Dissimilarités entre jeux de données
-
Listings

15 février 2017

1 Méta-attributs généraux des jeux de données

TABLE 1 – Méta-attributs simples des jeux de données

Méta-attribut	Description	$\delta^{\emptyset}(x)$
DefaultAccuracy	The predictive accuracy obtained by predicting the majority class.	0
Dimensionality	Number of attributes divided by the number of instances.	0
MajorityClassPercentage	Percentage of rows with the class with the most assigned index.	$\|x-1\|$
MajorityClassSize	The number of instances that have the majority class.	0
MinorityClassPercentage	Percentage of rows with the class with the least assigned index.	$\|x-1\|$
MinorityClassSize	The number of instances that have the minority class.	0
NumberOfBinaryFeatures	Count of binary attributes.	0
NumberOfClasses	The number of classes in the class attribute.	0
NumberOfFeatures	Number of attributes (columns) of the dataset.	0
NumberOfInstances	Number of instances (rows) of the dataset.	0
NumberOfInstancesWithMissingValues	Number of instances with at least one value missing.	0
NumberOfMissingValues	Number of missing values in the dataset.	0
NumberOfNumericFeatures	Count of categorical attributes.	0
NumberOfSymbolicFeatures	Count of nominal attributes.	0
PercentageOfBinaryFeatures	Percentage of binary attributes.	0
PercentageOfInstancesWithMissingValues	Percentage of instances with missing values.	$\|x-1\|$
PercentageOfMissingValues	Percentage of missing values.	$\|x-1\|$
PercentageOfNumericFeatures	Percentage of numerical attributes.	0
PercentageOfSymbolicFeatures	Percentage of nominal attributes.	0

Table 2 – Méta-attributs généraux décrivant les attributs numériques

Méta-attribut	Description	$\delta^{\emptyset}(x)$
MeanMeansOfNumericAtts	Mean of means among numeric attributes.	0
MeanStdDevOfNumericAtts	Mean standard deviation of numeric attributes.	$\|x\|$
MeanKurtosisOfNumericAtts	Mean kurtosis among numeric attributes.	$\|x+1,2\|$
MeanSkewnessOfNumericAtts	Mean skewness among numeric attributes.	$\|x\|$
MinMeansOfNumericAtts	Min of means among numeric attributes.	0
MinStdDevOfNumericAtts	Min standard deviation of numeric attributes.	$\|x\|$
MinKurtosisOfNumericAtts	Min kurtosis among numeric attributes.	$\|x+1,2\|$
MinSkewnessOfNumericAtts	Min skewness among numeric attributes.	$\|x\|$
MaxMeansOfNumericAtts	Max of means among numeric attributes.	0
MaxStdDevOfNumericAtts	Max standard deviation of numeric attributes.	$\|x\|$
MaxKurtosisOfNumericAtts	Max kurtosis among numeric attributes.	$\|x+1,2\|$
MaxSkewnessOfNumericAtts	Max skewness among numeric attributes.	$\|x\|$
Quartile1MeansOfNumericAtts	First quartile of means among numeric attributes.	0
Quartile1StdDevOfNumericAtts	First quartile of standard deviation of numeric attributes.	$\|x\|$
Quartile1KurtosisOfNumericAtts	First quartile of kurtosis among numeric attributes.	$\|x+1,2\|$
Quartile1SkewnessOfNumericAtts	First quartile of skewness among numeric attributes.	$\|x\|$
Quartile2MeansOfNumericAtts	Second quartile of means among numeric attributes.	0
Quartile2StdDevOfNumericAtts	Second quartile of standard deviation of numeric attributes.	$\|x\|$
Quartile2KurtosisOfNumericAtts	Second quartile of kurtosis among numeric attributes.	$\|x+1,2\|$
Quartile2SkewnessOfNumericAtts	Second quartile of skewness among numeric attributes.	$\|x\|$
Quartile3MeansOfNumericAtts	Third quartile of means among numeric attributes.	0
Quartile3StdDevOfNumericAtts	Third quartile of standard deviation of numeric attributes.	$\|x\|$
Quartile3KurtosisOfNumericAtts	Third quartile of kurtosis among numeric attributes.	$\|x+1,2\|$
Quartile3SkewnessOfNumericAtts	Third quartile of skewness among numeric attributes.	$\|x\|$

