

R Notebook

Parametros:

Measure = Area under the curve
Columns = sampling, weight_space, ruspool, learner
Performance = tuning_measure
Filter keys = NULL
Filter values = NULL

```
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.  
summary(ds)
```

```
##           learner      weight_space  
## classif.ksvm      :17100  Mode :logical  
## classif.randomForest:17100 FALSE:41040  
## classif.xgboost    :17100  TRUE :10260  
##                   NA's :0  
##  
##  
##  
##           measure      sampling      ruspool  
## 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.: 0.5924  1st Qu.: 0.3114  1st Qu.: 0.1648  
## Median : 0.9624  Median : 0.8193  Median : 0.5192  
## Mean   : 0.7570  Mean   : 0.6469  Mean   : 0.5099  
## 3rd Qu.: 0.9965  3rd Qu.: 0.9879  3rd Qu.: 0.8636  
## Max.    : 1.0000  Max.    : 1.0000  Max.    : 1.0000  
## NA's    :1761    NA's    :1761    NA's    :1761  
## 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      cardiotocography-10clases: 900  3rd Qu.:0.0500  
## Max.    :3      cardiotocography-3clases : 900  Max.    :0.0500
```

```
## NA's :1761 (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)){  
  ds = filter_at(ds, .vars = params$filter_keys, .vars_predicate = any_vars(. == params$filter_values))  
}
```

```
summary(ds)
```

```
##           learner      weight_space  
## classif.ksvm      :3420  Mode :logical  
## classif.randomForest:3420 FALSE:8208  
## classif.xgboost    :3420  TRUE :2052  
##                   NA's :0  
##  
##  
##  
##           measure      sampling      ruspool  
## Accuracy           : 0  ADASYN:2052  Mode :logical  
## Area under the curve :10260 FALSE :6156  FALSE:8208  
## F1 measure           : 0  SMOTE :2052  TRUE :2052  
## G-mean              : 0                   NA's :0  
## Matthews correlation coefficient: 0  
##  
##  
## tuning_measure  holdout_measure  holdout_measure_residual  
## Min. :0.3023  Min. :0.0000  Min. :0.0000  
## 1st Qu.:0.9030  1st Qu.:0.8213  1st Qu.:0.6821  
## Median :0.9949  Median :0.9770  Median :0.8800  
## Mean :0.9182  Mean :0.8816  Mean :0.8174  
## 3rd Qu.:0.9999  3rd Qu.:0.9996  3rd Qu.:0.9798  
## Max. :1.0000  Max. :1.0000  Max. :1.0000  
## NA's :384    NA's :384    NA's :384  
## iteration_count      dataset      imba.rate  
## Min. :1      abalone      : 180  Min. :0.0010  
## 1st Qu.:1      adult      : 180  1st Qu.:0.0100  
## Median :2      bank      : 180  Median :0.0300  
## Mean :2      car      : 180  Mean :0.0286  
## 3rd Qu.:3      cardiotocography-10clases: 180  3rd Qu.:0.0500  
## Max. :3      cardiotocography-3clases : 180  Max. :0.0500  
## NA's :384    (Other)      :9180
```

Computando as médias das iteracoes

```
ds = group_by(ds, learner , weight_space , measure , sampling , ruspool , 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 15
```

```

# Renomeando a variavel
df = df_tec_wide_residual

summary(df)

```

```

## ADASYN, FALSE, FALSE, classif.ksvm
## Min. :0.7759
## 1st Qu.:0.9994
## Median :1.0000
## Mean :0.9942
## 3rd Qu.:1.0000
## Max. :1.0000
## NA's :14
## ADASYN, FALSE, FALSE, classif.randomForest
## Min. :0.7697
## 1st Qu.:0.9988
## Median :0.9999
## Mean :0.9943
## 3rd Qu.:1.0000
## Max. :1.0000
## NA's :36
## ADASYN, FALSE, FALSE, classif.xgboost FALSE, FALSE, FALSE, classif.ksvm
## Min. :0.5148 Min. :0.4719
## 1st Qu.:0.9952 1st Qu.:0.7401
## Median :0.9994 Median :0.9261
## Mean :0.9818 Mean :0.8578
## 3rd Qu.:1.0000 3rd Qu.:0.9950
## Max. :1.0000 Max. :1.0000
## NA's :5
## FALSE, FALSE, FALSE, classif.randomForest

