Package 'ESKNN'

September 13, 2015

Type Package

Title Ensemble of Subset of K-Nearest Neighbours Classifiers for Classification and Class Membership Probability Estimation

Version 1.0

Date 2015-09-13

Author Asma Gul, Aris Perperoglou, Zardad Khan, Osama Mahmoud, Werner Adler, Miftahuddin Miftahuddin, and Berthold Lausen

Maintainer Asma Gul <agul@essex.ac.uk>

Description Functions for classification and group membership probability estimation are given. The issue of non-informative features in the data is addressed by utilizing the ensemble method. A few optimal models are selected in the ensemble from an initially large set of base k-nearest neighbours (KNN) models, generated on subset of features from the training data. A two stage assessment is applied in selection of optimal models for the ensemble in the training function.
The prediction functions for classification and class membership probability estimation returns class outcomes and class membership probability estimates for the test data.
The package includes measure of classification error and brier score, for classification and probability estimation tasks respectively.

Imports caret,stats

LazyLoad yes

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2015-09-13 09:22:47

R topics documented:

ESkNN-package 2
esknnClass 2
esknnProb 4
hepatitis 6
Predict.esknnClass 7

Predict.esknnProb 8
sonar 10

Index **12**

`ESkNN-package` *Ensemble of Subset of K-Nearest Neighbours Classifiers for Classification and Class Membership Probability Estimation*

Description

Functions for building an ensemble of optimal k-nearest neighbours (kNN) models for classification and class membership probability estimation are provided. To address the issue of non-informative features in the data. A set of base kNN models is generated and a subset of models is selected for the ensemble based on the individual and combined performance of these models. Out-of-bag data and an independent training data set is used for the performance assessment of models individually and collectively. Class labels and class membership probability estimates are returned by the prediction functions. Other measures such as confusion matrix, classification error rate, and brier scores etc, are also returned by the functions.

Details

Package:	ESKNN
Type:	Package
Version:	1.0
Date:	2015-09-13
License:	GPL (>= 2)

Author(s)

Asma Gul, Aris Perperoglou, Zardad Khan, Osama Mahmoud, Miftahuddin, Werner Adler, and Berthold Lausen Maintainer: Asma Gul <agul@essex.ac.uk>

References

Gul, A., Perperoglou, A., Khan, Z., Mahmoud, O., Miftahuddin, M., Adler, W. and Lausen, B.(2014),*Ensemble of subset of k-nearest neighbours classifiers*, Journal name to appear.

`esknnClass` *Train ensemble of subset of k-nearest neighbours classifiers for classification*

Description

Constructing m, models and search for the optimal models for classification.

Usage

```
esknnClass(xtrain, ytrain, k = NULL, q = NULL, m = NULL, ss = NULL)
```

Arguments

xtrain	A matrix or data frame of size n x d dimension where n is the number of traing observation and d is the number of features.
ytrain	A vector of class labels of the training data. Class labels should be factor of two levels (0,1) represented by variable Class in the data..
k	Number of nearest neighbours to be considered, when NULL then the default is set tok=3.
q	Percent of models to be selected from the initial set m.
m	Number of models to be generated in the first stage, when NULL the default is m=501.
ss	Feature subset size to be selected from d features for each bootstrap sample, when NULL the default is (number of features)/3.

Value

trainfinal	List of the extracted opimal models.
fsfinal	List of the features used in each selected models.

Author(s)

Asma Gul <agul@essex.ac.uk>

References

Gul, A., Perperoglou, A., Khan, Z., Mahmoud, O.,Miftahuddin, M., Adler, W. and Lausen, B.(2014), Ensemble of Subset of kNN Classifiers, Journal name to appear.