Table 3 – Méta-attributs généraux décrivant les attributs nominaux

Méta-attribut	Description	$\delta^{\emptyset}(x)$
ClassEntropy	Entropy of the target attribute.	0
EquivalentNumberOfAtts	An estimate of the amount of useful attributes.	0
NoiseToSignalRatio	An estimate of the amount of non-useful information in the attributes regarding the class.	0
MeanAttributeEntropy	Mean of entropy among attributes.	$\lvert x\rvert$
MeanMutualInformation	Mean of mutual information between the nominal attributes and the target attribute.	$\lvert x\rvert$
MinAttributeEntropy	Min of entropy among attributes.	$\lvert x\rvert$
MinMutualInformation	Min of mutual information between the nominal attributes and the target attribute.	$\lvert x\rvert$
MaxAttributeEntropy	Max of entropy among attributes.	$\lvert x\rvert$
MaxMutualInformation	Max of mutual information between the nominal attributes and the target attribute.	$\lvert x\rvert$
Quartile1AttributeEntropy	First quartile of entropy among attributes.	$\lvert x\rvert$
Quartile1MutualInformation	First quartile of mutual information between the nominal attributes and the target attribute.	$\lvert x\rvert$
Quartile2AttributeEntropy	Second quartile of entropy among attributes.	$\lvert x\rvert$
Quartile2MutualInformation	Second quartile of mutual information between the nominal attributes and the target attribute.	$\lvert x\rvert$
Quartile3AttributeEntropy	Third quartile of entropy among attributes.	$\lvert x\rvert$
Quartile3MutualInformation	Third quartile of mutual information between the nominal attributes and the target attribute.	$\lvert x\rvert$
MaxNominalAttDistinctValues	The maximum number of distinct values among attributes of the nominal type.	$\lvert x\rvert$
MinNominalAttDistinctValues	The minimal number of distinct values among attributes of the nominal type.	$\lvert x\rvert$
MeanNominalAttDistinctValues	Mean of number of distinct values among the attributes of the nominal type.	$\lvert x\rvert$
StdvNominalAttDistinctValues	Standard deviation of the number of distinct values among nominal attributes.	$\lvert x\rvert$

Table 4 – "Aire sous la courbe" des landmarkers

Méta-attribut	Description	$\delta^{\emptyset}(x)$
DecisionStumpAUC	Area Under ROC achieved by the landmarker weka.classifiers.trees.DecisionStump	$\lvert x-0,5\rvert$
J48AUC	Area Under ROC achieved by the landmarker weka.classifiers.trees.J48	$\lvert x-0,5\rvert$
JRipAUC	Area Under ROC achieved by the landmarker weka.classifiers.rules.Jrip	$\lvert x-0,5\rvert$
kNN_3NAUC	Area Under ROC achieved by the landmarker weka.classifiers.lazy.IBk -K 3	$\lvert x-0,5\rvert$
NaiveBayesAUC	Area Under ROC achieved by the landmarker weka.classifiers.bayes.NaiveBayes	$\lvert x-0,5\rvert$
NBTreeAUC	Area Under ROC achieved by the landmarker weka.classifiers.trees.NBTree	$\lvert x-0,5\rvert$
RandomTreeDepth3AUC	Area Under ROC achieved by the landmarker weka.classifiers.trees.RandomTree -depth 3	$\lvert x-0,5\rvert$
REPTreeDepth3AUC	Area Under ROC achieved by the landmarker weka.classifiers.trees.REPTree -L 3	$\lvert x-0,5\rvert$
SimpleLogisticAUC	Area Under ROC achieved by the landmarker weka.classifiers.functions.SimpleLogistic	$\lvert x-0,5\rvert$

