rep` Number of bootstrap iterations (default = 5000). A high number of iterations is needed for reliable estimations. If `rep` < 5000, estimation errors might occur.

`ci.method` Method used for bootstrap confidence interval estimation (default = "bca"; i.e., adjusted bootstrap percentile).

`formula` Performance index. One among: "f1" (default), "accuracy", "sensitivity", "specificity", "ppv" (or "precision"; i.e., positive predictive value), "npv" (negative predictive value), "plr" (positive likelihood ratio), "nlr" (hegative likelihood ratio), "fpr" (false positive rate), "fdr" (false discovery rate), "fnr" (flase negative rate), "fnc" (false negative cost = FN/(TP + TN), see Fragomeni et al. 2022).

`...` Currently ignored.

Value

A list of 2 objects:

1. "boot", an object of class `boot`, containing bootstrap replicates of the given statistic;
2. "ci", an object of class `bootci`, containing bootstrap confidence interval estimations.

Author(s)

Fernando Palluzzi <`fernando.palluzzi@gmail.com`>

References

Fragomeni SM, Moro F, Palluzzi F, Mascilini F, Rufini V, Collarino A, Inzani F, Giacò L, Scambia G, Testa AC, Garganese G (2022). Evaluating the risk of inguinal lymph node metastases before surgery using the Morphonode Predictive Model: a prospective diagnostic study. Ultrasound xx Xxxxxxxxxx xxx Xxxxxxxxxx. 00(0):000-000. <https://doi.org/00.0000/000000000000000000000>

Davison AC and Hinkley DV (1997). Bootstrap Methods and Their Application, Chapter 5. Cambridge University Press.

DiCiccio TJ and Efron B (1996). Bootstrap confidence intervals (with Discussion)._Statistical Science_, *11*, 189-228.

Efron B (1987). Better bootstrap confidence intervals (with Discussion). Journal of the American Statistical Association, *82*, 171-200. <https://doi.org/10.2307/2289144>

See Also

See `boot.ci` for further details.

Examples

```
# Predicted ultrasound phenotype from the default morphonode RFC1
y.hat <- predict(mpm.rfc$rfc$RFC1, mpm.rfc$validation$V1)

# Actual ultrasound phenotype values
y <- mpm.rfc$validation$V1$y

# Input preparation
Y <- data.frame(y.hat, y)

# F1 score bootstrap confidence intervals
```

```
F1 <- p.boot(Y)
print(F1$boot$t0)              # F1 score observed value
print(F1$ci$bca[4:5])          # F1 score bca confidence interval

# Accuracy bootstrap confidence intervals
A <- p.boot(Y, formula = "accuracy")
print(A$boot$t0)               # Accuracy observed value
print(A$ci$bca[4:5])           # Accuracy bca confidence interval
```

performance *Predictor performance calculation*

Description

Compute sensitivity, speciticity, precision, F1 score, and accuracy for a set of predictions.

Usage

```
performance(obs = NULL, pred = NULL, C = NULL, y = "0,1", ...)
```

Arguments

`obs`	A vector of observed values.
`pred`	A vector of predicted values.
`C`	2x2 contingency table (alternative to `obs` and `pred`). If `obs` and `pred` are given, the contingency table is automatically computed. The contingency table has the observed values as rows and the predicted ones as columns. By default, true negatives are located at position [1, 1], while true positives at [2, 2] (see below).
`y`	Contingenty table orientation. If `y = "0,1"` (default), true negatives are located at position [1, 1], while true positives at position [2, 2]. If `y = "1,0"`, these positions are inverted.
`...`	Currently ignored.

Value

A list of 6 objects:

1. "ctable", 2x2 contingency table of predicted vs. observed values;
2. "sensitivity", Se = TP/(TP + FN);
3. "specificity", Sp = TN/(TN + FP);
4. "precision", PPV = TP/(TP + FP), also called "positive predictive value";
5. "NPV", NPV = TN/(TN + FN), "negative predictive value";
6. "F1", F1 = 2*Se*PPV/(Se + PPV);
7. "accuracy", (TP + TN)/(TP + TN + FP + FN).

Author(s)

Fernando Palluzzi <fernando.palluzzi@gmail.com>

See Also

us.predict

Examples

```
# RBM performance
p <- predict(mpm.rbm$fit, dichotomize(mpm.us[2:15], asFactor = TRUE),
             type = "response")
y.hat <- ifelse(p > 0.5, 1, 0)
rbm <- performance(mpm.us$y, y.hat)

# Compare RBM and RFC F1 scores
print(rbm$F1)
print(mpm.rfc$performance$F1)
```

ranksim *Rank ultrasound profiles by similarity*

Description

Filter, rank, and return the k top-similar ultrasound profiles respect to the input one, by searchin across a reference dataset. This function implements the similarity profiling module (Moprhonode-SP).

Usage

```
ranksim(
  u,
  v = NULL,
  x = mpm.us,
  k = 5,
  p = 0.7,
  j = 2:15,
  d = c(2:6, 9, 10, 11),
  signature = NULL,
  check.data = TRUE,
  orderbyDistance = FALSE,
  ...
)
```

Arguments

`u` An ultrasound vector generated by `set.rfcdata`.

