

R Notebook

Parametros:

Measure = Matthews correlation coefficient
Columns = sampling, weight_space, underbagging
Performance = holdout_measure
Filter keys = imba.rate
Filter values = 0.01

```
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)){
  ds = filter_at(ds, .vars = params$filter_keys, .vars_predicate = any_vars(. == params$filter_values))
}
```

```
summary(ds)
```

```
##           learner      weight_space
## classif.ksvm      :600 Mode :logical
## classif.randomForest:600 FALSE:1440
## classif.rusboost   : 0 TRUE :360
## classif.xgboost    :600 NA's :0
##
##
##
##           measure      sampling      underbagging
## Accuracy           : 0 ADASYN: 360 Mode :logical
## Area under the curve : 0 FALSE :1080 FALSE:1440
## F1 measure          : 0 SMOTE : 360 TRUE :360
## G-mean              : 0 NA's :0
## Matthews correlation coefficient:1800
##
##
## tuning_measure      holdout_measure      holdout_measure_residual
## Min. : -0.00646 Min. : -0.1370 Min. : -0.06817
## 1st Qu.: 0.23261 1st Qu.: 0.0000 1st Qu.: 0.02011
## Median : 0.82014 Median : 0.3764 Median : 0.19200
## Mean : 0.64070 Mean : 0.4285 Mean : 0.29498
## 3rd Qu.: 0.99730 3rd Qu.: 0.8152 3rd Qu.: 0.49996
## Max. : 1.00000 Max. : 1.0000 Max. : 1.00000
## NA's :69 NA's :69 NA's :69
## iteration_count      dataset      imba.rate
## Min. :1      abalone : 45 Min. :0.01
## 1st Qu.:1      adult : 45 1st Qu.:0.01
## Median :2      bank : 45 Median :0.01
## Mean :2      car : 45 Mean :0.01
## 3rd Qu.:3      cardiocography-10clases: 45 3rd Qu.:0.01
## Max. :3      cardiocography-3clases : 45 Max. :0.01
## NA's :69      (Other) :1530
```

