

Package: RPWNSC (via r-universe)

July 21, 2026

Type Package

Title Run Purity Weighted Nearest Shrunk Centroid Feature Selection

Version 0.1.0

Author Amjad Ali [aut, cre], Zardad Khan [aut], Saeed Aldahmani [aut]

Maintainer Amjad Ali <amjadali@uaeu.ac.ae>

Description Provides a classifier independent filter method for high-dimensional gene-expression feature selection. The Run Purity Weighted Nearest Shrunk Centroid ('RPWNSC') score ranks genes by multiplying a run based purity score, which measures the compactness of class labels after sorting samples by each feature, by a scaled nearest shrunk centroid score, which measures standardized class centroid separation relative to within class variation. The top ranked features can then be used with downstream classifiers without wrapper search, feature clustering, or classifier dependent training. 'Amjad Ali, Zardad Khan, Saeed Aldahmani' (2026) <doi:10.1016/j.mlwa.2026.100947>.

License GPL-3

Encoding UTF-8

LazyData false

Depends R (>= 3.5.0)

Imports stats

NeedsCompilation no

Repository <https://cran.r-universe.dev>

Date/Publication 2026-07-21 09:50:11 UTC

RemoteUrl <https://github.com/cran/RPWNSC>

RemoteRef HEAD

RemoteSha 30a941659232b59e8b44c2a3d21722c20b0d9070

Contents

RPWNSC-package	2
Huntington	3
RPWNSC	4
Index	6

RPWNSC-package	<i>RPWNSC: Run Purity Weighted Nearest Shrunkn Centroid Feature Selection</i>
----------------	---

Description

The RPWNSC package provides a filter based feature selection method for high-dimensional gene-expression classification.

Details

The package implements the Run Purity Weighted Nearest Shrunkn Centroid (RPWNSC) score. For each feature, RPWNSC combines a run purity score, which evaluates the compactness of class labels after sorting samples by that feature, with a scaled nearest shrunkn centroid score, which evaluates standardized class centroid separation relative to within-class variation.

The final feature score is the product of the two components. This design favors features that simultaneously provide compact class label ordering and strong class discrimination. The method is classifier independent and does not require wrapper based search, feature clustering, or classifier dependent training during the ranking step.

The main exported function is RPWNSC. The package also includes the Huntington example dataset.

Author(s)

Amjad Ali, Zardad Khan, and Saeed Aldahmani

References

Ali, A., Khan, Z., and Aldahmani, S. (2026). A run purity weighted nearest shrunkn centroid feature selection for high-dimensional gene-expression classification. *Machine Learning with Applications*, 25, 100947.

See Also

[RPWNSC](#), [Huntington](#)

Huntington

Huntington's Disease Gene-Expression Dataset

Description

High-dimensional Huntington's disease gene-expression dataset included for demonstrating the RP-WNSC feature selection method.

Usage

```
data(Huntington)
```

Format

A list with two components:

X Numeric matrix of predictors with 31 samples in rows and 22,283 gene-expression features in columns.

Y Factor response vector with 31 class labels and 2 classes.

Details

The dataset follows the high-dimensional setting considered in the RPWNSC paper: a small number of observations and a very large number of gene-expression features. It is intended as a package example for testing feature ranking, selected gene extraction, and downstream classification workflows.

References

Borovecki, F., Lovrecic, L., Zhou, J., Jeong, H., Then, F., Rosas, H. D., Hersch, S. M., Hogarth, P., Bouzou, B., Jensen, R. V., and Krainc, D. (2005). Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. *Proceedings of the National Academy of Sciences*, 102(31), 11023–11028.

Examples

```
data(Huntington)
dim(Huntington$X)
table(Huntington$Y)
```

RPWNSC

*Run Purity Weighted Nearest Shrunk Centroid Feature Selection***Description**

Ranks features using the Run Purity Weighted Nearest Shrunk Centroid (RPWNSC) score and returns the top ranked feature indices.

Usage

```
RPWNSC(X, Y, topK = 10)
```

Arguments

X	A numeric matrix or data frame of predictors with samples in rows and features in columns.
Y	A response vector or factor with one class label for each row of X. At least two classes are required.
topK	Number of top ranked features to return. If topK is larger than the number of features, all features are returned in ranked order.

Details

RPWNSC is a classifier independent filter feature selection method for high-dimensional classification data, especially gene-expression data. The method combines two complementary criteria: a run based label compactness score and a standardized nearest shrunk centroid style separation score.

For each feature, the samples are first sorted according to the feature value. The ordered class labels are divided into contiguous runs, where a run is a maximal consecutive block of samples from the same class. For class c and feature j , the class wise run purity score is

$$Q_{cj} = \frac{\sum_{r \in R_{cj}} l_{rj}^2}{n_{cj}^2}.$$

Here, l_{rj} is the length of run r , and n_{cj} is the number of samples from class c in the ordered sequence. This score equals 1 when all samples of a class appear in one contiguous run and decreases when the class labels are fragmented. The feature level run purity score is the average over the classes present in the ordered sequence:

$$R_j = \frac{1}{|C_j|} \sum_{c \in C_j} Q_{cj}.$$

The centroid component follows the standardized centroid difference idea used in nearest shrunk centroid methods. The method computes the pooled within class standard deviation, a stabilizing

constant, and standardized class centroid differences. The nearest shrunken centroid feature score is

$$A_j = \sqrt{\sum_{c=1}^C d_{cj}^2}.$$

The centroid scores are scaled to the interval $[0, 1]$ across all features, giving $\tilde{A}_j$.

The final RPWNSC score is the product

$$T_j = R_j \tilde{A}_j.$$

This multiplicative score acts as a soft AND rule: a feature receives a high rank only when it has both compact class label ordering and strong standardized class centroid separation. Features with strong centroid separation but mixed sorted labels, or compact sorted labels but weak centroid separation, are penalized.

Non-finite predictor values are replaced by the corresponding feature median before scoring. Feature ranking uses only the supplied X and Y ; it does not train a classifier, perform wrapper search, cluster features, or run iterative optimization. In the rare case where all final scores are zero, the function falls back to run purity and then to the scaled centroid score so that a valid ranking is always returned.

Value

A list with the following components:

topK Integer vector containing the indices of the selected top ranked features.

Scores Numeric vector containing the final RPWNSC score for every feature.

References

Ali, A., Khan, Z., and Aldahmani, S. (2026). A run purity weighted nearest shrunken centroid feature selection for high-dimensional gene-expression classification. *Machine Learning with Applications*, 25, 100947.

Examples

```
data(Huntington)

# Small demonstration using the first 100 genes.
out <- RPWNSC(Huntington$X[, 1:100], Huntington$Y, topK = 10)

# Selected feature indices.
out$topK

# RPWNSC scores for every feature.
out$Scores
```

Index

- * **datasets**

- Huntington, [3](#)

- * **methods**

- RPWNSC, [4](#)

- * **multivariate**

- RPWNSC, [4](#)

- * **package**

- RPWNSC-package, [2](#)

Huntington, [2](#), [3](#)

RPWNSC, [2](#), [4](#)

RPWNSC-package, [2](#)