```

```

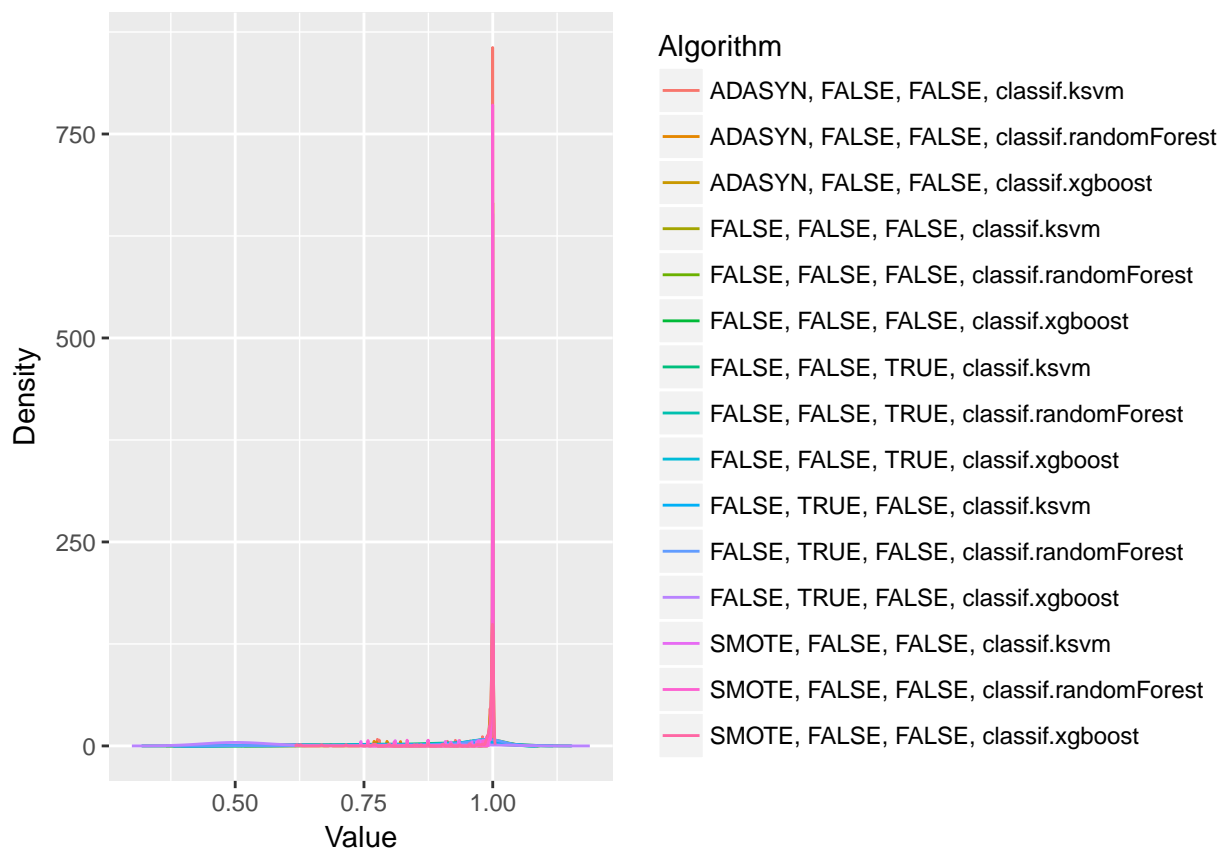
## Min.      :0.4564
## 1st Qu.:0.8885
## Median :0.9849
## Mean    :0.9242
## 3rd Qu.:0.9989
## Max.    :1.0000
## NA's    :7
## FALSE, FALSE, FALSE, classif.xgboost FALSE, FALSE, TRUE, classif.ksvm
## Min.      :0.4689      Min.      :0.5126
## 1st Qu.:0.8932      1st Qu.:0.7878
## Median :0.9773      Median :0.8779
## Mean    :0.9258      Mean    :0.8602
## 3rd Qu.:0.9982      3rd Qu.:0.9530
## Max.    :1.0000      Max.    :1.0000
## NA's    :4
## FALSE, FALSE, TRUE, classif.randomForest
## Min.      :0.4610
## 1st Qu.:0.8957
## Median :0.9755
## Mean    :0.9252
## 3rd Qu.:0.9963
## Max.    :1.0000
## NA's    :12
## FALSE, FALSE, TRUE, classif.xgboost FALSE, TRUE, FALSE, classif.ksvm
## Min.      :0.4862      Min.      :0.4719
## 1st Qu.:0.8968      1st Qu.:0.7345
## Median :0.9668      Median :0.9261
## Mean    :0.9168      Mean    :0.8563
## 3rd Qu.:0.9946      3rd Qu.:0.9950
## Max.    :1.0000      Max.    :1.0000
## NA's    :3          NA's    :5
## FALSE, TRUE, FALSE, classif.randomForest
## Min.      :0.4652
## 1st Qu.:0.8964
## Median :0.9825
## Mean    :0.9224
## 3rd Qu.:0.9985
## Max.    :1.0000
## NA's    :11
## FALSE, TRUE, FALSE, classif.xgboost SMOTE, FALSE, FALSE, classif.ksvm
## Min.      :0.4886      Min.      :0.7440
## 1st Qu.:0.5000      1st Qu.:0.9994
## Median :0.5000      Median :1.0000
## Mean    :0.6680      Mean    :0.9949
## 3rd Qu.:0.9144      3rd Qu.:1.0000
## Max.    :1.0000      Max.    :1.0000
## NA's    :5
## SMOTE, FALSE, FALSE, classif.randomForest
## Min.      :0.7577
## 1st Qu.:0.9990
## Median :0.9999
## Mean    :0.9952
## 3rd Qu.:1.0000
## Max.    :1.0000