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] 120    5

# Renomeando a variavel
df = df_tec_wide_residual

summary(df)
```

```
## ADASYN, FALSE, FALSE FALSE, FALSE, FALSE FALSE, FALSE, TRUE
## Min.   :-0.0145      Min.   :-0.005742  Min.   :-0.02588
## 1st Qu.: 0.0000      1st Qu.: 0.000000  1st Qu.: 0.14800
## Median : 0.3365      Median : 0.465389  Median : 0.27777
## Mean   : 0.4247      Mean    : 0.465610  Mean    : 0.35146
## 3rd Qu.: 0.8023      3rd Qu.: 0.872325  3rd Qu.: 0.50993
## Max.   : 1.0000      Max.    : 1.000000  Max.    : 1.00000
## NA's   :10          NA's     :3         NA's     :1
## FALSE, TRUE, FALSE SMOTE, FALSE, FALSE
## Min.   :-0.005031  Min.   :-0.01672
## 1st Qu.: 0.000000  1st Qu.: 0.00000
## Median : 0.467017  Median : 0.38977
## Mean   : 0.460559  Mean    : 0.44164
## 3rd Qu.: 0.858511  3rd Qu.: 0.83054
## Max.   : 1.000000  Max.    : 1.00000
## NA's   :4          NA's     :5
```

Verificando a média de cada coluna selecionada

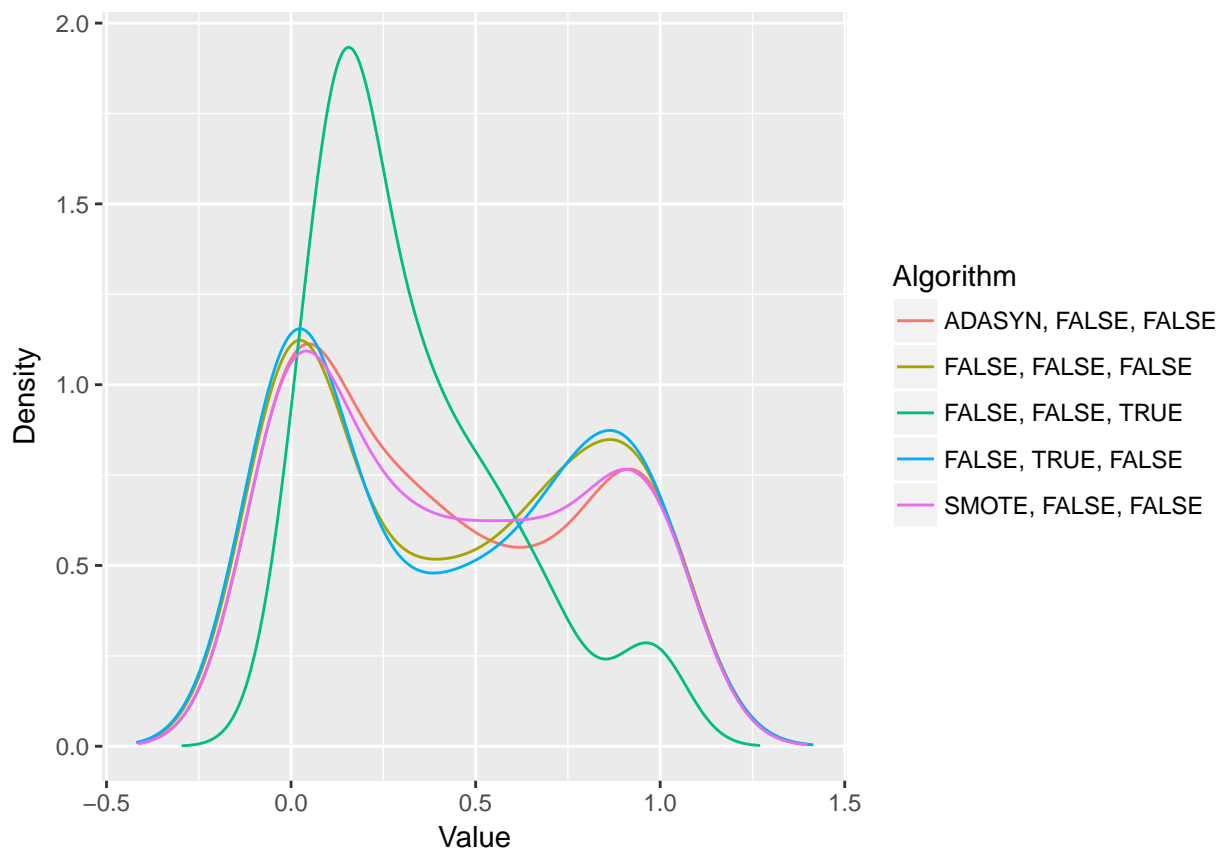
```
for(i in (1:dim(df)[2])){
  #print(df[,i])
}
```

```
print(paste("Media da coluna ", colnames(df)[i], " = ", mean(df[,i], na.rm = TRUE), sep=""))
}
```

```
## [1] "Media da coluna ADASYN, FALSE, FALSE = 0.424664743950473"
## [1] "Media da coluna FALSE, FALSE, FALSE = 0.465610453584806"
## [1] "Media da coluna FALSE, FALSE, TRUE = 0.351458120128611"
## [1] "Media da coluna FALSE, TRUE, FALSE = 0.460559366861479"
## [1] "Media da coluna SMOTE, FALSE, FALSE = 0.441642007742059"
```

Fazendo teste de normalidade

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



Testando as diferencas

```
friedmanTest(df)
```

```
##
## Friedman's rank sum test
##
## data: df
## Friedman's chi-squared = 2.2783, df = 4, p-value = 0.6847
```

Testando as diferenças par a par

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

```
##      ADASYN, FALSE, FALSE FALSE, FALSE, FALSE FALSE, FALSE, TRUE
## [1,]          FALSE          FALSE          FALSE
## [2,]          FALSE          FALSE          FALSE
## [3,]          FALSE          FALSE          FALSE
## [4,]          FALSE          FALSE          FALSE
## [5,]          FALSE          FALSE          FALSE
##      FALSE, TRUE, FALSE SMOTE, FALSE, FALSE
## [1,]          FALSE          FALSE
## [2,]          FALSE          FALSE
## [3,]          FALSE          FALSE
## [4,]          FALSE          FALSE
## [5,]          FALSE          FALSE
```

Plotando grafico de Critical Difference

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

