[Supplementary Document] Exploring Opportunistic Meta-knowledge to Reduce Search Spaces for Automated Machine Learning

Tien-Dung Nguyen[1], David Jacob Kedziora[2], Katarzyna Musial[2], and Bogdan Gabrys[2]

Advanced Analytics Institute, University of Technology Sydney, Sydney, Australia
TienDung.Nguyen-2@student.uts.edu.au[1]
{David.Kedziora,Katarzyna.Musial-Gabrys,Bogdan.Gabrys}@uts.edu.au[2]

This document provides more details about the experiments for the paper entitled *Exploring Opportunistic Meta-knowledge to Reduce Search Spaces for Automated Machine Learning*.

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

Dataset	**r30**	**baseline**	**avatar**
abalone	CustomReplaceMissingValues → RandomSubset → Resample → Logistic → Bagging	SimpleLogistic	SMO
adult	J48	PART	PART
amazon	CustomReplaceMissingValues → Normalize → RandomSubset → NaiveBayesMultinomial → RandomSubSpace	NaiveBayes	NaiveBayesMultinomial
car	SMO→ MultiClassClassifier	SMO	SMO
cifar10small	RandomForest → MultiClassClassifier	DecisionStump	RandomForest
convex	RandomForest → AdaBoostM1	VotedPerceptron	RandomForest
dexter	DecisionStump → AdaBoostM1	NaiveBayesMultinomial	SGD
dorothea	OneR → RandomSubSpace	PART	DecisionStump
gcredit	LMT → Bagging	SMO	SMO
gisette	VotedPerceptron → RandomSubSpace	VotedPerceptron	VotedPerceptron
kddcup	-	DecisionStump	ClassBalancer → RemoveOutliers → InterquartileRange → AttributeSelection → Resample → PART
krvskp	Jrip → AdaBoostM1	RandomForest	J48
madelon	PrincipalComponents → IBk → LogitBoost	Jrip	Jrip
mnist	CustomReplaceMissingValues → Center → J48 → AdaBoostM1	-	PART
secom	J48 → AdaBoostM1	ClassBalancer → EMImputation → Normalize → PrincipalComponents → Kstar → MultiClassClassifier	SimpleLogistic
semeion	CustomReplaceMissingValues → PrincipalComponents → SMO	SMO	SMO
shuttle	RandomForest → AdaBoostM1	RandomForest	RandomForest
waveform	RemoveOutliers → InterquartileRange → Normalize → SMO → AttributeSelectedClassifier	LMT	SimpleLogistic
winequality	RandomForest → AdaBoostM1	RandomForest	Kstar
yeast	RandomForest → 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	CustomReplaceMissingValues → Normalize → RandomSubset → SimpleLogistic	ClassBalancer → RemoveOutliers → InterquartileRange → Normalize → RandomSubset → SimpleLogistic	REPTree	RandomForest	PART
adult	RemoveOutliers → InterquartileRange → Center → Logistic	Logistic	NaiveBayes	PART	PART
amazon	ClassBalancer → RemoveOutliers → InterquartileRange → Center → Logistic	SMO	NaiveBayesMultinomial	NaiveBayesMultinomial	NaiveBayesMultinomial
car	SMO	SMO	SMO	SMO	SMO
cifar10small	RandomSubset → PeriodicSampling → Logistic	NaiveBayesMultinomial	RandomForest	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	SpreadSubsample → CustomReplaceMissingValues → Normalize → RandomSubset → Resample → VotedPerceptron	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	SpreadSubsample → CustomReplaceMissingValues → Center → DecisionStump	NaiveBayes	NaiveBayes	NaiveBayes	OneR
gcredit	NaiveBayes	MultilayerPerceptron	SMO	MultilayerPerceptron	MultilayerPerceptron
gisette	VotedPerceptron	VotedPerceptron	VotedPerceptron	VotedPerceptron	VotedPerceptron
kddcup	MultilayerPerceptron	Ibk	DecisionStump	-	DecisionStump
krvskp	LMT	J48	J48	J48	J48
madelon	Jrip	Jrip	Jrip	Jrip	Jrip
mnist	SMO	-	SMO	RandomForest	SMO
secom	ClassBalancer → CustomReplaceMissingValues → RemoveOutliers → InterquartileRange → Normalize → PeriodicSampling → DecisionStump	SpreadSubsample → Standardize → RandomSubset → PeriodicSampling → ZeroR	Kstar	VotedPerceptron	Kstar
semeion	Logistic	SMO	SMO	SMO	SMO
shuttle	RandomForest	RandomForest	RandomForest	RandomForest	RandomForest
waveform	Logistic	SimpleLogistic	SimpleLogistic	SimpleLogistic	SimpleLogistic
winequality	RandomForest	RandomForest	RandomForest	Kstar	Kstar
yeast	RandomForest	RandomForest	RandomForest	RandomForest	RandomForest

Table 4. The best ML pipelines found using different methods to design configuration spaces (continued).

