

R Notebook

Parametros:

Measure = Matthews correlation coefficient
Columns = learner
Performance = holdout_measure_residual
Filter keys = sampling, weight_space, underbagging
Filter values = FALSE, FALSE, FALSE

```
library("scmamp")  
library(dplyr)
```

Tratamento dos dados

Carregando data set compilado

```
ds = read.csv("/home/rodrigo/Dropbox/UNICAMP/IC/estudo_cost_learning/SummaryResults/summary_compilation.  
ds = filter(ds, learner != "classif.rusboost")  
summary(ds)
```

```
##           learner      weight_space  
## classif.ksvm      :17100  Mode :logical  
## classif.randomForest:17100 FALSE:41040  
## classif.rusboost   :    0  TRUE :10260  
## classif.xgboost    :17100  NA's :0  
##  
##  
##  
##           measure      sampling      underbagging  
## Accuracy           :10260  ADASYN:10260  Mode :logical  
## Area under the curve :10260  FALSE :30780  FALSE:41040  
## F1 measure           :10260  SMOTE :10260  TRUE :10260  
## G-mean              :10260           NA's :0  
## Matthews correlation coefficient:10260  
##  
##  
## tuning_measure  holdout_measure  holdout_measure_residual  
## Min.   :-0.1277  Min.   :-0.2120  Min.   :-0.4658  
## 1st Qu.: 0.6911  1st Qu.: 0.4001  1st Qu.: 0.1994  
## Median : 0.9700  Median : 0.8571  Median : 0.5581  
## Mean   : 0.7903  Mean   : 0.6718  Mean   : 0.5298  
## 3rd Qu.: 0.9975  3rd Qu.: 0.9900  3rd Qu.: 0.8755  
## Max.   : 1.0000  Max.   : 1.0000  Max.   : 1.0000  
## NA's   :1077    NA's   :1077    NA's   :1077  
## iteration_count      dataset      imba.rate  
## Min.   :1          abalone      : 900  Min.   :0.0010  
## 1st Qu.:1          adult      : 900  1st Qu.:0.0100  
## Median :2          bank      : 900  Median :0.0300  
## Mean   :2          car      : 900  Mean   :0.0286
```

```
## 3rd Qu.:3      cardiocography-10clases: 900 3rd Qu.:0.0500
## Max. :3      cardiocography-3clases : 900 Max. :0.0500
## NA's :1077 (Other) :45900
```

Filtrando pela metrica

```
ds = filter(ds, measure == params$measure)
```

Filtrando o data set

```
if(params$filter_keys != 'NULL' && !is.null(params$filter_keys)){
  dots = paste0(params$filter_keys, " == '",params$filter_values, "'")
  ds = filter_(ds, .dots = dots)
}
```

```
summary(ds)
```

```
##          learner      weight_space
## classif.ksvm      :684 Mode :logical
## classif.randomForest:684 FALSE:2052
## classif.rusboost   : 0 NA's :0
## classif.xgboost    :684
##
##
##
##          measure      sampling      underbagging
## Accuracy          : 0 ADASYN: 0 Mode :logical
## Area under the curve : 0 FALSE :2052 FALSE:2052
## F1 measure          : 0 SMOTE : 0 NA's :0
## G-mean              : 0
## Matthews correlation coefficient:2052
##
##
## tuning_measure      holdout_measure      holdout_measure_residual
## Min. : -0.02717 Min. : -0.05263 Min. : -0.4571
## 1st Qu.: 0.05889 1st Qu.: 0.00000 1st Qu.: 0.0000
## Median : 0.49619 Median : 0.56796 Median : 0.1669
## Mean : 0.47012 Mean : 0.47690 Mean : 0.2775
## 3rd Qu.: 0.81018 3rd Qu.: 0.86046 3rd Qu.: 0.4871
## Max. : 1.00000 Max. : 1.00000 Max. : 1.0000
## NA's :18 NA's :18 NA's :18
## iteration_count      dataset      imba.rate
## Min. :1 abalone : 36 Min. :0.0010
## 1st Qu.:1 adult : 36 1st Qu.:0.0100
## Median :2 bank : 36 Median :0.0300
## Mean :2 car : 36 Mean :0.0286
## 3rd Qu.:3 cardiocography-10clases: 36 3rd Qu.:0.0500
## Max. :3 cardiocography-3clases : 36 Max. :0.0500
## NA's :18 (Other) :1836
```