`v` An ultrasound profile generated by `set.rbmdata` (default = NULL). If this argument is not NULL, the core similarity function will be switched from "cosine" to "jaccard".

`x` Reference ultrasound data.frame. If no reference is specified, the default simulated dataset (object `mpm.us`) will be used.

`k` Numeric value defining the number of top-k profiles to return after similarity ranking (default = 5).

`p` Continuous value from 0 to 1 representing the minimum similarity value for an ultrasound profile to be included in the output (default = 0.7).

`j` Indices of the features (columns) in `x` used to compute profile similarity (default = 2:15).

`d` Indices of the features (columns) in `x` used to compute euclidean distance (default = c(2:6, 9, 10, 11)).

`signature` One among "LMR", "MMR", "HMR", "MET" (default = NULL). Resctrict the similarity search to a given metastatic risk signature (if NULL, no signature restriction is applied).

`check.data` Logical value. If TRUE (default), the input data type is checked.

`orderbyDistance` Logical value. If TRUE, the k top-similar profiles are finally ordered by increasing euclidean distance (default = FALSE).

`...` Currently ignored.

Details

The input ultrasound profile is compared to each entry in the reference dataset by computing pairwise similarity. By default, cosine similarity is used, while jaccard similarity is enabled if a binary vector is given. The hits are then filtered by minimum similarity (by default, > 0.7) and pairwise euclidean distance between them is computed. Results are then ranked by either decreasing similarity (default) or increasing distance.

Value

A list of 4 objects:

1. "signature", metastatic risk signature (MRS);
2. "p", MRS-associated malignancy risk (evaluated as positive predictive value, according to Fragomeni et al. 2022);
3. "ci95", 95
4. "y.uss", naive guess of the outcome (0: non-malignant, 1: malignant), based on the MRS (this will be less accurate than the RFC-based prediction).

Author(s)

Fernando Palluzzi <`fernando.palluzzi@gmail.com`>

References

Fragomeni SM, Moro F, Palluzzi F, Mascilini F, Rufini V, Collarino A, Inzani F, Giacò L, Scambia G, Testa AC, Garganese G (2022). Evaluating the risk of inguinal lymph node metastases before surgery using the Morphonode Predictive Model: a prospective diagnostic study. Ultrasound xx Xxxxxxxxxx xxx Xxxxxxxxxx. 00(0):000-000. <https://doi.org/00.0000/0000000000000000000>

Leydesdorff L (2005). Similarity Measures, Author Cocitation Analysis,and Information Theory. In: JASIST 56(7), pp.769-772. <https://doi.org/10.48550/arXiv.0911.4292>

See Also

See `us.predict` to launch all morphonode modules at once. See also `topsim` for a simple similarity search.

Examples

```
# Prepare a simulated malignant ultrasound profile
x <- new.profile(us.simulate(y = 1))
u <- set.rfcdata(x, ref = mpm.us[, 2:15], levels = mpm.levels)
v <- set.rbmdata(x, ref = mpm.us[, 2:15], levels = mpm.levels)
print(u)
print(v)

# Rank using cosine similarity and the default simulated reference
Rc <- ranksim(u$ultrasound)

# Rank using jaccard similarity and the default simulated reference
Rj <- ranksim(u$ultrasound, v$ultrasound)

# Compare the two rankings
print(Rc)
print(Rj)
```

---

rfc *Build a random forest classifier*

Description

Wrapper for the `GenerateLearningsets` `classification` functions.

Usage

```
rfc(data, status, k = 5, method = "CV", ...)
```

Arguments

`data`	An (n, m) matrix or data.frame with no outcome attribute.
`status`	A vector of length n, containing the outcome.
`k`	number of cross-validation iterations (default = 5).
`method`	One of the `GenerateLearningsets` methods, including: "LOOCV", "CV", "MCCV", and "bootstrap" (default = "CV").
`...`	Currently ignored.

Value

A list of objects of class "cloutput" and "clvarseloutput", respectively.

Author(s)

Fernando Palluzzi <fernando.palluzzi@gmail.com>

References

Slawski M, Daumer M, Boulesteix AL. CMA - a comprehensive Bioconductor package for supervised classification with high dimensional data. BMC Bioinformatics 9, 439 (2008). <https://doi.org/10.1186/1471-2105-9-439>

See Also

See `us.predict` to launch all morphonode modules at once. See also `topsim` for a simple similarity search.

Examples

```
# Extract a subset of 300 subjects and an outcome vector of length 30
# from the default simulated dataset

x <- mosaic::sample(mpm.us, 300, replace = FALSE, prob = NULL)
y <- x$y
x <- x[, 2:15]
dim(x)
length(y)

# Build a 5-fold cross-validation object

CV <- rfc(x, status = y)

# Performances of the first of five predictors

CV1 <- CV[[1]]
P <- performance(obs = CV1@y, pred = CV1@yhat)
print(P)
```

rfc.predict *Random Forest Classifier-based prediction*

Description

Core function of the random forest classifier for malignancy prediction (Morphonode-RFC module). Given an ultrasound vector generated by `set.rfcdata`, this function yields a prediction of malignancy (y = 1) or non-malignancy (y = 0).

