

Package ‘maxtwin’

October 29, 2026

Title A Robust Deterministic Data Splitting Method

Version 0.1.0

Description Provides a deterministic method for splitting data into training and testing sets that maximizes the distance between training and testing points while preserving their distributions. See Joseph and Cho (in press) for details. This work is supported by U.S. National Science Foundation grant DMS-2310637.

License GPL (>= 3)

Encoding UTF-8

Depends R (>= 3.5.0)

Imports FNN, twinning (>= 1.1)

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

RoxygenNote 7.3.3

NeedsCompilation no

Author Youngseo Cho [aut, cre],
V. Roshan Joseph [aut]

Maintainer Youngseo Cho <ycho397@gatech.edu>

Repository CRAN

Date/Publication 2026-10-29 11:40:02 UTC

Contents

initialmaxtwin	2
maxtwin	3
Index	5

initialmaxtwin	<i>Initial MaxTwin (Step 1)</i>
----------------	---------------------------------

Description

Selects testing points by maximizing nearest-neighbor distance within each Voronoi cell induced by the twinning set of Vakayil and Joseph (2022). This is Step 1 of the algorithm of Joseph and Cho (in press), without iterative improvement of the separation distance.

Usage

```
initialmaxtwin(
  X,
  Y = NULL,
  splitratio,
  initialization = TRUE,
  standardization = TRUE
)
```

Arguments

<code>X</code>	Matrix or data frame of inputs (N x p). Columns may be numeric or factors.
<code>Y</code>	Optional numeric or factor vector of outputs (length N). If supplied, it is included in twinning. If omitted, splitting is based on X alone.
<code>splitratio</code>	Fraction of data for the testing set. Must be 1/r where r is an integer ≥ 2 (e.g., 0.5, 0.25, 0.2, 0.1). This constraint comes from the twinning algorithm.
<code>initialization</code>	Logical. If TRUE, start twinning from the point closest to the data centroid.
<code>standardization</code>	Logical. If TRUE, standardize the encoded data used for twinning before selection.

Value

A vector of testing point indices.

References

Vakayil, A. and Joseph, V.R. (2022). Data Twinning. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(5), 598–610. doi:10.1002/sam.11574.

Joseph, V.R. and Cho, Y. (in press). A Robust Deterministic Data Splitting Method for Model Validation. *Journal of Computational and Graphical Statistics*.

maxtwin

*MaxTwin Algorithm***Description**

Implements the robust deterministic data-splitting algorithm of Joseph and Cho (in press). Step 1 selects testing points using twinning, Voronoi partitioning, and maximum nearest-neighbor distance. Step 2 iteratively improves separation if needed.

Usage

```
maxtwin(X, Y = NULL, splitratio, initialization = TRUE, standardization = TRUE)
```

Arguments

X	Matrix or data frame of inputs (N x p). Columns may be numeric or factors.
Y	Optional numeric or factor vector of outputs (length N). If supplied, it is included in twinning. If omitted, splitting is based on X alone.
splitratio	Fraction of data for the testing set. Must be 1/r where r is an integer ≥ 2 (e.g., 0.5, 0.25, 0.2, 0.1). This constraint comes from the twinning algorithm.
initialization	Logical. If TRUE, start twinning from the point closest to the data centroid.
standardization	Logical. If TRUE, standardize the encoded data used for twinning before selection.

Details

See Joseph and Cho (in press) for details.

Note that splitratio must be 1/r for integer $r \geq 2$, due to the constraint of the underlying twinning algorithm.

Value

A vector of testing point indices.

References

Joseph, V.R. and Cho, Y. (in press). A Robust Deterministic Data Splitting Method for Model Validation. *Journal of Computational and Graphical Statistics*.

Examples

```
set.seed(42)
X <- matrix(runif(200), ncol = 2)
Y <- sin(X[,1]) + cos(X[,2])
idx <- maxtwin(X, Y, splitratio = 0.2)
X_test  <- X[idx, ]
X_train <- X[-idx, ]
```

Index

`initialmaxtwin`, [2](#)

`maxtwin`, [3](#)