Package 'SubLasso'

February 19, 2015

Type Package

Title Gene selection using Lasso for Microarray data with user-defined genes fixed in model.

Version 1.0

Depends R(>= 2.10), glmnet, psych, gplots

Date 2014-03-01

Author Fengfeng Zhou, Youxi Luo, Qinghan Meng, Ruiquan Ge, Guoqin Mai, Jikui Liu

Maintainer Qinghan Meng <qinghan.meng@gmail.com>

Description The package implements a convenient procedure for microarray study, which is to do gene selection and classification simultaneously for binary outcomes. Users needn't to tune the parameters and can fix any genes that they desire to keep in the model. The K-folds cross validation results are returned.

License GPL

NeedsCompilation no

Repository CRAN

Date/Publication 2014-03-26 17:34:58

R topics documented:

SubLasso-package *SubLasso package*

Description

SubLasso package

Details

Package:	SubLasso
Type:	Package
Version:	1.0
Date:	2013-10-01
License:	GPL-2

A convient procedure for microarray studies. It can do feature selection and classification simultaneously for binary outcomes . K-folds cross validation results were returned for users. Users needn't to adjust the tune parameter and can fix any features that they desire to keep in the model.

Author(s)

Author: Fengfeng Zhou, Youxi Luo, Qinghan Meng, Ruiquan Ge, Guoqin Mai, Jikui Liu
Maintainer: Fengfeng Zhou <ff.zhou@siat.ac.cn>

References

1. Smyth GK: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology 2005, 3:Article3.
2. Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America 2001, 98(9):5116-5121.
3. Culhane AC, Perriere G, Considine EC, Cotter TG, Higgins DG: Between-group analysis of microarray data. Bioinformatics 2002, 18(12):1600-1608.
4. Parodi S, Muselli M, Fontana V, Bonassi S: ROC curves are a suitable and flexible tool for the analysis of gene expression profiles. Cytogenetic and genome research 2003, 101(1):90-91.
5. Wang Q, Wei F, Lv G, Li C, Liu T, Hadjipanayis CG, Zhang G, Hao C, Bellail AC: The association of TP53 mutations with the resistance of colorectal carcinoma to the insulin-like growth factor-1 receptor inhibitor picropodophyllin. Bmc Cancer 2013, 13(1):521.
6. Cortot AB, Younes M, Martel-Planche G, Guibert B, Isaac S, Souquet PJ, Commo F, Girard P, Fouret P, Brambilla E et al: Mutation of TP53 and Alteration of p14 Expression in EGFR- and KRAS-Mutated Lung Adenocarcinomas. Clinical lung cancer 2013.
7. Singamsetty GK, Malempati S, Bhogadhi S, Kondreddy R, Govatati S, Tangudu NK, Govatati S, Kuraganti AK, Bhanoori M, Kassetty K: TP53 alterations and colorectal cancer predisposition in south Indian population: A case-control study. Tumour biology : the journal of the International

Society for Oncodevelopmental Biology and Medicine 2013.
8. Chung R, Whaley J, Kley N, Anderson K, Louis D, Menon A, Hettlich C, Freiman R, Hedley-Whyte ET, Martuza R et al: TP53 gene mutations and 17p deletions in human astrocytomas. Genes, chromosomes & cancer 1991, 3(5):323-331.
9. Eeles RA, Bartkova J, Lane DP, Bartek J: The role of TP53 in breast cancer development. Cancer surveys 1993, 18:57-75.
10. Ding J, Shi J, Wu FX: SVM-RFE based feature selection for tandem mass spectrum quality assessment. International journal of data mining and bioinformatics 2011, 5(1):73-88.
11. Mundra PA, Rajapakse JC: SVM-RFE with MRMR filter for gene selection. IEEE transactions on nanobioscience 2010, 9(1):31-37.
12. Tang Y, Zhang YQ, Huang Z: Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM 2007, 4(3):365-381.
13. Zhou X, Tuck DP: MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. Bioinformatics 2007, 23(9):1106-1114.
14. Niijima S, Kuhara S: Recursive gene selection based on maximum margin criterion: a comparison with SVM-RFE. BMC bioinformatics 2006, 7:543.
15. Ding Y, Wilkins D: Improving the performance of SVM-RFE to select genes in microarray data. BMC bioinformatics 2006, 7 Suppl 2:S12.
16. Duan KB, Rajapakse JC, Wang H, Azuaje F: Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE transactions on nanobioscience 2005, 4(3):228-234.
17. Tibshirani R: Regression shrinkage and selection via the Lasso. J Roy Stat Soc B Met 1996, 58(1):267-288.
18. Friedman J, Hastie T, Tibshirani R: Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010, 33(1):1-22.

`Colon` *Gene expression data from Alon et al. (1999)*

Description

Gene expression data (2000 genes for 62 samples) from the microarray experiments of Colon tissue samples of Alon et al. (1999).

Usage

```
data(Colon)
```

Details

This data set contains 62 samples with 2000 genes: 40 tumor tissues, coded 2 and 22 normal tissues, coded 1.