See Also

Predict.esknnClass

Examples

```
# Load the data

  data(hepatitis)
  data <- hepatitis

# Divide the data into testing and training parts
```

```
  Class <- data[,names(data)=="Class"]
  data$Class<-as.factor(as.numeric(Class)-1)
  train <- data[sample(1:nrow(data),0.7*nrow(data)),]
  test <- data[-(sample(1:nrow(data),0.7*nrow(data))),]
  ytrain<-train[,names(train)=="Class"]
  xtrain<-train[,names(train)!="Class"]
  xtest<-test[,names(test)!="Class"]
  ytest <- test[,names(test)=="Class"]

# Trian esknnClass

  model<-esknnClass(xtrain, ytrain,k=NULL)

# Predict on test data

  resClass<-Predict.esknnClass(model,xtest,ytest,k=NULL)

# Returning Objects are predicted class labels, confusion matrix and classification error

  resClass$PredClass
  resClass$ConfMatrix
  resClass$ClassError
```

---

esknnProb *Train the ensemble of subset of k-nearest neighbours classifiers for estimation of class membership probabilty.*

Description

This function selects a subset of optimal models from a set of `m` models, initially generated on bootstrap sample with a random feature subset from the training data, for class membership probability estimation. The values for the hyper parameters, for example subset size of the best models from the total initial `m` models, can be specified by the user otherwise the default values are considered.

Usage

```
esknnProb(xtrain, ytrain, k = NULL, q = NULL, m = NULL, ss = NULL)
```

Arguments

`xtrain`	A matrix or data frame of size `n x d` dimension where `n` is the number of traing observation and `d` is the number of features.
`ytrain`	A vector of class labels for the training data. Class labels should be factor of two levels (0,1) represented by variable `Class` in the data.
`k`	Number of nearest neighbours to be considered, when `NULL` then the default is set to`k=3`.
`q`	Percent of models to be selected from the initial set `m`.

m	Number of models to be generated in the first stage, when NULL the default is m=501.
ss	Feature subset size to be selected from d features for each bootstrap sample, when NULL the default is (number of features)/3.

Value

trainfinal	List of the extracted opimal models.
fsfinal	List of the features used in each selected models.

Author(s)

Asma Gul <agul@essex.ac.uk>

References

Gul, A., Perperoglou, A., Khan, Z., Mahmoud, O.,Miftahuddin, M., Adler, W. and Lausen, B.(2014),Ensemble of Subset of kNN Classifiers, Journal name to appear.

See Also

Predict.esknnProb

Examples

```
# Load the data

  data(sonar)
  data <- sonar

# Divide the data into testing and training

  Class <- data[,names(data)=="Class"]
  data$Class<-as.factor(as.numeric(Class)-1)
  train <- data[sample(1:nrow(data),0.7*nrow(data)),]
  test <- data[-(sample(1:nrow(data),0.7*nrow(data))),]
  ytrain<-train[,names(train)=="Class"]
  xtrain<-train[,names(train)!="Class"]
  xtest<-test[,names(test)!="Class"]
  ytest <- test[,names(test)=="Class"]

# Trian esknnProb on training data

  model<-esknnProb(xtrain, ytrain,k=NULL)

# Predict on test data

  resProb<-Predict.esknnProb(model,xtest,ytest,k=NULL)

## Returning Objects
```

```
resProb$PredProb
resProb$BrierScore
```

`hepatitis` *Hepatitis data set*

Description

This data set is about hepatitus disease. The data set is obtained from UCI machine learning repository. There are 155 observations in total, however this data set consists of 80 observations after removing the observations with missing values. There are 19 features/ attributes where 13 attributes are binary while 6 attributes are discrete valued. The observations are catogarized in two classes classes die and live. There are 13 observations in class "die" and "67" in class live.

Usage

```
data(hepatitis)
```

Format

A data frame with 80 observations on the following 20 variables.

`Age` age of the patients in years, from 20 to 80 years.

`Sex` Gender of patient, a factor at two levels coded by 1 (male) and 2(female)

`Steroid` Steroid treatment, a factor at two levels coded by 1(yes) and 2(no) .

`Antivirals` Antivirals medication, a factor at two levels 1 (yes) and 2 (no).