Usage

```
rfc.predict(u, rfc, recover = TRUE, ...)
```

Arguments

`u`	An ultrasound vector generated by `set.rfcdata`.
`rfc`	Random forest classifier as an object of class `randomForest`. The default classifier `mpm.rfc$rfc` can be used (see details).
`recover`	Logical value. If TRUE (default) the predictors with least out-of-bag error get the highest priority, in case of a tie.
`...`	Currently ignored.

Details

The default classifier (`rfc` = `mpm.rfc$rfc`) is a set of 5 RFCs are used to predict subject's phenotype (0: non-malignant, 1: malignant). Each RFC is trained through a 5-fold nested cross-validation procedure over 10000 random trees, with 3/14 randomly chosen variables per tree branching. Each of the 5 RFCs achieves and independent prediction and the majority wins. The input is the default simulated dataset (object `mpm.us`), of 948 subjects (508 non-malignant and 440 malignant profiles) and 14 ultrasound features. The dataset includes also the expected phenotype (y), the related metastatic risk signature (signature), and the Brier score (E) calculated during the cross-validation procedure.

Value

A list of 3 objects:

1. "y.hat", final prediction;
2. "decisions", prediction of each single RFC;
3. "oob.error", out-of-bag error of each classifier in the ensemble.

Author(s)

Fernando Palluzzi <`fernando.palluzzi@gmail.com`>

References

Fragomeni SM, Moro F, Palluzzi F, Mascilini F, Rufini V, Collarino A, Inzani F, Giacò L, Scambia G, Testa AC, Garganese G (2022). Evaluating the risk of inguinal lymph node metastases before surgery using the Morphonode Predictive Model: a prospective diagnostic study. Ultrasound xx Xxxxxxxxxx xxx Xxxxxxxxxx. 00(0):000-000. <https://doi.org/00.0000/0000000000000000000>

Liaw A, Wiener M. Classification and Regression by randomForest (2002). R News, 2(3):18-22. <https://doi.org/10.1023/A:1010933404324>

See Also

See `us.predict` to launch all morphonode modules at once. See also `topsim` for a simple similarity search.

Examples

```
# Prepare a simulated malignant ultrasound profile
x <- new.profile(us.simulate(y = 1))
u <- set.rfcdata(x, ref = mpm.us[, 2:15], levels = mpm.levels)
print(u)

# Predict subject's phenotype
P <- rfc.predict(u$ultrasound, rfc = mpm.rfc$rfc)
print(P)
```

---

set.missing *Cope with possible missing data in the ultrasound profile*

Description

This function imputes missing data in the ultrasound profile, creating a new profile with imputed missing values. If no missing values are found, it will simply send a message and return the input profile.

Usage

```
set.missing(
  v,
  ref = NULL,
  levels = NULL,
  con = 1:2,
  cat = 3:14,
  missing = -1,
  na = NA,
  asNumeric = TRUE,
  ...
)
```

Arguments

`v` An ultrasound profile generated by `new.profile`, i.e., a list of two objects, containing a vector of length 14 (`list$ultasound`), corresponding to the input ultrasound profile (14 ultrasound variables or "features"), and a vector of missing value indices (`list$missing`).

`ref` A data.frame representing the reference dataset. The ultrasound profile will be attached to the reference dataset before the imputation. This argument is required to impute missing features.

`levels` A list of length 14, corresponding to the levels of each ultrasound variable. Needed for categorical variables (factors); for continuous variables, it should assume the nominal value of 0. The default levels variable `mpm.levels` can be used.

`con` Vector of the indices corresponding to continuous variables in the `list$ultasound` vector (default = 1:2).

`cat` Vector of the indices corresponding to categorical variables in the `list$ultasound` vector (default = 3:14).

`missing` Value used to mark missing data (default = -1).

`na` Value used for "not available" data (default = NA). This will be used to dubstitute `missing` within the ultrasound vector before the imputation.

`asNumeric` Logical value used to convert every value in the ultrasound vector to class "numeric". This argument is used only if `ref = NULL` and `levels = NULL`.

`...` Currently ignored.

Details

Automatic imputation is necessary to improve RFC-based (malignancy prediction) and RBM-based (metastatic risk evaluation) estimations. Imputation is currently forbidden for short axis and cortical thickness (i.e., the first two ultrasound features), since they have a critical role in the prediction, estimation and signature detection processes. Hence, their actual value must be entered for a reliable prediction. Although permitted, the imputation is discouraged for the following three features: nodal core sign (i.e., hilum presence), perinodal hyperechogenic ring (i.e., the presence of inflammatory stroma), and cortical interruption (i.e., extracapsular spread). These features define a strongly metastatic profile with possible multiple metastases (i.e., the "MET" signature) that are hardly imputable from the other ultrasound variables.

Value

An ultrasound profile with imputed missing values.