```

```
## NA's :26
## SMOTE, FALSE, FALSE, classif.xgboost
## Min. :0.6201
## 1st Qu.:0.9947
## Median :0.9994
## Mean :0.9802
## 3rd Qu.:1.0000
## Max. :1.0000
##
```

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 = 1679.7, df = 14, p-value < 2.2e-16
```

Testando as diferencas par a par

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

```
##      ADASYN, FALSE, FALSE, classif.ksvm
## [1,]                                     FALSE
## [2,]                                     FALSE
## [3,]                                     TRUE
## [4,]                                     TRUE
## [5,]                                     TRUE
## [6,]                                     TRUE
## [7,]                                     TRUE
## [8,]                                     TRUE
## [9,]                                     TRUE
## [10,]                                    TRUE
## [11,]                                    TRUE
## [12,]                                    TRUE
## [13,]                                    FALSE
## [14,]                                    FALSE
## [15,]                                    TRUE
##      ADASYN, FALSE, FALSE, classif.randomForest
## [1,]                                     FALSE
## [2,]                                     FALSE
## [3,]                                     FALSE
## [4,]                                     TRUE
## [5,]                                     TRUE
## [6,]                                     TRUE
## [7,]                                     TRUE
## [8,]                                     TRUE
## [9,]                                     TRUE
## [10,]                                    TRUE
## [11,]                                    TRUE
## [12,]                                    TRUE
## [13,]                                    TRUE
## [14,]                                    FALSE
## [15,]                                    FALSE
##      ADASYN, FALSE, FALSE, classif.xgboost
## [1,]                                     TRUE
## [2,]                                     FALSE
## [3,]                                     FALSE
## [4,]                                     TRUE
## [5,]                                     TRUE
## [6,]                                     TRUE
## [7,]                                     TRUE
## [8,]                                     TRUE
## [9,]                                     TRUE
## [10,]                                    TRUE
## [11,]                                    TRUE
## [12,]                                    TRUE
## [13,]                                    TRUE
## [14,]                                    FALSE
## [15,]                                    FALSE
```

```

##      FALSE, FALSE, FALSE, classif.ksvm
## [1,]      TRUE
## [2,]      TRUE
## [3,]      TRUE
## [4,]     FALSE
## [5,]      TRUE
## [6,]      TRUE
## [7,]     FALSE
## [8,]      TRUE
## [9,]     FALSE
## [10,]    FALSE
## [11,]     TRUE
## [12,]     TRUE
## [13,]     TRUE
## [14,]     TRUE
## [15,]     TRUE
##      FALSE, FALSE, FALSE, classif.randomForest
## [1,]      TRUE
## [2,]      TRUE
## [3,]      TRUE
## [4,]      TRUE
## [5,]     FALSE
## [6,]     FALSE
## [7,]      TRUE
## [8,]     FALSE
## [9,]      TRUE
## [10,]     TRUE
## [11,]    FALSE
## [12,]     TRUE
## [13,]     TRUE
## [14,]     TRUE
## [15,]     TRUE
##      FALSE, FALSE, FALSE, classif.xgboost
## [1,]      TRUE
## [2,]      TRUE
## [3,]      TRUE
## [4,]      TRUE
## [5,]     FALSE
## [6,]     FALSE
## [7,]      TRUE
## [8,]     FALSE
## [9,]      TRUE
## [10,]     TRUE
## [11,]    FALSE
## [12,]     TRUE
## [13,]     TRUE
## [14,]     TRUE
## [15,]     TRUE
##      FALSE, FALSE, TRUE, classif.ksvm
## [1,]      TRUE
## [2,]      TRUE
## [3,]      TRUE
## [4,]     FALSE
## [5,]      TRUE

```

