[Supplementary Document] Configuration Space Reduction in Automated Machine Learning Using Relative Landmarking

Tien-Dung Nguyen 1, Bogdan Gabrys 2, and Katarzyna Musial^2

This document provides more details about the experiments for the paper entitled Configuration Space Reduction in Automated Machine Learning Using Relative Landmarking.

Table 1. The best ML pipelines found using different methods to design configuration spaces.

Dataset	r30	baseline	avatar	
abalone	$ \begin{array}{l} {\rm CustomReplaceMissingValues} \\ \rightarrow {\rm RandomSubset} \\ \rightarrow {\rm Resample} \\ \rightarrow {\rm Logistic} \\ \rightarrow {\rm Bagging} \end{array} $	SimpleLogistic	SMO	
adult	J48	PART	PART	
amazon	$ \begin{array}{l} CustomReplaceMissingValues \\ \rightarrow Normalize \\ \rightarrow RandomSubset \\ \rightarrow NaiveBayesMultinomial \\ \rightarrow RandomSubSpace \end{array} $	malize domSubset VeBayesMultinomial NaiveBayes		
car	$SMO \rightarrow MultiClassClassifier$	SMO	SMO	
cifar10small	$ \begin{array}{l} {\rm RandomForest} \\ {\rm \rightarrow \ MultiClassClassifier} \end{array} $	DecisionStump	RandomForest	
convex	$RandomForest \rightarrow AdaBoostM1$	VotedPerceptron	RandomForest	
dexter	$\begin{array}{c} {\rm DecisionStump} \\ \rightarrow {\rm AdaBoostM1} \end{array}$	NaiveBayesMultinomial	SGD	
dorothea	$\begin{array}{l} \text{OneR} \\ \rightarrow \text{RandomSubSpace} \end{array}$	PART	DecisionStump	
gcredit	$LMT \rightarrow Bagging$	SMO	SMO	
gisette	$ \begin{array}{l} VotedPerceptron \\ \rightarrow RandomSubSpace \end{array} $	VotedPerceptron	VotedPerceptron	
kddcup	-	DecisionStump	$ \begin{array}{l} {\rm ClassBalancer} \\ \rightarrow {\rm RemoveOutliers} \\ \rightarrow {\rm InterquartileRange} \\ \rightarrow {\rm AttributeSelection} \\ \rightarrow {\rm Resample} \rightarrow {\rm PART} \end{array} $	
krvskp	$Jrip \rightarrow AdaBoostM1$	RandomForest	J48	
madelon	$ \begin{array}{l} Principal Components \\ \rightarrow IBk \\ \rightarrow Logit Boost \end{array} $	Jrip	Jrip	
mnist	$ \begin{array}{l} CustomReplaceMissingValues \\ \rightarrow Center \rightarrow J48 \rightarrow AdaBoostM1 \end{array} $	-	PART	
secom	$\begin{array}{l} \rm J48 \\ \rightarrow \rm AdaBoostM1 \end{array}$	$ \begin{array}{l} {\rm ClassBalancer} \\ {\rm \rightarrow \ EMImputation \rightarrow \ Normalize} \\ {\rm \rightarrow \ PrincipalComponents} \\ {\rm \rightarrow \ Kstar \rightarrow \ MultiClassClassifier} \end{array} $	SimpleLogistic	
semeion	$ \begin{array}{l} {\rm CustomReplaceMissingValues} \\ {\rm \rightarrow \ PrincipalComponents \ \rightarrow \ SMO} \end{array} $	SMO	SMO	
shuttle	$RandomForest \rightarrow AdaBoostM1$	RandomForest	RandomForest	
waveform	$ \begin{array}{l} {\rm RemoveOutliers} \\ \rightarrow {\rm InterquartileRange} \rightarrow {\rm Normalize} \\ \rightarrow {\rm SMO} \rightarrow {\rm AttributeSelectedClassifier} \end{array} $	LMT	SimpleLogistic	
winequality	$RandomForest \rightarrow AdaBoostM1$	RandomForest	Kstar	
yeast	$RandomForest \rightarrow Bagging$	RandomForest	RandomForest	

Table 2. The best ML pipelines found using different methods to design configuration spaces (continued).