Value

A list with the following elements:

X	a (62 x 2000) matrix giving the expression levels of 2000 genes for the 62 Colon tissue samples. Each row corresponds to a patient, each column to a gene.
Y	a numeric vector of length 62 giving the type of tissue sample (tumor or normal).
gene.names	a vector containing the names of the 2000 genes for the gene expression matrix X.

Source

The data are described in Alon et al. (1999) and can be freely downloaded from http://microarray.princeton.edu/oncology/affydata/index.html.

References

Alon, U. and Barkai, N. and Notterman, D.A. and Gish, K. and Ybarra, S. and Mack, D. and Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Natl. Acad. Sci. USA,**96**(12), 6745–6750.

Examples

```
# load plsgenomics library
library(SubLasso)

# load data set
data(Colon)

# how many samples and how many genes ?
dim(Colon$X)

# how many samples of class 1 and 2 respectively ?
sum(Colon$Y==1)
sum(Colon$Y==2)
```

Golub_Merge *Gene expression data from Alon et al. (1999)*

Description

Golub_Merge data

Usage

```
data(Golub_Merge)
```

Value

X	X
y	y

References

Alon, U. and Barkai, N. and Notterman, D.A. and Gish, K. and Ybarra, S. and Mack, D. and Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Natl. Acad. Sci. USA,**96**(12), 6745–6750.

Examples

```
# load plsgenomics library
library(SubLasso)

# load data set
data(Golub_Merge)
```

---

predict.sublasso *Predict method for SubLasso fits.*

Description

Similar to other predict methods, this functions predicts fitted values, logits, coefficients and more from a fitted "SubLasso" object.

Usage

```
## S3 method for class 'sublasso'
predict(object, xpred, type, s, ...)
```

Arguments

object	Fitted "SubLasso" model object.
xpred	Matrix of new values for x at which predictions are to be made. Must have the same row with x
type	type=c("link","response","class"). Default is "class";Type of prediction required. Type "link" gives the linear predictors; Type "response" gives the fitted probabilities; Type "class" produces the class label corresponding to the maximum probability.
s	Value(s) of the penalty parameter lambda at which predictions are required. Default is obtained by CV method.
...	Currently not used

Value

Predy	it depends on type.

SubLasso *Logistic model via Lasso penalty with a subset features*

Description

Fit a logistic model via Lasso penalty. A subset features can be fixed in the model.

Usage

```
SubLasso(X, y, subset, nfold)
```

Arguments

X	gene expression matrix, column is sample, row is gene(probe sets).
y	category vector, 1 (positive, illness) or 0 (negative, normal).
subset	gene (probe sets) names must be included in the model; Default is null set.
nfold	number of cross-validation; Default is 5.

Details

some details

Value

selname	features selected by the model.
w	the coefficient (weight) of feature in the model
valid	sensitivity (Sn), specificity (Sp), Accuracy (Acc), and Matthews correlation coefficient(Mcc)
description	the description statistics of selected features by group.
correlation	the correlations between all selected features.

Author(s)

Youxi Luo

References

Friedman, J., Hastie, T. and Tibshirani, R. (2008) Regularization Paths for Generalized Linear Modelsvia Coordinate Descent, http://www.stanford.edu/~hastie/Papers/glmnet.pdf Journal of Statistical Software, Vol. 33(1), 1-22 Feb 2010. http://www.jstatsoft.org/v33/i01/

See Also

glmnet

Examples

```
#screen device is not support in examples but SubLasso function use it to visualize results.
#when you want to test examples, please uncomment following code.

##### Example 1
#data(Golub_Merge)
#X <- Golub_Merge$X
#y <- Golub_Merge$y
#f1=SubLasso(X,y,nfold=10)

## predict.sublasso(f1,X[1:10,]) ##error predicted x
#predy=predict.sublasso(f1,X)
#predy=predict.sublasso(f1,X,type="class")
#predy=predict.sublasso(f1,X,type="link")
#predy=predict.sublasso(f1,X,type="response")
#predy=predict.sublasso(f1,X,type="response",s=0.05)
#subset=f1$selname
#f2=SubLasso(X,y,subset,nfold=10)

#subset=row.names(X)[1:10]
#f3=SubLasso(X,y,subset,nfold=10)

#predy=predict.sublasso(f3,X)
#predy=predict.sublasso(f3,X,type="class")
#predy=predict.sublasso(f3,X,type="link")
#predy=predict.sublasso(f3,X,type="response")
#predy=predict.sublasso(f3,X,type="response",s=0.05)

###Example 2

#data(Colon)
#X<-t(Colon$X)
#y_tmp<-Colon$Y
#y<-ifelse(y_tmp==1,1,0)
#f1=SubLasso(X,y,nfold=10)
#subset=f1$selname
#f2=SubLasso(X,y,subset,nfold=10)
#subset=row.names(X)[30:40]
#f3=SubLasso(X,y,subset,nfold=10)
```

Index