```

## [6,] TRUE
## [7,] FALSE
## [8,] TRUE
## [9,] TRUE
## [10,] FALSE
## [11,] TRUE
## [12,] FALSE
## [13,] TRUE
## [14,] TRUE
## [15,] TRUE
## FALSE, FALSE, TRUE, classif.randomForest
## [1,] TRUE
## [2,] TRUE
## [3,] TRUE
## [4,] TRUE
## [5,] FALSE
## [6,] FALSE
## [7,] TRUE
## [8,] FALSE
## [9,] FALSE
## [10,] TRUE
## [11,] FALSE
## [12,] TRUE
## [13,] TRUE
## [14,] TRUE
## [15,] TRUE
## FALSE, FALSE, TRUE, classif.xgboost FALSE, TRUE, FALSE, classif.ksvm
## [1,] TRUE TRUE
## [2,] TRUE TRUE
## [3,] TRUE TRUE
## [4,] FALSE FALSE
## [5,] TRUE TRUE
## [6,] TRUE TRUE
## [7,] TRUE FALSE
## [8,] FALSE TRUE
## [9,] FALSE FALSE
## [10,] FALSE FALSE
## [11,] FALSE TRUE
## [12,] TRUE TRUE
## [13,] TRUE TRUE
## [14,] TRUE TRUE
## [15,] TRUE TRUE
## FALSE, TRUE, FALSE, classif.randomForest
## [1,] TRUE
## [2,] TRUE
## [3,] TRUE
## [4,] TRUE
## [5,] FALSE
## [6,] FALSE
## [7,] TRUE
## [8,] FALSE
## [9,] FALSE
## [10,] TRUE
## [11,] FALSE

```



```

## [12,] TRUE
## [13,] TRUE
## [14,] TRUE
## [15,] TRUE
## FALSE, TRUE, FALSE, classif.xgboost
## [1,] TRUE
## [2,] TRUE
## [3,] TRUE
## [4,] TRUE
## [5,] TRUE
## [6,] TRUE
## [7,] FALSE
## [8,] TRUE
## [9,] TRUE
## [10,] TRUE
## [11,] TRUE
## [12,] FALSE
## [13,] TRUE
## [14,] TRUE
## [15,] TRUE
## SMOTE, FALSE, FALSE, classif.ksvm
## [1,] FALSE
## [2,] TRUE
## [3,] TRUE
## [4,] TRUE
## [5,] TRUE
## [6,] TRUE
## [7,] TRUE
## [8,] TRUE
## [9,] TRUE
## [10,] TRUE
## [11,] TRUE
## [12,] TRUE
## [13,] FALSE
## [14,] TRUE
## [15,] TRUE
## SMOTE, FALSE, FALSE, classif.randomForest
## [1,] FALSE
## [2,] FALSE
## [3,] FALSE
## [4,] TRUE
## [5,] TRUE
## [6,] TRUE
## [7,] TRUE
## [8,] TRUE
## [9,] TRUE
## [10,] TRUE
## [11,] TRUE
## [12,] TRUE
## [13,] TRUE
## [14,] FALSE
## [15,] FALSE
## SMOTE, FALSE, FALSE, classif.xgboost
## [1,] TRUE

```

```
## [2,] FALSE
## [3,] FALSE
## [4,] TRUE
## [5,] TRUE
## [6,] TRUE
## [7,] TRUE
## [8,] TRUE
## [9,] TRUE
## [10,] TRUE
## [11,] TRUE
## [12,] TRUE
## [13,] TRUE
## [14,] FALSE
## [15,] FALSE
```

Plotando grafico de Critical Difference

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

