The Kendall Interaction Filter for Variable Interaction Screening in Ultra High Dimensional Classification Problems

Youssef Anzarmou, Abdallah Mkhadri & Karim Oualkacha

Introduction

This is a package implementing the *Kendall Interaction Filter* (*KIF*), an efficient interaction screening method aiming to select the relevant couples to the classification task in the high dimensional data frame. The measure *KIF* is presented in the paper "The Kendall Interaction Filter for Variable Interaction Screening in Ultra High Dimensional Classification Problems". It ranges from 0 to 1 and has several advantages:

- It is model-free.
- It has the ability to process both continuous and categorical features.
- It has the sure screening property.
- It is heredity-assumption free.
- It is robust against heavy-tailed distributions.

The **KIF** package implements two methods; namely **KIF** and **KIFall**. The method **KIF** takes a couple as an input and returns its *Kendall Interaction Filter* as an output while the method **KIFall** takes as an input the complete dataset and returns the most relevant couples, to the classification task, as an output.

Installation

To install **KIF** package, run:

```
library(devtools)
devtools::install_github("KarimOualkacha/KIF", build_vignettes = TRUE)

library(KIF)
## Loading required package : mvtnorm
## Loading required package : ccaPP
## Loading required package : parallel
## Loading required package : pcaPP
## Loading required package : robustbase
```

Quick start example

Toy example

We generate a toy dataset to illustrate the usage of the functions **KIF** and **KIFall**. the dataset has 200 observations and 500 explanatory variables. It is a two class example where each class has 100 observations.

```
library(mvtnorm)
set.seed(1)
n1 <- 100
n2 <- 100
n <- n1 +n2
```

```
p <- 500
sigma <- diag(p)
sigma[upper.tri(sigma)] <- 0.2
sigma[lower.tri(sigma)] <- 0.2
sigma1 <- sigma
sigma2 <- sigma
sigma1[1,2] <- 0.8
sigma1[2,1] <- 0.8
sigma1[3,4] <- 0.8
sigma1[4,3] <- 0.8
sigma2[3,4] <- -0.8
sigma2[4,3] <- -0.8
mean1 <- c(rep(0,p))
mean2 <- c(rep(0,p))
Sample <- rbind(rmvnorm(n1, mean1, sigma1), rmvnorm(n2, mean2, sigma2))
y <- c(rep(1,n1), rep(0,n2))
```

The relevenant couples, to the classification task, are "1,2" and "3,4". Couple "3,4" is more relevant than "1,2".

KIF function

The **KIF** function requires as arguments a pair of explanatory variables and the label variables and returns as an output the corresponding *Kendall Interaction Filter* score.

```
couple12 <- Sample[,1:2]
couple34 <- Sample[,3:4]
out12 <- KIF(couple12,y)
out34 <- KIF(couple34,y)
```

The result is:

```
out12
## [1] 0.199798
out34
## [1] 0.5165657
```

Kendall Interaction Filter score of couple "3,4" is higher than that of couple "1,2", as expected.

KIFall function

The **KIFall** function requires as arguments the dataset, the label variable, the number of cores to use for parallelization and the number of of pairs to select among the first selected couples. It returns as an output the couples selected as relevant ones based on thier decreasing *Kendall Interaction Filter* scores order.

```
outall <- KIFall(Sample, y, 2, 10)
outall
## [1] "3,4" "3,263" "1,2" "59,475" "42,350" "162,303" "393,483" "42,101"  "223,437" "246,366"
```

Couples "1,2" and "3,4" are both among the first 10 selected couples, impliying that *Kendall Interaction Filter* indeed has the ability to select the relevant couples.