Table 5 – Taux d'erreur des landmarkers

Méta-attribut	Description	$\delta^{\emptyset}(x)$
DecisionStumpErrRate	Error rate achieved by the landmarker weka.classifiers.trees.DecisionStump	$\|x-1\|$
J48ErrRate	Error rate achieved by the landmarker weka.classifiers.trees.J48	$\|x-1\|$
JRipErrRate	Error rate achieved by the landmarker weka.classifiers.rules.Jrip	$\|x-1\|$
kNN_3NErrRate	Error rate achieved by the landmarker weka.classifiers.lazy.IBk -K 3	$\|x-1\|$
NaiveBayesErrRate	Error rate achieved by the landmarker weka.classifiers.bayes.NaiveBayes	$\|x-1\|$
NBTreeErrRate	Error rate achieved by the landmarker weka.classifiers.trees.NBTree	$\|x-1\|$
RandomTreeDepth3ErrRate	Error rate achieved by the landmarker weka.classifiers.trees.RandomTree -depth 3	$\|x-1\|$
REPTreeDepth3ErrRate	Error rate achieved by the landmarker weka.classifiers.trees.REPTree -L 3	$\|x-1\|$
SimpleLogisticErrRate	Error rate achieved by the landmarker weka.classifiers.functions.SimpleLogistic	$\|x-1\|$

Table 6 – Kappa de Cohen des landmarkers

Méta-attribut	Description	$\delta^{\emptyset}(x)$
DecisionStumpKappa	Kappa coefficient achieved by the landmarker weka.classifiers.trees.DecisionStump	$\|x\|$
J48Kappa	Kappa coefficient achieved by the landmarker weka.classifiers.trees.J48	$\|x\|$
JRipKappa	Kappa coefficient achieved by the landmarker weka.classifiers.rules.Jrip	$\|x\|$
kNN_3NKappa	Kappa coefficient achieved by the landmarker weka.classifiers.lazy.IBk -K 3	$\|x\|$
NaiveBayesKappa	Kappa coefficient achieved by the landmarker weka.classifiers.bayes.NaiveBayes	$\|x\|$
NBTreeKappa	Kappa coefficient achieved by the landmarker weka.classifiers.trees.NBTree	$\|x\|$
RandomTreeDepth3Kappa	Kappa coefficient achieved by the landmarker weka.classifiers.trees.RandomTree -depth 3	$\|x\|$
REPTreeDepth3Kappa	Kappa coefficient achieved by the landmarker weka.classifiers.trees.REPTree -L 3	$\|x\|$
SimpleLogisticKappa	Kappa coefficient achieved by the landmarker weka.classifiers.functions.SimpleLogistic	$\|x\|$

2 Méta-attributs des attributs

Table 7 – Méta-attributs simples des attributs

Méta-attribut	Description	$\delta^{\emptyset}(x)$
ValuesCount	Number of values.	$\|x-1\|$
NonMissingValuesCount	Number of non missing values.	$\|x\|$
MissingValuesCount	Number of missing values.	0
Distinct	Number of distinct values.	$\|x-1\|$
AverageClassCount	Average count of occurrences among different classes.	0
Entropy	Entropy of the values.	$\|x-1\|$
MostFequentClassCount	Count of the most probable class.	0
LeastFequentClassCount	Count of the least probable class.	0
ModeClassCount	Mode of the number of distinct values.	0
MedianClassCount	Median of the number of distinct values.	0
PearsonCorrellationCoefficient	Pearson Correlation Coeeficient of the values with the target attribute.	$\|x\|$
SpearmanCorrelationCoefficient	Spearman Correlation Coeeficient of the values with the target attribute.	$\|x\|$
CovarianceWithTarget	Covariance of the values with the target attribute.	$\|x\|$

Table 8 – Méta-attributs spécifiques aux attributs nominaux

Méta-attribut	Description	$\delta^{\emptyset}(x)$
UniformDiscrete	Result of Pearson's chi-squared test for discrete uniform distribution.	$\|x-1\|$
ChiSquareUniformDistribution	Statistic value for the Pearson's chi-squared test.	$\|x\|$
RationOfDistinguishing CategoriesByKolmogorov SmirnoffSlashChiSquare	Ratio of attribute values that after sub-setting the dataset to that attribute value leads to different distribution of the target as indicated by the Kolmogorov-Smirnoff test.	0
RationOfDistinguishing CategoriesByUtest	Ratio of attribute values that after sub-setting the dataset to that attribute value leads to different distribution of the target as indicated by the Mann-Whitney U-test.	0

