

R Notebook

Parametros:

```
Measure = Accuracy
Columns = learner
Performance = holdout_measure
Filter keys = sampling, weight_space, underbagging, imba.rate
Filter values = FALSE, FALSE, FALSE, 0.001
```

```
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.csv")
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      :120  Mode :logical
## classif.randomForest:120 FALSE:360
## classif.rusboost   : 0  NA's :0
## classif.xgboost    :120
##
##
##
##          measure      sampling  underbagging
## Accuracy           :360  ADASYN: 0  Mode :logical
## Area under the curve : 0  FALSE :360  FALSE:360
## F1 measure          : 0  SMOTE : 0  NA's :0
## G-mean              : 0
## Matthews correlation coefficient: 0
##
##
## tuning_measure  holdout_measure  holdout_measure_residual
## Min. :0.9857  Min. :0.9800  Min. :0.03881
## 1st Qu.:0.9900  1st Qu.:0.9909  1st Qu.:0.32577
## Median :0.9920  Median :0.9929  Median :0.73202
## Mean :0.9933  Mean :0.9941  Mean :0.64527
## 3rd Qu.:0.9961  3rd Qu.:0.9974  3rd Qu.:0.97711
## Max. :1.0000  Max. :1.0000  Max. :1.00000
## NA's :6      NA's :6      NA's :6
## iteration_count      dataset      imba.rate
## Min. :1      abalone      : 9  Min. :0.001
## 1st Qu.:1      adult      : 9  1st Qu.:0.001
## Median :2      bank      : 9  Median :0.001
## Mean :2      car      : 9  Mean :0.001
## 3rd Qu.:3      cardiocography-10clases: 9  3rd Qu.:0.001
## Max. :3      cardiocography-3clases : 9  Max. :0.001
## NA's :6      (Other) :306
```

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] 40 3
```

```
# Renomeando a variavel
df = df_tec_wide_residual

head(df)
```

```
##   classif.ksvm classif.randomForest classif.xgboost
## 1    0.9898629         0.9910555         0.9910555
## 2    0.9892883              NA         0.9918993
## 3    0.9900000         0.9900000         0.9900000
## 4    0.9979101         1.0000000         1.0000000
## 5    0.9941667         0.9916667         0.9925000
## 6    0.9964820         1.0000000         0.9982410
```

```
summary(df)
```

```
##   classif.ksvm   classif.randomForest classif.xgboost
## Min.   :0.9893   Min.   :0.9883         Min.   :0.9850
## 1st Qu.:0.9902   1st Qu.:0.9911         1st Qu.:0.9916
## Median :0.9916   Median :0.9947         Median :0.9943
## Mean   :0.9931   Mean   :0.9947         Mean   :0.9945
## 3rd Qu.:0.9960   3rd Qu.:0.9991         3rd Qu.:0.9981
## Max.   :1.0000   Max.   :1.0000         Max.   :1.0000
## NA's    :2
```

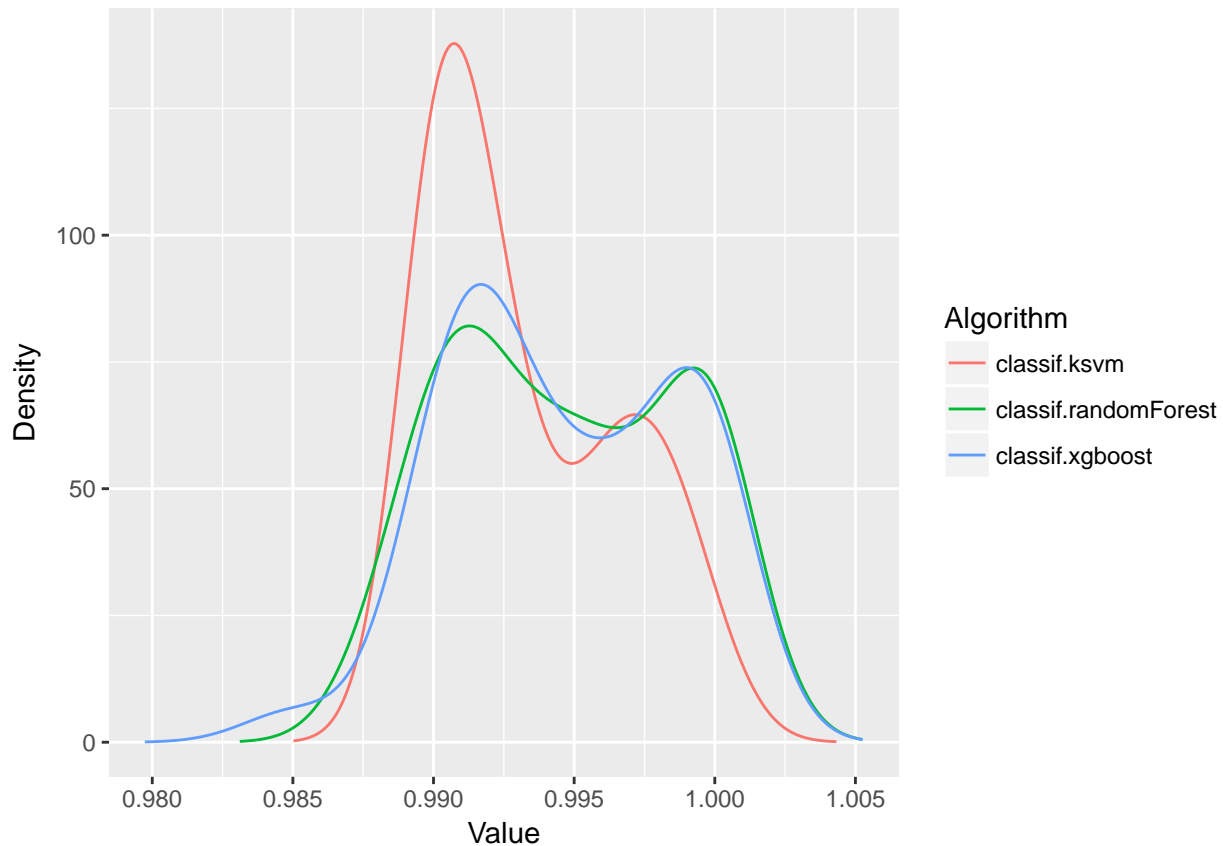
Verificando a média de cada coluna selecionada

```
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 = 0.993062611875091"
## [1] "Media da coluna classif.randomForest = 0.994689893128772"
## [1] "Media da coluna classif.xgboost = 0.994517387753698"
```

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 = 5.55, df = 2, p-value = 0.06235
```

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                  FALSE          FALSE
## [2,]          FALSE                  FALSE          FALSE
## [3,]          FALSE                  FALSE          FALSE
```

Plotando os ranks

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

```
##      classif.ksvm classif.randomForest      classif.xgboost
##              2.275              1.975              1.750
```

Plotando grafico de Critical Difference

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