Author(s)

Fernando Palluzzi <`fernando.palluzzi@gmail.com`>

See Also

See `set.rfcdata` for random forest (RFC) data preparation and `set.rbmdata` for robust binomial model (RBM) data preparation.

Examples

```
# Create an ultrasound profile with missing values

u <- new.profile(c(10.0, 6.3, 1, 0, 0, 0, -1, 1, 2, 2, 3, 1, -1, -1))
print(u)

# Fix missing values with the default simulated dataset as reference
# (ultrasound features only: \code{mpm.us} attributes 2 to 15).
# Default levels are provided by the \code{mpm.levels} object.

v <- set.missing(u, ref = mpm.us[, 2:15], levels = mpm.levels)
print(v)
```

---

set.rbmdata *Ultrasound profile preparation for metastatic risk estimation*

Description

Prepare a new ultrasound profile for metastatic risk estimation using robust binomial modeling. This function includes missing values check and fix (see set.missing).

Usage

```
set.rbmdata(
  u,
  levels,
  short = 8,
  cortical = 2,
  ist = 1,
  ecs = 1,
  hab = 0,
  eco = 1,
  vp = c(1, 2, 3),
  vfl = c(2, 3, 4),
  ct = c(2),
  fid = c(1, 2, 3),
  cmid = c(2, 3, 4),
  shape = c(3),
  cs = 3,
  grouping = c(2, 3),
  ref = NULL,
  asFactor = FALSE,
  ...
)
```

Arguments

u New ultrasound profile generated by new.profile.

levels A list of length 14, corresponding to the levels of each ultrasound variable. Needed for categorical variables (factors); for continuous variables, it should assume the nominal value of 0. The default levels variable mpm.levels can be used.

short Numeric value corresponding to the short axis cutoff in millimeters (default = 8).

cortical Numeric value corresponding to the cortical thickness cutoff in millimeters (default = 2).

ist Dichotomous value 0, 1 for the presence of inflammatory stroma (perinodal hyperechogenic ring; default = 1).

ecs Dichotomous value 0, 1 for the presence of extracapsular spread (cortical interruption; default = 1).

hab Dichotomous value 0, 1 for the absence of the hilum (nodal core sign; default = 0).

eco Dichotomous value 0, 1 for heterogeneous echogenicity (echostructure; default = 1).

vp Categorical value (integers from 0 to 4) associated to a high-risk vascular flow architecture pattern (default = c(1, 2, 3)).

vfl Categorical value (integers from 0 to 4) associated to a high-risk vascular flow localization (default = c(2, 3, 4)).

ct Categorical value (integers from 0 to 4) associated to a high-risk cortical thickening (default = 2).

fid Categorical value (integers from 0 to 3) associated to a high-risk focal intranodal deposit (default = c(1, 2, 3)).

cmid Categorical value (integers from 0 to 4) associated to a high-risk cortical-medullar interface distortion (default = c(2, 3, 4)).

shape Categorical value (integers from 1 to 3) associated to a high-risk shape (default = 3).

cs Ordinal value (integers from 1 to 5) associated to a high-risk color score (default = 3).

grouping Categorical value (integers from 1 to 3) associated to a high-risk grouping (default = c(2, 3)).

ref Reference ultrasound features dataset as a (n, 14) data.frame, with n being the number of subjects (rows). The default simulated dataset mpm.us can be used.

asFactor Logical value. If TRUE, data.frame columns are converted to factors (default = FALSE).

... Currently ignored.

Value

A list of 2 objects:

1. "ultrasound", dichotomized ultrasound features vector;
2. "missing", indices of missing values (empty if no missing values are found).

Author(s)

Fernando Palluzzi <fernando.palluzzi@gmail.com>

References

Fragomeni SM, Moro F, Palluzzi F, Mascilini F, Rufini V, Collarino A, Inzani F, Giacò L, Scambia G, Testa AC, Garganese G (2022). Evaluating the risk of inguinal lymph node metastases before surgery using the Morphonode Predictive Model: a prospective diagnostic study. Ultrasound xx Xxxxxxxxxx xxx Xxxxxxxxxx. 00(0):000-000. <https://doi.org/00.0000/00000000000000000000>

See Also

See set.rfcdata for random forest classifier data preparation. See also us.predict to launch all morphonode modules at once.

Examples

```
# Prepare a simulated malignant ultrasound profile
x <- new.profile(us.simulate(y = 1))
print(x)

# Set the new profile for RFC prediction
v <- set.rbmdata(x, ref = mpm.us[, 2:15], levels = mpm.levels)
print(v)
```

set.rfcdata *Ultrasound profile preparation for random forest classification*

Description

Prepare a new ultrasound profile for RFC prediction. This function includes missing values check and fix (see set.missing).

Usage

```
set.rfcdata(u, levels = NULL, ref = NULL, ...)
```

Arguments

u	New ultrasound profile generated by new.profile.
levels	A list of length 14, corresponding to the levels of each ultrasound variable. Needed for categorical variables (factors); for continuous variables, it should assume the nominal value of 0. The default levels variable mpm.levels can be used.
ref	Reference ultrasound features dataset as a (n, 14) data.frame, with n being the number of subjects (rows). The default simulated dataset mpm.us can be used.
...	Currently ignored.