Table 9 – Méta-attributs spécifiques aux attributs numériques

Méta-attribut	Description	$\delta^{\emptyset}(x)$
IsUniform	Whether statistical test did or not reject that the attribute values corresponds to a uniform distribution.	$\|x-1\|$
IntegersOnly	Whether attribute values are only integers.	0
Min	Minimal value of the attribute values.	0
Max	Maximal value of the attribute values.	0
Kurtosis	Kurtosis of the values.	$\|x+1,2\|$
Mean	Mean of the values.	0
Skewness	Skewness of the values.	$\|x\|$
StandardDeviation	Standard deviation of the values.	$\|x\|$
Variance	Variance of the values.	$\|x\|$
Mode	Mode of the values.	0
Median	Median of the values.	0
ValueRange	Difference between maximum and minimum of the values.	$\|x\|$
LowerOuterFence	Lower outer fence of the values.	0
HigherOuterFence	Higher outer fence of the values.	0
LowerQuartile	Lower quartile of the values.	0
HigherQuartile	Higher quartile of the values.	0
HigherConfidence	Higher confidence interval of the values.	0
LowerConfidence	Lower confidence interval of the values.	0
PositiveCount	Number of positive values.	$\|x\|$
NegativeCount	Number of negative values.	$\|x\|$

Table 10 – Méta-attributs normalisés selon le nombre d'instances

Méta-attribut	Description	$\delta^{\emptyset}(x)$
MissingValues	1 if count of missing values is greater than 0, 0 otherwise.	$\|x-1\|$
AveragePercentageOfClass	Percentage of the occurrences among classes.	$\|x-1\|$
PercentageOfMissing	Percentage of missing values in the attribute.	$\|x-1\|$
PercentageOfNonMissing	Percentage of non missing values in the attribute.	$\|x\|$
PercentageOfMostFrequentClass	Percentage of the most frequent class.	$\|x-1\|$
PercentageOfLeastFrequentClass	Percentage of the least frequent class.	$\|x-1\|$
ModeClassPercentage	Percentage of mode of class count calculated as ModeFrequentClassCount / ValuesCount.	$\|x-1\|$
MedianClassPercentage	Percentage of median of class count calculated as MedianFrequentClassCount / ValuesCount.	$\|x-1\|$

Table 11 – Méta-attributs normalisés spécifiques aux attributs numériques

Méta-attribut	Description	$\delta^{\emptyset}(x)$
PositivePercentage	Percentage of positive values.	$\|x\|$
NegativePercentage	Percentage of negative values.	$\|x\|$
HasPositiveValues	1 if attribute has at least one positive value, 0 otherwise.	$\|x\|$
HasNegativeValues	1 if attribute has at least one negative value, 0 otherwise.	$\|x\|$

3 Algorithmes de la *baseline*

Table 12 – Algorithmes d'apprentissage traditionnels de la *baseline*

Implémentation Weka	Description
GaussianProcesses	Gaussian Processes for regression. See [**?**].
LinearRegression	Linear regression for prediction. Uses the Akaike criterion for model selection, and is able to deal with weighted instances. See [**?**].
RBFRegressor	Radial basis function networks, trained in a fully supervised manner by minimizing squared error with the BFGS method. See [**?**].
SMOreg	Sequential minimal optimization algorithm for training a support vector regression model. See [**?**].
RandomForest	Ensemble of decision trees outputting the mean prediction of the individual trees. See [**?**].
KStar	Instance-based classifier where the class of a test instance is based upon the class of those training instances similar to it, as determined by an entropy-based distance function. See [**?**].
M5Rules	Generates a decision list for regression problems using separate-and-conquer. Builds a model tree using M5 In each iteration and makes the best leaf into a rule. See [**?**].