Dataset	**M-k1**	**M-k4**	**M-k8**	**M-k10**	**M-k19**
abalone	RandomForest	RandomForest	PART	REPTree	MultilayerPerceptron
adult	ClassBalancer → CustomReplaceMissingValues → RemoveOutliers → InterquartileRange → Standardize → RandomForest	PART	J48	J48	J48
amazon	RandomForest	J48	-	NaiveBayes	NaiveBayesMultinomial
car	RandomForest	J48	LMT	LMT	SMO
cifar10small	RandomForest	-	NaiveBayes	RandomForest	NaiveBayes
convex	RandomForest	RandomForest	RandomForest	RandomForest	RandomForest
dexter	Resample → RandomForest	RandomForest	SimpleLogistic	SimpleLogistic	SimpleLogistic
dorothea	ClassBalancer → Standardize → PrincipalComponents → RandomForest	-	-	-	-
gcredit	RandomForest	RandomForest	RandomForest	RandomForest	SMO
gisette	SpreadSubsample → CustomReplaceMissingValues → RandomForest	RandomForest	SimpleLogistic	JRip	JRip
kddcup	Standardize → ReservoirSample→ RandomForest	PART	PART	PART	DecisionStump
krvskp	RandomForest	J48	J48	J48	RandomForest
madelon	RandomForest	JRip	JRip	JRip	JRip
mnist	RandomForest	RandomForest	-	RandomForest	RandomForest
secom	RandomForest	RandomForest	RandomForest	RandomForest	KStar
semeion	RandomForest	RandomForest	RandomForest	RandomForest	SMO
shuttle	RandomForest	RandomForest	RandomForest	RandomForest	RandomForest
waveform	RandomForest	RandomForest	SimpleLogistic	LMT	LMT
winequality	RandomForest	RandomForest	RandomForest	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*.

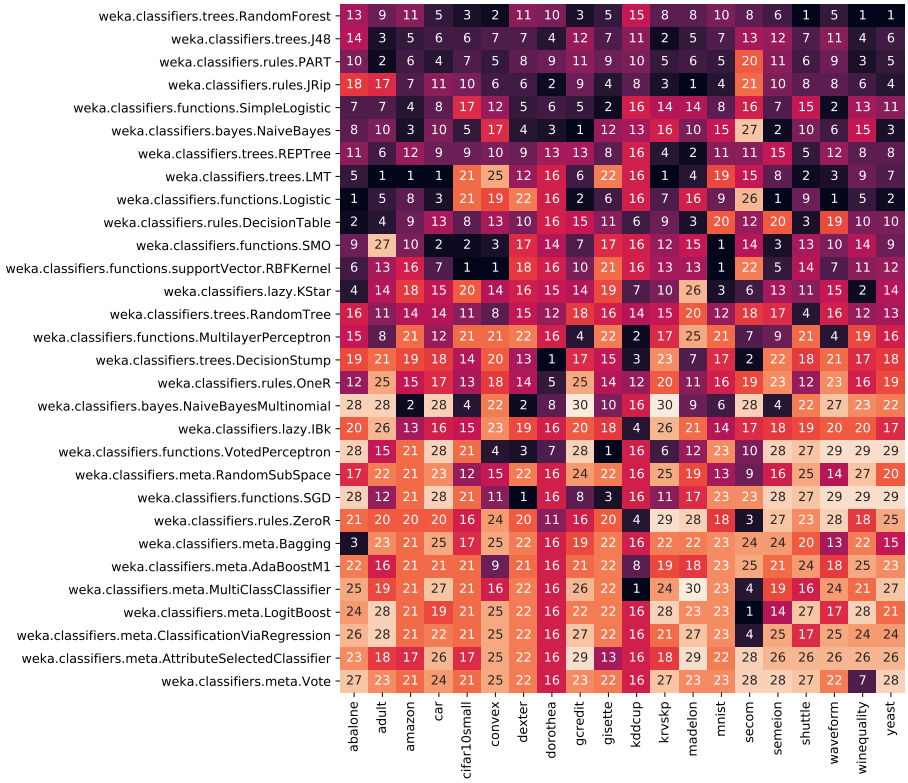

Fig. 1. Ranking of predictors based on mean error rate of their pipelines that is extracted from historical runs of 20 datasets within 2 hours optimisation time.