Value

A list of 2 objects:

1. "ultrasound", ultrasound features vector;
2. "missing", indices of missing values (empty if no missing values are found).

Author(s)

Fernando Palluzzi <fernando.palluzzi@gmail.com>

References

Fragomeni SM, Moro F, Palluzzi F, Mascilini F, Rufini V, Collarino A, Inzani F, Giacò L, Scambia G, Testa AC, Garganese G (2022). Evaluating the risk of inguinal lymph node metastases before surgery using the Morphonode Predictive Model: a prospective diagnostic study. Ultrasound xx Xxxxxxxxxx xxx Xxxxxxxxxx. 00(0):000-000. <https://doi.org/00.0000/0000000000000000000>

See Also

See set.rbmdata for robust binomial model data preparation. See also us.predict to launch all morphonode modules at once.

Examples

```
# Prepare a simulated malignant ultrasound profile
x <- new.profile(us.simulate(y = 1))

# Set the new profile for RFC prediction
u <- set.rfcdata(x, ref = mpm.us[, 2:15], levels = mpm.levels)
print(u)
```

`show.mpmenv` *Morphonode ultrasound variables description*

Description

Shows a description of the utrasound variables used by the **morphonode** package.

Usage

```
show.mpmenv(brief = TRUE, ...)
```

Arguments

`brief`	Logical value enabling a compact view (default = TRUE).
`...`	Currently ignored.

Author(s)

Fernando Palluzzi <`fernando.palluzzi@gmail.com`>

Examples

```
show.mpmenv()
```

`similarity` *Compute different similarity coefficients*

Description

Compute the similarity between two vectors.

Usage

```
similarity(x, y, f = "cosine", ceiling = 1e+07, ...)
```

Arguments

`x`	A numeric vector.
`y`	A numeric vector.
`f`	Similarity function. One among "cosine" (default), "jaccard" (for dichotomous vectors only), "pearson", "spearman", "kendall", or "euclidean".
`ceiling`	If `f = "euclidean"`, the similarity is computed as 1/distance. This argument limits the similarity to a very high value, in case the euclidean distance is equal to 0. The default value is 1E+07.
`...`	Currently ignored.

Value

A numeric value corresponding to the similarity coefficient.

Author(s)

Fernando Palluzzi <fernando.palluzzi@gmail.com>

References

Leydesdorff L (2005). Similarity Measures, Author Cocitation Analysis,and Information Theory. In: JASIST 56(7), pp.769-772. <https://doi.org/10.48550/arXiv.0911.4292>

See Also

jaccard, euclidean

Examples

```
# Sample two random ultrasound profiles from the default dataset
x <- mosaic::sample(mpm.us, 1, replace = FALSE, prob = NULL)
x <- as.numeric(x[, 2:15])
y <- mosaic::sample(mpm.us, 1, replace = FALSE, prob = NULL)
y <- as.numeric(y[, 2:15])

# Compute the cosine similarity
r <- similarity(x, y)
print(r)
```

topsim *Similarity search*

Description

Vector similarity search across a reference dataset.

Usage

```
topsim(
  v,
  x = mpm.us,
  f = "cosine",
  k = 5,
  p = 0.7,
  features = 2:15,
  check.data = TRUE,
  ...
)
```

Arguments

`v` An input vector of length equal to the number of columns of the reference data.frame (see below).

`x` Reference data.frame. If no reference is specified, the default simulated dataset (object `mpm.us`) will be used.

`f` Similarity function. Available functions: "cosine", "jaccard", "euclidean", "pearson", "spearman", "kendall" (default = "cosine").

`k` Numeric value defining the number of top-k profiles to return after similarity ranking (default = 5).

`p` Continuous value from 0 to 1 representing the minimum similarity value for an ultrasound profile to be included in the output (default = 0.7).

`features` Indices of the features (columns) in `x` used to compute similarity (default = 2:15).

`check.data` Logical value. If TRUE (default), the input data type is checked.

`...` Currently ignored.

Value

A list of 4 objects:

1. "signature", metastatic risk signature (MRS);
2. "p", MRS-associated malignancy risk (evaluated as positive predictive value, according to Fragomeni et al. 2022);
3. "ci95", 95
4. "y.uss", naive guess of the outcome (0: non-malignant, 1: malignant), based on the MRS (this will be less accurate than the RFC-based prediction).

Author(s)

Fernando Palluzzi <`fernando.palluzzi@gmail.com`>

References

Leydesdorff L (2005). Similarity Measures, Author Cocitation Analysis,and Information Theory. In: JASIST 56(7), pp.769-772. <https://doi.org/10.48550/arXiv.0911.4292>

See Also

See `us.predict` to launch all morphonode modules at once. See also `ranksim` for ultrasound profile similarity ranking.