4 Meta-Dataset

Table 13 – Chaines de traitement weka employés comme classifieurs dans les expériences comparatives

A1DE
AdaBoostM1_DecisionStump
Bagging_REPTree
BayesNet_K2
BFTree
ConjunctiveRule
Dagging_DecisionStump
DecisionStump
DecisionTable_BestFirst
END_ND_J48
FT
FURIA
Grading_ZeroR
HoeffdingTree
HyperPipes
IB1
IBk
J48
J48graft
JRip
KStar
LADTree
LibLINEAR
LibSVM
LMT
Logistic
LogitBoost_DecisionStump
LWL_DecisionStump
MultiBoostAB_DecisionStump
NaiveBayes
NBTree
OLM
OneR
PART
RacedIncrementalLogitBoost_DecisionStump
RandomForest
RandomSubSpace_REPTree
RandomTree
RBFClassifier
REPTree
Ridor
RotationForest_PrincipalComponents_J48
SimpleCart
SimpleLogistic
SMO_PolyKernel
SMO_RBFKernel
VFI
ZeroR

TABLE 14 – Jeux de données OpenML utilisés dans les expériences comparatives

anneal
anneal
kr-vs-kp
labor
audiology
autos
lymph
balance-scale
breast-cancer
mfeat-fourier
breast-w
mfeat-karhunen
mfeat-morphological
car
mfeat-zernike
cmc
mushroom
nursery
colic
optdigits
credit-a
page-blocks
credit-g
pendigits
postoperative-patient-data
dermatology
segment
diabetes
ecoli
sonar
glass
soybean
haberman
spambase
tae
heart-c
tic-tac-toe
heart-h
heart-statlog
vehicle
hepatitis
vote
ionosphere
waveform-5000
iris
BNG(cmc,nominal,55296)
electricity
primary-tumor
solar-flare
solar-flare
adult
yeast
satimage
abalone
braziltourism
eucalyptus
BNG(breast-w)
meta_all
meta_batchincremental
meta_ensembles
meta_instanceincremental
lung-cancer
wine
hypothyroid
shuttle-landing-control
Australian
sick
monks-problems-1
monks-problems-2
monks-problems-3
SPECT
SPECTF
grub-damage
pasture
squash-stored
squash-unstored
white-clover
aids
JapaneseVowels
ipums_la_97-small
analcatdata_boxing2
prnn_crabs
analcatdata_boxing1
analcatdata_lawsuit
irish
analcatdata_broadwaymult
analcatdata_authorship
analcatdata_asbestos
analcatdata_creditscore
prnn_synth
analcatdata_cyyoung8092
schizo
confidence
analcatdata_dmft
profb
lupus
analcatdata_germangss
prnn_viruses
biomed

rmftsa_sleepdata
sleuth_ex2016
sleuth_ex2015
visualizing_livestock
diggle_table_a2
fruitfly
fri_c3_1000_25
fri_c3_100_50
rmftsa_ladata
veteran
abalone
pwLinear
pol
fri_c4_1000_25
analcatdata_vineyard
bank8FM
fri_c2_100_5
2dplanes
analcatdata_supreme
visualizing_slope
fri_c1_250_5
baskball
fri_c0_250_50
machine_cpu
ailerons
cpu_small
visualizing_environmental
space_ga
sleep
fri_c3_1000_10
rmftsa_sleepdata
fri_c1_1000_5
fri_c3_250_5
auto_price
fri_c1_250_25
servo
analcatdata_wildcat
fri_c3_500_5
pm10
fri_c4_1000_10
puma32H
wisconsin
fri_c0_100_5
sleuth_ex1605
autoPrice
meta
cpu_act
fri_c2_100_10
fri_c0_250_10
analcatdata_apnea3
analcatdata_apnea2
fri_c1_500_50
analcatdata_apnea1
fri_c3_100_25
fri_c1_250_50
strikes
quake
fri_c0_250_25
disclosure_x_bias
fri_c2_100_25
fri_c0_250_5
sleuth_ex1714
bodyfat
fri_c1_500_25
rabe_265
rabe_266
fri_c3_100_10
newton_hema
wind_correlations
cleveland
triazines
fri_c1_100_10
elusage
diabetes_numeric
fri_c2_500_5
fri_c3_250_10
fri_c2_250_25
disclosure_x_tampered
cpu
fri_c4_1000_50
cholesterol
fri_c0_1000_5
pyrim
pbcseq
delta_ailerons
hutsof99_logis
fri_c4_500_50
fri_c3_1000_50
kin8nm
fri_c0_100_10
mushroom
pbc
rmftsa_ctoarrivals
fri_c1_100_25
fri_c3_1000_5
chscase_vine2
chscase_vine1
puma8NH
diggle_table_a1