Computando as médias das iteracoes

```
ds = group_by(ds, learner , weight_space , measure , sampling , underbagging , dataset , imba.rate)
ds = summarise(ds, tuning_measure = mean(tuning_measure), holdout_measure = mean(holdout_measure),
               holdout_measure_residual = mean(holdout_measure_residual))

ds = as.data.frame(ds)
```

Criando dataframe

```
# Dividindo o ds em n, um para cada técnica
splited_df = ds %>% group_by_at(.vars = params$columns) %>% do(vals = as.data.frame(.)) %>% select(vals)

# Juntando cada uma das partes horizontalmente em um data set
df_tec_wide = do.call("cbind", splited_df)

# Renomeando duplicacao de nomes
colnames(df_tec_wide) = make.unique(colnames(df_tec_wide))

# Selecionando apenas as medidas da performance escolhida
df_tec_wide_residual = select(df_tec_wide, matches(paste("^", params$performance, "$|", params$performance)))

# Renomeando colunas
new_names = NULL
for(i in (1:length(splited_df))){
  id = toString(sapply(splited_df[[i]][1, params$columns], as.character))
  new_names = c(new_names, id)
}
colnames(df_tec_wide_residual) = new_names

# Verificando a dimensao do df
dim(df_tec_wide_residual)
```

```
## [1] 228 3
```

```
# Renomeando a variavel
df = df_tec_wide_residual
```

```
head(df)
```

```
##   classif.ksvm classif.randomForest classif.xgboost
## 1  0.03701282      0.005300235      0.005300235
## 2  0.03701282      0.005300235      0.005300235
## 3  0.05529460      0.000000000      0.021544091
## 4  0.08009145      0.000000000      0.017705096
## 5  0.09710987      0.323620415      0.302563452
## 6  0.09710987      NA           0.302563452
```

```
summary(df)
```

```
##   classif.ksvm      classif.randomForest classif.xgboost
## Min.      :-0.26935 Min.      :-0.34919 Min.      :-0.3993
## 1st Qu.: 0.00000 1st Qu.: 0.05113 1st Qu.: 0.0723
## Median : 0.08433 Median : 0.22407 Median : 0.2587
## Mean    : 0.19174 Mean    : 0.30998 Mean    : 0.3317
## 3rd Qu.: 0.27632 3rd Qu.: 0.52248 3rd Qu.: 0.5229
## Max.    : 0.99489 Max.    : 1.00000 Max.    : 0.9974
## NA's    :6
```