Examples

```
# Prepare a simulated malignant ultrasound profile
x <- new.profile(us.simulate(y = 1))
u <- set.rfcdata(x, ref = mpm.us[, 2:15], levels = mpm.levels)

# Top-similar profiles
sim <- topsim(u$ultrasound)
print(sim)
```

---

us.predict *Morhonode Predictive Model (MPM) launcher*

Description

The `us.predict` function launches the 4 MPM modules: (i) malignancy prediction (Morphonode-RFC), (ii) malignancy risk estimation (Morphonode-RBM), (iii) malignancy risk signature detection (Morphonode-DT), and (iv) similarity profiling (Morphonode-SP). The MPM structure is described in Fragomeni et al. (2022). See also the details section.

Usage

```
us.predict(
  x,
  f = "cosine",
  levels = NULL,
  ref = NULL,
  rfc = NULL,
  rbm = NULL,
  k = 5,
  features = 2:15,
  orderbyDistance = FALSE,
  uncertainty = "loss",
  b0 = c(0, 0.028, 0.013),
  b = 0.055,
  rho = 0.9,
  wmax = 1,
  verbose = TRUE,
  ...
)
```

Arguments

`x`	Ultrasound profile generated by the function `new.profile`).
`f`	Similarity profiling core function: one between "cosine" (default) and "jaccard". The former directly compares ultrasound profiles, while the latter uses dichotomized versions of them (see also `dichotomize`).

levels A list of length 14, corresponding to the levels of each ultrasound variable. Needed for categorical variables (factors); for continuous variables, it should assume the nominal value of 0. If NULL (default), the internal mpm.levels variable will be used.

ref Reference ultrasound features dataset as a (n, 14) data.frame, with n being the number of subjects (rows). If NULL (default), the internal mpm.us variable will be used.

rfc Random forest classifier as an object of class randomForest. If NULL (default), the internal mpm.rfc variable will be used.

rbm Robust binomial model as a fitted model object of class glm. If NULL (default), the internal mpm.rbm variable will be used.

k Numeric value defining the number of top-k profiles to return after similarity ranking (default = 5).

features Indices of the features to be used for the similarity profiling (default = 2:15).

orderbyDistance

Logical value enabling sorting by increasing euclidean distance of the top-similar ultrasound profiles (default = FALSE).

uncertainty Function used to compute the RFC prediction error. It can be one among "loss" (default) and "similarity". The former estimates the error on a new prediction based on a parametric linear relationship between the loss function and the observed (reference dataset) error. The latter estimates the error as the average of the observed errors of the top-3 similar profiles from the reference dataset (non-parametric).

b0 Baseline uncertainty. A vector of three values representing the intercept parameter when uncertainty = "loss". This value depends on the metastatic risk signature and it is fixed to 0 for the LMR one. The b0 values for MMR, HMR, and MET signatures are the first, second, and third element of the b0 argument, respectively (default = c(0, 0.028, 0.013)).

b Uncertainty coefficient. A numeric value indicating the linear coefficient for the loss-to-error conversion equation.

rho Numeric value between 0 and 1 denoting the minimum required similarity coefficient to make a naive estimation of the true outcome y = 0, 1. This guess is needed as a reference value for uncertainty calculation (default = 0.9).

wmax Nominal value for the self-correlation coefficient (for visualization purposes only; default = 1).

verbose If TRUE (default), a user-frienly summary of the prediction and estimation results is printed to screen.

... Currently ignored.

Details

The MPM classifier (Morphonode-RFC module) is based on an ensemble of 5 RFCs. Each RFC is trained over 10000 random trees, with 3/14 randomly chosen variables per tree branching. The 5 RFCs yield independent predictions and the majority wins. This module provides a dichotomous phenotype classification in malignant (y = 1) and non-malignant (y = 0), and an estimation of the prediction error (E). Similarly, the Morphonode-RBM module provides a continuous estimate of malignancy (i.e., p = malignancy risk), through a binomial model with robust bootstrap standard error estimation (5000 bootstrap iterations). Optimal cutpoint estimations define two thresholds for p (three risk intervals): low risk (p < 0.23), moderate risk (0.23 <= p <= 0.29), and high risk (p > 0.29). In addition, Morphonode-DT model defines four metastatic risk signatures, strongly associated with the metastasis rate in the corresponding subjects. LMR (low metastatic risk) and MMR (moderate metastatic risk) signatures are associated with none-to-low metastasis rates. Conversely, HMR (high metastatic risk) and MET (metastatic) signatures are associated with a high risk of single and multiple metastatic events (lymph nodes), respectively. Finally, the Morphonode-SP module ranks ultrasound profiles from the reference dataset (by default, the internal simulated dataset) by similarity with respect to the input profile. This provides a supplementary support to the classification process, having only a secondary role compared to the other three modules. Generally, the majority of similar profiles should have the same outcome (y) as the input one.

Value

A list of 5 objects:

1. "prediction" (Morphonode-RFC module), a list including:
   - `y.hat`: the final malignancy prediction,
   - `decisions`: the predictions of each RFC in the ensemble,
   - `oob.err`: out-of-bag errors of each RFC in the ensemble;
2. "E", estimated overall prediction error (Morphonode-RFC module);
3. "p", estimated malignancy risk(Morphonode-RBM module);
4. "signature", metastatic risk signature (Morphonode-DT module);
5. "profiles", data.frame containing the top-k similar profiles sorted by similarity (Morphonode-SP module). This data.frame includes:
   - ID: numeric value identifying a subject,
   - the 14 ultrasound features characterizing each subject,
   - y: the observed outcome,
   - E: subject-level estimated prediction error (Brier score),
   - R: similarity coefficient with the input profile,
   - D: euclidean distance from the input profile.