diggle_table_a2
delta_elevators
chatfield_4
house_16H
cal_housing
houses
fri_c1_500_10
boston_corrected
sensory
disclosure_x_noise
fri_c1_100_5
fri_c2_250_10
autoMpg
fri_c3_250_25
bank32nh
analcatdata_vehicle
sleuth_case1102
fri_c1_1000_50
fri_c4_500_25
kdd_el_nino-small
autoHorse
stock
analcatdata_runshoes
house_8L
breastTumor
fri_c0_1000_10
elevators
wind
schlvote
fri_c0_1000_25
fri_c0_100_50
analcatdata_gsssexsurvey
housing
fishcatch
fri_c4_500_10
bolts
hungarian
vinnie
auto93
sleuth_ex2016
fri_c4_250_10
sleuth_ex2015
fri_c2_1000_50
visualizing_livestock
fri_c4_100_25
fri_c2_500_10
fri_c1_500_5
pollen
boston
fri_c3_250_50
rabe_131
analcatdata_chlamydia
fri_c1_100_50
fri_c2_250_50
fri_c4_100_10
fri_c2_500_25
mu284
mv
pollution
fri_c0_500_5
transplant
no2
mbagrade
fri_c0_500_50
fri_c0_100_25
cloud
sleuth_case1202
visualizing_hamster
rabe_148
chscase_geyser1
fri_c3_500_25
hip
analcatdata_negotiation
chscase_census6
fried
sleuth_case2002
fri_c2_1000_25
fri_c0_1000_50
chscase_census5
chscase_census4
chscase_census3
chscase_census2
fri_c1_1000_10
fri_c2_250_5
fri_c2_1000_5
fri_c2_1000_10
plasma_retinol
fri_c3_100_5
fri_c1_1000_25
fri_c4_250_50
fri_c2_500_50
analcatdata_seropositive
fri_c2_100_50
visualizing_soil
visualizing_galaxy
fri_c0_500_25
disclosure_z
fri_c4_100_50
fri_c4_250_25

socmob
fri_c1_250_10
fri_c3_500_10
fri_c3_500_50
lowbwt
fri_c0_500_10
echoMonths
kidney
visualizing_ethanol
arsenic-male-bladder
quake
arsenic-female-bladder
arsenic-female-lung
arsenic-male-lung
tae
braziltourism
segment
nursery
postoperative-patient-data
analcatdata_broadwaymult
mfeat-morphological
heart-h
pasture
cars
analcatdata_birthday
iris
analcatdata_authorship
mfeat-fourier
squash-stored
wine
hayes-roth
autos
kdd_JapaneseVowels
letter
waveform-5000
optdigits
kdd_internet_usage
heart-c
cmc
squash-unstored
analcatdata_marketing
fl2000
anneal
eucalyptus
car
kdd_ipums_la_97-small
vehicle
mfeat-zernike
prnn_fglass
balance-scale
audiology
hypothyroid
kdd_ipums_la_98-small
primary-tumor
glass
lymph
white-clover
dermatology
ecoli
analcatdata_challenger
analcatdata_dmft
confidence
kdd_ipums_la_99-small
pendigits
mfeat-karhunen
page-blocks
soybean
analcatdata_germangss
grub-damage
ada_prior
ada_agnostic
eye_movements
kc1-top5
mozilla4
jEdit_4.2_4.3
pc4
pc3
jm1
mc2
cm1_req
mc1
ar1
ar3
ar4
ar5
kc2
ar6
kc3
kc1-binary
kc1
pc1
pc2
mw1
jEdit_4.0_4.2
desharnais
datatrieve
teachingAssistant
pc1_req