`Fatigue` Fatigue is a frequent and disabling symptom reported by patients with chronic hepatitis, a factor at two levels 1 (yes) and 2 (no).

`Malaise` Malaise one of the symptoms of hepatitis, a factor at two levels 1 (yes) and 2 (no).

`Anorexia` Anorexia, loss of appetite, a factor at two levels 1 (yes) and 2 (no).

`LiverBig` The size of liver increased or fatty, a factor at two levels 1 (yes) and 2 (no).

`LiverFirm` A factor at two levels 1 (yes) and 2 (no).

`SpleenPalpable` Splenomegaly is an enlargement of the spleen, a factor at two levels 1 (yes) and 2 (no).

`Spiders` Enlarged blood vessels that resemble little spiders,a factor at two levels 1 (yes) and 2 (no).

`Ascites` Ascites is the presence of excess fluid in the peritoneal cavity, a factor at two levels 1(yes) and 2(no)).

`Varices` a factor at two levels 1(yes) and 2(no)).

`Bilirubin` Bilirubin is a substance made when the body breaks down old red blood cells, continuous feature

`AlkPhosphate` Alkaline phosphatase is an enzyme made in liver cells and bile ducts, a discrete valued feature reveals level Alkaline phosphatase.

`Sgot` A discrete valued feature.

`AlbuMin` A continous feature.

`ProTime` A discrete valued feature.

`Histology` a factor at two levels 1 (yes) and 2 (no).

`Class` a factor at two levels 1(Die) or 2(Live).

Source

This data set is available on: https://archive.ics.uci.edu/ml/datasets/Hepatitis

Examples

```
data(hepatitis)
str(hepatitis)
```

Predict.esknnClass *Class predictions from ensemble of subset of k-nearest neighbours*

Description

Classification prediction for test data on the trained `esknnClass` object for.

Usage

```
Predict.esknnClass(optModels, xtest, ytest=NULL, k = NULL)
```

Arguments

`optModels`	An object of esknnClass
`xtest`	A matrix or data frame test set features/attributes.
`ytest`	Optional: A vector of lenth `m` consisting of class labels for the test data. Should be binary (0,1), reprenting by a variable `Class` in the data. If provided then confusion matrix and classification error rate is returned.
`k`	Number of nearest neighbors considered. The same value is considered as for training in `esknnClass`.

Value

`predClass`	A vector of predicted classes of test set observations.
`ConfMatrix`	Confusion matrix return a matrix of cross classification counts based on the estimated class labels and the true class labels of test observations. This matrix is returned if ytest is given.
`ClassError`	Classification error rate of the clssifier for test set observations. This is returned if ytest is provided.

Author(s)

Asma Gul <agul@essex.ac.uk>

References

Gul, A., Perperoglou, A., Khan, Z., Mahmoud, O.,Miftahuddin, M., Adler, W. and Lausen, B.(2014),Ensemble of Subset of kNN Classifiers, Journal name to appear.

See Also

esknnClass

Examples

```
# Load the data

  data(hepatitis)
  data <- hepatitis

# Spliting the data into testing and training parts.

  Class <- data[,names(data)=="Class"]
  data$Class<-as.factor(as.numeric(Class)-1)
  train <- data[sample(1:nrow(data),0.7*nrow(data)),]
  test <- data[-(sample(1:nrow(data),0.7*nrow(data))),]
  ytrain<-train[,names(train)=="Class"]
  xtrain<-train[,names(train)!="Class"]
  xtest<-test[,names(test)!="Class"]
  ytest <- test[,names(test)=="Class"]

# Trian esknnClass using training data

model<-esknnClass(xtrain, ytrain,k=NULL)

# Predict on test data

resClass<-Predict.esknnClass(model,xtest,ytest,k=NULL)

# Returning Objects are predicted class labels, confusion matrix and classification error

  resClass$predClass
  resClass$ConfMatrix
  resClass$ClassError
```

Predict.esknnProb *Prediction function returning class membership probability estimates*

Description

This function provides class membership probability estimates for the test set observations.