Author(s)

Fernando Palluzzi <`fernando.palluzzi@gmail.com`>

References

Fragomeni SM, Moro F, Palluzzi F, Mascilini F, Rufini V, Collarino A, Inzani F, Giacò L, Scambia G, Testa AC, Garganese G (2022). Evaluating the risk of inguinal lymph node metastases before surgery using the Morphonode Predictive Model: a prospective diagnostic study. Ultrasound xx Xxxxxxxxxx xxx Xxxxxxxxxx. 00(0):000-000. <https://doi.org/00.0000/0000000000000000000>

Liaw A, Wiener M. Classification and Regression by randomForest (2002). R News, 2(3):18-22. <https://doi.org/10.1023/A:1010933404324>

See Also

See `new.profile` to create a new ultrasound profile. See also `us.simulate` for ultrasound data simulation.

Examples

```
# Create a simulated malignant ultrasound profile
x <- new.profile(us.simulate(y = 1))

# Lauch the Morhonode Predictive Model
u <- us.predict(x)
```

`us.simulate` *Ultrasound profile simulation*

Description

Ultrasound profile simulation based on a reference dataset.

Usage

```
us.simulate(
  reps = 1,
  x = mpm.us,
  y = NULL,
  signature = NULL,
  features = 2:15,
  header = NULL,
  ...
)
```

Arguments

`reps` Number of simulated vectors to be generated (default = 1).

`x` Reference ultrasound features data.frame. By default, a dataset of 948 simulated profiles (object `mpm.us`) is used.

`y` An optional numeric value defining the phenotype to be simulated: 0 for non-malignant and 1 for malignant (default = NULL).

`signature` An optional string defining the metastatic risk signature to be simulated. One among: "LMR" (low metastatic risk), "MMR" (moderate metastatic risk), "HMR" (high metastatic risk), "MET" (metastatic signature). Defaults to NULL.

`features` A vector of indices defining the ultasound features within the reference data.frame (default = 2:15).

`header` A vector of new names for the ultrasound features vector. If NULL (default), feature names will be imported from the reference.

`...` Currently ignored.

Value

A simulated ultrasound features vector.

Author(s)

Fernando Palluzzi <fernando.palluzzi@gmail.com>

See Also

See `new.profile` for ultrasound profile creation and `us.predict` for subject classification.

Examples

```
# Simulate a generic ultrasound profile
u <- us.simulate()
print(u)

# Simulate a non-malignant ultrasound profile
u0 <- us.simulate(y = 0)
print(u0)

# Simulate a malignant ultrasound profile
u1 <- us.simulate(y = 1)
print(u1)

# Simulate a high metastatic risk profile (single metastasis event)
u.hmr <- us.simulate(signature = "HMR")
print(u.hmr)

# Simulate a metastatic ultrasound profile (multiple metastasis events)
u.met <- us.simulate(signature = "MET")
print(u.met)
```

`uss` *Ultrasound signatures (uss) detection*

Description

This function implements the decision tree described in Fragomeni et al. 2022 (Morphonode-DT module) to detect metastatic risk signatures (MRSs). The first split detects subjects with a "MET" (metastatic) signature. These individuals show at least one among the three metastatic markers (see details) and a high risk of malignancy (86-100 usually coming with multiple metastatic lymph nodes. The other decision tree branches are defined on the base of five key ultrasound features (see details), with a malignancy risk (MR) signature ranging from low ("LMR", 0-10 to high ("HMR", 52-90 in most malignancies. The main goal of MRSs is to predict single-metastatic event malignancies (HMR) and multiple-metastatic event malignancies (MET).

Usage

```
uss(
  x,
  dichotomous = FALSE,
  dct = 2,
  short = 8,
  cortical = 2,
  ist = 1,
  ecs = 1,
  hab = 0,
  eco = 1,
  vp = c(1, 2, 3),
  vfl = c(2, 3, 4),
  ct = c(2),
  fid = c(1, 2, 3),
  cmid = c(2, 3, 4),
  shape = c(3),
  cs = 3,
  grouping = c(2, 3),
  ...
)
```

Arguments

`x` An (n, 14) ultrasound features data.frame, where n is the number of subjects.

`dichotomous` Logical value. The first step of this function is to dichotomize the input data.frame. Set `dichotomous = TRUE` if the input data.frame is already dichotomous (default = FALSE).

`dct` Numeric value. Moderate risk signature predicts a baseline risk of malignancy equal to 0.16 (CI95 Fragomeni et al. (2022), if at least 2 other ultrasound features than those used in the decision tree (referred to as diagnostic covariates) are above their optimal threshold, the malignancy risk increases to 0.55 (CI95 as MMR1, in contrast to the basal MMR signature (MMR0). The argument `dct` controls the number of diagnostic covariates needed to switch from MMR0 to MMR1 (default = 2).

