

R Notebook

Parametros:

Measure = G-mean

Columns = sampling, weight_space, underbagging

Performance = holdout_measure_residual

Filter keys = imba.rate

Filter values = 0.03

```
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      :990 Mode :logical
## classif.randomForest:990 FALSE:2376
## classif.rusboost   : 0 TRUE :594
## classif.xgboost    :990 NA's :0
##
##
##
##          measure      sampling      underbagging
## Accuracy           : 0 ADASYN: 594 Mode :logical
## Area under the curve : 0 FALSE :1782 FALSE:2376
## F1 measure          : 0 SMOTE : 594 TRUE :594
## G-mean              :2970 NA's :0
## Matthews correlation coefficient: 0
##
##
## tuning_measure  holdout_measure  holdout_measure_residual
## Min. :0.0000 Min. :0.0000 Min. :0.0000
## 1st Qu.:0.6338 1st Qu.:0.2132 1st Qu.:0.1828
## Median :0.9453 Median :0.7348 Median :0.4920
## Mean :0.7583 Mean :0.6032 Mean :0.4882
## 3rd Qu.:0.9933 3rd Qu.:0.9533 3rd Qu.:0.8073
## Max. :1.0000 Max. :1.0000 Max. :1.0000
## NA's :48 NA's :48 NA's :48
## iteration_count      dataset      imba.rate
## Min. :1 abalone : 45 Min. :0.03
## 1st Qu.:1 adult : 45 1st Qu.:0.03
## Median :2 annealing : 45 Median :0.03
## Mean :2 arrhythmia : 45 Mean :0.03
## 3rd Qu.:3 balance-scale: 45 3rd Qu.:0.03
## Max. :3 bank : 45 Max. :0.03
## NA's :48 (Other) :2700
```

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 columnas
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] 198 5
```

```
# Renomeando a variavel
df = df_tec_wide_residual

head(df)
```

```
## ADASYN, FALSE, FALSE FALSE, FALSE, FALSE FALSE, FALSE, TRUE
## 1 0.33453836 0.27960010 0.6303596
## 2 0.37818667 0.37894057 0.7581606
## 3 0.53318651 0.32769371 0.7253062
## 4 0.00000000 0.00000000 0.2877302
## 5 0.06267509 0.30230195 0.3524105
## 6 0.28934649 0.06265226 0.5829965
## FALSE, TRUE, FALSE SMOTE, FALSE, FALSE
## 1 0.25068766 0.3256847
## 2 0.38112845 0.3875660
## 3 0.32769371 0.4649288
## 4 0.00000000 0.0000000
## 5 0.30230195 0.2206087
## 6 0.06265226 0.1151854
```

```
summary(df)
```

```
## ADASYN, FALSE, FALSE FALSE, FALSE, FALSE FALSE, FALSE, TRUE
## Min. :0.0000 Min. :0.0000 Min. :0.09623
## 1st Qu.:0.1997 1st Qu.:0.1280 1st Qu.:0.47810
## Median :0.4121 Median :0.3333 Median :0.71620
## Mean :0.4718 Mean :0.4094 Mean :0.68741
## 3rd Qu.:0.7529 3rd Qu.:0.6961 3rd Qu.:0.91132
## Max. :1.0000 Max. :1.0000 Max. :0.99992
```