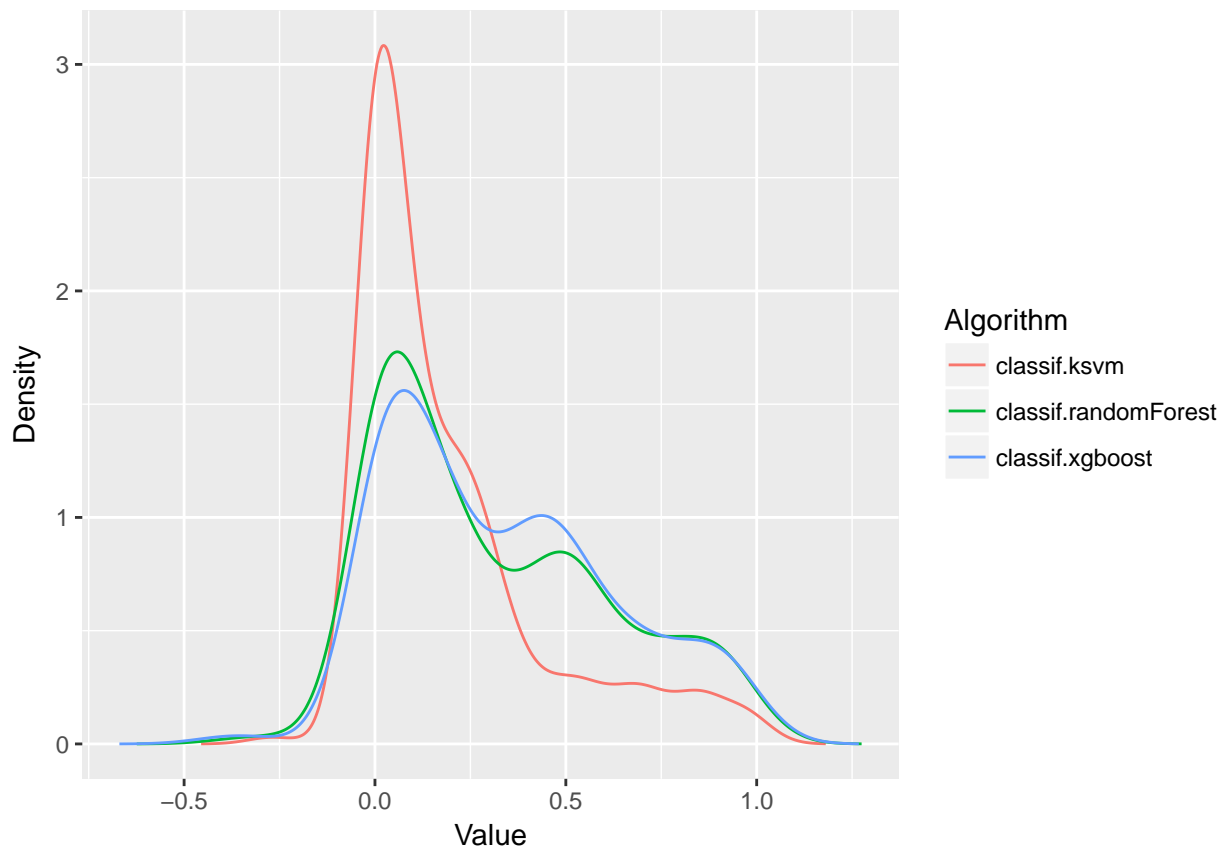
Verificando a média de cada coluna selecionada

```
for(i in (1:dim(df)[2])){
  print(paste("Media da coluna ", colnames(df)[i], " = ", mean(df[,i], na.rm = TRUE), sep=""))
}
```

```
## [1] "Media da coluna classif.ksvm = 0.19173654981395"
## [1] "Media da coluna classif.randomForest = 0.309979346079034"
## [1] "Media da coluna classif.xgboost = 0.331714283128261"
```

Fazendo teste de normalidade

```
plotDensities(data = na.omit(df))
```



Testando as diferenças

```
friedmanTest(df)
```

```
##
## Friedman's rank sum test
##
## data: df
## Friedman's chi-squared = 70.243, df = 2, p-value = 5.551e-16
```

Testando as diferenças par a par

```
test <- nemenyiTest(df, alpha=0.05)
abs(test$diff.matrix) > test$statistic
```

```
##      classif.ksvm classif.randomForest classif.xgboost
## [1,]          FALSE                TRUE             TRUE
## [2,]           TRUE                FALSE             TRUE
## [3,]           TRUE                TRUE             FALSE
```

Plotando os ranks

```
print(colMeans(rankMatrix(df)))
```

```
##      classif.ksvm classif.randomForest      classif.xgboost
##      2.407895      1.967105      1.625000
```

Plotando grafico de Critical Difference

```
result = tryCatch({
  plotCD(df, alpha=0.05, cex = 0.35)
}, error = function(e) {})
```