Usage

```
Predict.esknnProb(optModels, xtest, ytest, k = NULL)
```

Arguments

optModels	An object of class esknnProb.
xtest	A matrix or data frame test set features/attributes.
ytest	Optional: A vector of class labels for the test data. Class labels should be factor of two levels (0,1) represented by variable Class in the data. The Brier score is returned if this vector is given.
k	Number of nearest neighbors considered. The same value should be considered as for training in esknnProb

Value

PredProb	A vector of estimated class membership probabilities of test set observations.
BrierScore	A vector of Brier Score based on the estimated probabilities and true class label of test set observations. This vector is returned if ytest is given.

Author(s)

Asma Gul <agul@essex.ac.uk>

References

ul, A., Perperoglou, A., Khan, Z., Mahmoud, O.,Miftahuddin, M., Adler, W. and Lausen, B.(2014),Ensemble of Subset of kNN Classifiers, Journal name to appear.

See Also

esknnProb

Examples

```
# Load the data

data(sonar)
data <- sonar

# Divide the data into testing and training parts

  Class <- data[,names(data)=="Class"]

# Class Varible must be a factor in (0,1)
```

```
  data$Class<-as.factor(as.numeric(Class)-1)
  train <- data[sample(1:nrow(data),0.7*nrow(data)),]
  test <- data[-(sample(1:nrow(data),0.7*nrow(data))),]
  ytrain<-train[,names(train)=="Class"]
  xtrain<-train[,names(train)!="Class"]
  xtest<-test[,names(test)!="Class"]
  ytest <- test[,names(test)=="Class"]

# Trian esknnProb

  model<-esknnProb(xtrain, ytrain,k=NULL)

# Predict on test data

  resProb<-Predict.esknnProb(model,xtest,ytest,k=NULL)

## Returning Objects

  resProb$PredProb
  resProb$BrierScore
```

---

sonar *Sonar, Mines vs. Rocks.*

Description

This data set is a collection of sonar signals, coded as 60 continuous attributes on 208 observations. The sonar signals are obtained from a variety of different aspect angles, spanning 90 degrees for mines and 180 degrees for rocks. The task is classification of sonar signals in two catagories, signals bounced off a "rock" or a "metal cylinder". Each pattern in the data is a set of 60 numbers (continous) in the range 0.0 to 1.0, where each number represents the energy within a particular frequency band, integrated over a certain period of time. From total 208 observations, 111 obtained by bouncing sonar signals off a metal cylinder at various angles and under various conditions, is labled with "M" and 97 patterns obtained from rocks under similar conditions is labled with "R".

Usage

```
data(sonar)
```

Format

A data frame with 208 observations on 60 features/attributes in two classes. All the features are numerical and the class is nominal.

Source

This data set is available on: ftp://ftp.ics.uci.edu/pub/machine-learning-databases http://sci2s.ugr.es/keel/dataset.php?cod=85

References

Gorman, R. P., and Sejnowski, T. J. (1988). "Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets" in Neural Networks, Vol. 1, pp. 75-89.

Friedrich Leisch & Evgenia Dimitriadou (2010). mlbench: Machine Learning Benchmark Problems. R package version 2.1-1.

Examples

```
data(sonar)
str(sonar)
```

Index

∗Topic **Predict.esknnProb**
    Predict.esknnProb, 8
∗Topic **datasets**
    hepatitis, 6
    sonar, 10
∗Topic **esknnClass**
    esknnClass, 2
    esknnProb, 4
∗Topic **esknn**
    esknnClass, 2
    esknnProb, 4
    Predict.esknnProb, 8
∗Topic **packages**
    Predict.esknnClass, 7
∗Topic **package**
    ESkNN-package, 2

ESkNN *(*ESkNN-package*)*, 2
ESkNN-package, 2
esknnClass, 2, *8*
esknnProb, 4, *9*

hepatitis, 6

Predict.esknnClass, *3*, 7
Predict.esknnProb, *5*, 8

sonar, 10