```
## NA's :9          NA's :1          NA's :1
## FALSE, TRUE, FALSE SMOTE, FALSE, FALSE
## Min. :0.0000    Min. :0.0000
## 1st Qu.:0.1229   1st Qu.:0.1995
## Median :0.3320   Median :0.4103
## Mean :0.4025     Mean :0.4686
## 3rd Qu.:0.6868   3rd Qu.:0.7698
## Max. :1.0000     Max. :1.0000
## NA's :2          NA's :3
```

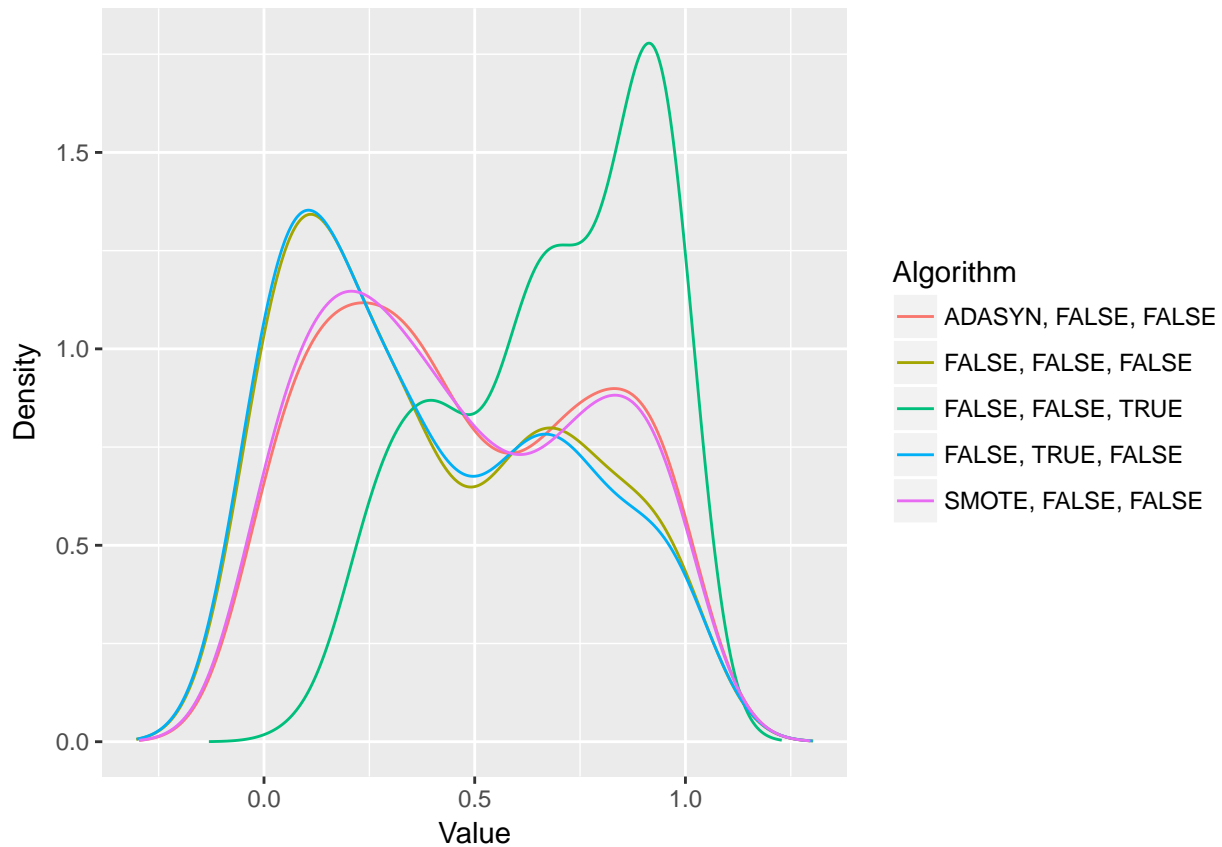
Verificando a média de cada coluna selecionada

```
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 ADASYN, FALSE, FALSE = 0.471761029066261"
## [1] "Media da coluna FALSE, FALSE, FALSE = 0.4094071402036"
## [1] "Media da coluna FALSE, FALSE, TRUE = 0.687413650906424"
## [1] "Media da coluna FALSE, TRUE, FALSE = 0.402470606643995"
## [1] "Media da coluna SMOTE, FALSE, FALSE = 0.46857544352466"
```

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 = 317.58, df = 4, 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 FALSE, FALSE, FALSE FALSE, FALSE, TRUE
## [1,]          FALSE          TRUE          TRUE
## [2,]          TRUE          FALSE          TRUE
## [3,]          TRUE          TRUE          FALSE
## [4,]          TRUE          FALSE          TRUE
## [5,]          FALSE          TRUE          TRUE
##      FALSE, TRUE, FALSE SMOTE, FALSE, FALSE
## [1,]          TRUE          FALSE
```

```
## [2,]          FALSE          TRUE
## [3,]          TRUE          TRUE
## [4,]          FALSE          TRUE
## [5,]          TRUE          FALSE
```

Plotando os ranks

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

```
## ADASYN, FALSE, FALSE  FALSE, FALSE, FALSE  FALSE, FALSE, TRUE
##           3.063131           3.707071           1.363636
##  FALSE, TRUE, FALSE  SMOTE, FALSE, FALSE
##           3.909091           2.957071
```

Plotando grafico de Critical Difference

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

