SURF

Lihui Liu

June 20, 2018

```
SURF(Xo,y,X=NULL,fold=10,Alpha=1,prop=0.1,weights=FALSE,B=1000,C=200,
   ncores=1,display.progress=TRUE,family=stats::binomial(link="logit"),pval=0.05)
```

Performs variable selection based on subsampling, ranking forward selection. Xo is the matrix of predictor variables. y is the response variable. Currently only binary responses using logistic regression are supported. X is a matrix of additional predictors which should be scaled to have sum 1 prior to analysis. fold is the number of folds for cross-validation. Alpha is the parameter for the elastic net method used in the subsampling procedure: the default value of 1 corresponds to LASSO. prop is the proportion of variables to remove in the each subsample. weights indicates whether observations should be weighted by class size. When the class sizes are unbalanced, weighting observations can improve results. B is the number of subsamples to use for ranking the variables. C is the number of permutations to use for estimating the critical value of the null distribution. If the doParallel package is installed, the function can be run in parallel by setting ncores to the number of threads to use. If the default value of 1 is used, or if the doParallel package is not installed, the function does not run in parallel. display.progress indicates whether the function should display messages indicating its progress. family is a family variable for the glm fitting. Note that the glmnet package does not permit the use of non-standard link functions, so will always use the default link function. However, the glm fitting will use the specified link. The default is binomial with logistic regression, because this is a common use case. pval is the p-value for inclusion of a variable in the model. Under the null case, the number of false positives will be geometrically distributed with this as probability of success, so if this parameter is set to p, the expected number of false positives should be $\frac{p}{1-p}$.

Example:

```
> library(SuRF)
> set.seed(1234)
> X<-rnorm(400)
> dim(X)<-c(80,5)
> v<-X%*%c(0,0.9,0,0,-1.1)
> X<-as.data.frame(X)
> names(X)<-c("var1","var2","var3","var4","var5")
```

```
> Y<-rbinom(80,1,exp(v)/(1+exp(v)))
> fitting<-SURF(X,Y,B=10,C=100,display.progress=FALSE)
> fitting

     variable deviance           cut-off            p-values
[1,] "var5"   "13.0678076018364" "6.68345066958575" "0"
[2,] ""       "4.57377902860846" "6.14536913391936" "0.14"
```