Dataset	L-k1	L-k4	L-k8	L-k10	L-k19
abalone		$ \begin{array}{l} \operatorname{ClassBalancer} \\ \to \operatorname{RemoveOutliers} \\ \to \operatorname{InterquartileRange} \\ \to \operatorname{Normalize} \\ \to \operatorname{RandomSubset} \\ \to \operatorname{SimpleLogistic} \end{array} $	REPTree	RandomForest	PART
adult	$ \begin{array}{l} {\rm RemoveOutliers} \\ {\rm \rightarrow \ InterquartileRange} \\ {\rm \rightarrow \ Center \ \rightarrow \ Logistic} \end{array} $	Logistic	NaiveBayes	PART	PART
amazon	$ \begin{array}{l} {\rm ClassBalancer} \\ \rightarrow {\rm RemoveOutliers} \\ \rightarrow {\rm InterquartileRange} \\ \rightarrow {\rm Center} \\ \rightarrow {\rm Logistic} \end{array} $	SMO	NaiveBayesMultinomial	NaiveBayesMultinomial	NaiveBayesMultinomial
car	SMO	SMO	SMO	SMO	SMO
cifar10small	$ \begin{array}{l} {\rm RandomSubset} \\ {\rm \rightarrow \ PeriodicSampling \ \rightarrow \ Logistic} \end{array} $	Naive Bayes Multinomial		NaiveBayes	-
convex	-	Jrip	SMO	RandomForest	RandomForest
dexter	SMO	SMO	SMO	SMO	SimpleLogistic
dorothea	-	-	-	DecisionStump	DecisionStump
gcredit	LMT	NaiveBayes	NaiveBayes	SMO	SMO
gisette	Logistic	-	Jrip	RandomForest	Logistic
kddcup	-	-	-	-	-
krvskp	SMO	SMO	J48	J48	RandomForest
madelon	$ \begin{aligned} & SpreadSubsample \\ & \rightarrow CustomReplaceMissingValues \\ & \rightarrow Normalize \rightarrow RandomSubset \\ & \rightarrow Resample \rightarrow VotedPerceptron \end{aligned} $	Jrip	Jrip	Jrip	Jrip
mnist	-	-	-	SMO	SimpleLogistic
secom	-	ZeroR	ZeroR	DecisionTable	LMT
semeion	SMO	SMO	SMO	SMO	SMO
shuttle	SMO	Jrip	PART	RandomForest	RandomForest
waveform	SMO	SMO	SimpleLogistic	SimpleLogistic	LMT
winequality	SMO	Kstar	SMO	RandomForest	Kstar
yeast	Logistic	SMO	RandomForest	RandomForest	RandomForest

Table 3. The best ML pipelines found using different methods to design configuration spaces (continued).

Dataset	O-k1	O-k4	O-k8	O-k10	O-k19
abalone	Logistic	DecisionTable	SimpleLogistic	PART	MultilayerPerceptron
adult	-	PART	J48	J48	PART
amazon	-	NaiveBayesMultinomial	NaiveBayesMultinomial	NaiveBayesMultinomial	NaiveBayesMultinomial
car	LMT	SMO	SMO	SMO	SMO
cifar10small	-	RandomForest	RandomForest	NaiveBayes	RandomForest
convex	-	RandomForest	SMO	RandomForest	RandomForest
dexter	SGD	SGD	VotedPerceptron	SGD	SimpleLogistic
dorothea	$ \begin{aligned} & SpreadSubsample \\ & \rightarrow CustomReplaceMissingValues \\ & \rightarrow Center \rightarrow DecisionStump \end{aligned} $	NaiveBayes	v		OneR
gcredit	NaiveBayes	MultilayerPerceptron		MultilayerPerceptron	MultilayerPerceptron
gisette	VotedPerceptron	VotedPerceptron	VotedPerceptron	VotedPerceptron	VotedPerceptron
kddcup	MultilayerPerceptron	Ibk	DecisionStump	-	DecisionStump
krvskp	LMT	J48	J48	J48	J48
madelon		Jrip	Jrip	Jrip	Jrip
mnist	SMO	-	SMO	RandomForest	SMO
secom	$ \begin{array}{l} {\rm ClassBalancer} \\ \rightarrow {\rm CustomReplaceMissingValues} \\ \rightarrow {\rm RemoveOutliers} \rightarrow {\rm InterquartileRange} \\ \rightarrow {\rm Normalize} \rightarrow {\rm PeriodicSampling} \\ \rightarrow {\rm DecisionStump} \end{array} $	\rightarrow PeriodicSampling \rightarrow ZeroR	Kstar	m VotedPerceptron	Kstar
semeion	Logistic	SMO	-	SMO	SMO
shuttle	RandomForest	RandomForest		RandomForest	RandomForest
waveform	Logistic	SimpleLogistic		SimpleLogistic	SimpleLogistic
winequality		RandomForest			Kstar
yeast	RandomForest	RandomForest	RandomForest	RandomForest	RandomForest

To facilitate the study, we use prior evaluations of AutoWeka4MCPS with 2 hours optimisation time, 1GB memory, using SMAC as the ML pipeline composition and optimisation method over 20 datasets. We extracts the mean error rate of predictors based on the error rate of their ML pipelines within 20 datasets in the prior evaluations. Based on the mean error rate of the predictors, we generate the ranking of predictors across 20 datasets, as shown in Figure 1. We can see that there is no predictor has the best ranking across all datasets. For example, the component Logistic has the top ranking in the cases of the datasets abalone, semeion and waveform. However, this component ranks 19^{th} and 26^{th} in the case of the datasets convex and secom.

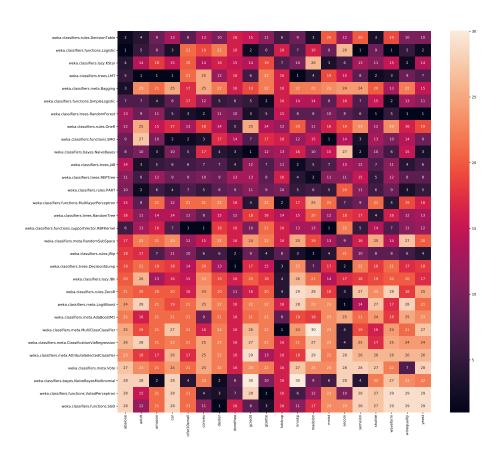


Fig. 1. Ranking of predictors based on mean error rate of their pipelines that is extracted from historical runs of 20 datasets within 2 hours optimisation time.