`short` Numeric value corresponding to the short axis cutoff in millimeters (default = 8).

`cortical` Numeric value corresponding to the cortical thickness cutoff in millimeters (default = 2).

`ist` Dichotomous value 0, 1 for the presence of inflammatory stroma (perinodal hyperechogenic ring; default = 1).

`ecs` Dichotomous value 0, 1 for the presence of extracapsular spread (cortical interruption; default = 1).

`hab` Dichotomous value 0, 1 for the absence of the hilum (nodal core sign; default = 0).

`eco` Dichotomous value 0, 1 for heterogeneous echogenicity (echostructure; default = 1).

`vp` Categorical value (integers from 0 to 4) associated to a high-risk vascular flow architecture pattern (default = c(1, 2, 3)).

`vfl` Categorical value (integers from 0 to 4) associated to a high-risk vascular flow localization (default = c(2, 3, 4)).

`ct` Categorical value (integers from 0 to 4) associated to a high-risk cortical thickening (default = 2).

`fid` Categorical value (integers from 0 to 3) associated to a high-risk focal intranodal deposit (default = c(1, 2, 3)).

`cmid` Categorical value (integers from 0 to 4) associated to a high-risk cortical-medullar interface distortion (default = c(2, 3, 4)).

`shape` Categorical value (integers from 1 to 3) associated to a high-risk shape (default = 3).

`cs` Ordinal value (integers from 1 to 5) associated to a high-risk color score (default = 3).

`grouping` Categorical value (integers from 1 to 3) associated to a high-risk grouping (default = c(2, 3)).

`...` Currently ignored.

Details

The core method of the Morphonode-DT model is implemented in this function. A series of binary branching points define the metastatic risk signature (MRS) of the subject. The first branch point is based on the evaluation of three metastatic markers: the absence of the nodal core sign (hilum), the presence of the perinodal hyperechogenic ring, and the presence of cortical interruption. If at least one of these conditions are true, the ultrasound profile has a high malignancy risk (86-100 multiple metastatic lymph nodes (Fragomeni et al. 2022). This is referred to as the metastatic (MET) signature. A cortical thickness below 2 mm defines a low metastatic risk (LMR) signature (0.04, CI95 malignant. Based on the values of short axis, vascular flow architecture pattern, cortical thickening, and vascular flow localization, two more signatures are defined: (i) moderate metastatic risk (MMR; 0.16, CI95 CI95 mostly malignant, but chraracterized by a single metastatic event. MRSs should be always compared to the output of the random forest classifier (Morphonode-RFC module) and robust binomial model (Morphonode-RBM module). The main advantage of MRSs is the prediction of either multiple (MET signature) or single (HMR signature) metastasis events.

Value

A list of 4 objects:

1. "signature", metastatic risk signature (MRS);
2. "p", MRS-associated malignancy risk (evaluated as positive predictive value, according to Fragomeni et al. 2022);
3. "ci95", 95
4. "y.uss", naive guess of the outcome (0: non-malignant, 1: malignant) based on the MRS (this will be less accurate than the RFC-based prediction).

Author(s)

Fernando Palluzzi <fernando.palluzzi@gmail.com>

References

Fragomeni SM, Moro F, Palluzzi F, Mascilini F, Rufini V, Collarino A, Inzani F, Giacò L, Scambia G, Testa AC, Garganese G (2022). Evaluating the risk of inguinal lymph node metastases before surgery using the Morphonode Predictive Model: a prospective diagnostic study. Ultrasound xx Xxxxxxxxxx xxx Xxxxxxxxxx. 00(0):000-000. <https://doi.org/00.0000/00000000000000000000>

See Also

See us.predict to launch all morphonode modules at once.

Examples

```
# Extract 5 random subjects from the default simulated dataset
x <- mosaic::sample(mpm.us[, 2:15], 5, replace = FALSE, prob = NULL)
print(x)

# Assign a metastatic risk signature to each subject in the dataset
mrs <- uss(x)
x$signature <- mrs$signature
print(x)
```

---

vpart *Data partitioning utility*

Description

Extract a random partition from an input dataset.

Usage

```
vpart(data, p, shuffle = FALSE, ...)
```

Arguments

data	An input data.frame.
p	Proportion of rows to be extracted from the input dataset.
shuffle	A logical value. If TRUE, the input rows are randomly shuffled before data partitioning.
...	Currently ignored.

Value

A list of 2 data.frames:

1. "training.set", the portion of the input data defined by p;
2. "validation.set", the portion of the input data defined by 1-p.

Author(s)

Fernando Palluzzi <fernando.palluzzi@gmail.com>

See Also

buildPredictor,

Examples

```
# Extract a subset of 300 subjects and an outcome vector of length 30
# from the default simulated dataset

x <- mosaic::sample(mpm.us, 300, replace = FALSE, prob = NULL)

# Data partitioning

x <- vpart(x, p = 0.75)
print(dim(x$training.set))
print(dim(x$validation.set))
```

Index