Supplementary Material of Selection of Data Science Pipeline on Sparse Training Records

Yawen Chen[†], Yile Chen[†], Zeyi Wen[‡], Jian Chen[†], Jin Huang[§]
[†]South China University of Technology, [‡]Hong Kong University of Science and Technology (Guangzhou),
[§]South China Normal University
ywchenscut@gmail.com, jireh.x6@gmail.com, wenzeyi@ust.hk, ellachen@scut.edu.cn,huangjin@m.scnu.edu.cn

APPENDIX A
INFORMATION OF BENCHMARKS

We show more information on the data sets and algorithms in *110-classifiers* and *OpenML* benchmarks in the following.

- *110-classifiers* benchmark: The meta-features used for the data sets in this benchmark are provided by Pymfe library [1] which are listed in Table II. The algorithms used are listed as Tabel I. More details can be found in the paper [2].
- *OpenML* benchmark: The meta-features used for the data sets in this benchmark are provided by OpenML platform which are summerized in Table III. We collect 681 different combinations of pre-processing, algorithms and hyper-parameters that have been run on at least 10 data sets on OpenML. The number of the data sets that have been trained with the collected combinations is 1069, and we keep 956 data sets of them whose meta-features are available on the OpenML platform.
The IDs of the collected 956 data sets on OpenML platform are:
[2-16, 18, 20, 22-32, 34-44, 46, 48-57, 59-62, 70-78, 115-144, 146-164, 171, 172, 179-188, 195, 210, 244-269, 271-279, 285, 293, 300, 307, 310-313, 316, 327-329, 333-340, 342, 343, 346, 350, 351, 354, 357, 373, 375, 377, 378, 381-

TABLE I: Algorithms in *110-classifiers*.

category	algorithm
decision trees	rpart_R, DecisionStump_weka , C5.0Tree_caret, RandomSubSpace_weka, NBTree_weka, RandomTree_weka, REPTree_weka, rpart_caret
discriminant analysis	lda_R, lda2_caret, rrlda_R, sda_caret, slda_caret, PenalizedLDA_R, sddaLDA_R, sddaQDA_R, fda_caret, fda_R, mda_R, rda_R, hdda_R, pda_caret
ensemble (Bagging)	Bagging_weka, ldaBag_R, nnetBag_R, MetaCost_weka
ensemble (Boosting)	logitboost_R, RacedIncrementalLogitBoost_weka, LogitBoost_weka, AdaBoostM1_weka, C5.0_caret, MultiBoostAB_IBk_weka, MultiBoostAB_OneR_weka, MultiBoostAB_PART_weka, MultiBoostAB_RandomTree_weka, MultiBoostAB_REPTree_weka, MultiBoostAB_weka, Bagging_DecisionStump_weka, Bagging_HyperPipes_weka, Bagging_J48_weka, Bagging_LWL_weka, Bagging_MultilayerPerceptron_weka, MultiBoostAB_MultilayerPerceptron_weka, MultiBoostAB_NaiveBayes_weka, Bagging_OneR_weka, Bagging_NaiveBayes_weka, Bagging_PART_weka, MultiBoostAB_Logistic_weka, Bagging_RandomTree_weka
ensemble (Forest)	parRF_caret, RotationForest_weka
ensemble (others)	RandomCommittee_weka, OrdinalClassClassifier_weka, MultiScheme_weka, MultiClassClassifier_weka, END_weka, Vote_weka, CostSensitiveClassifier_weka, Dagging_weka
general linear/ rule models	PART_weka, PART_caret, C5.0Rules_caret, JRip_caret, OneR_weka, OneR_caret, DTNB_weka, ZeroR_weka, gcvEarth_caret, DecisionTable_weka, ConjunctiveRule_weka, glm_R, glmnet_R, mlm_R, bayesglm_caret, SimpleLogistic_weka, multinom_caret
navie Bayes	naiveBayes_R, NaiveBayes_weka, NaiveBayesUpdateable_weka, BayesNet_weka
nearest neighbors	knn_R, knn_caret, IBk_weka, IB1_weka, spls_R, simpls_R
neural networks	rbf_caret, mlp_matlab, mlp_caret, cascor_C, avNNet_caret, nnet_caret, pcaNNet_caret, MultilayerPerceptron_weka, elm_matlab, mlp_C, lvq_R, bdk_R, dkp_C, dpp_C
SVMs	LibSVM_weka, LibLINEAR_weka, SMO_weka
other methods	pam_caret, HyperPipes_weka, FilteredClassifier_weka, ClassificationViaClustering_weka, AttributeSelectedClassifier_weka, ClassificationViaRegression_weka, VFI_weka

401, 443, 444, 446, 448, 450-455, 457-459, 461-470, 472, 474-477, 479-481, 488, 554, 679, 682, 683, 685, 694, 713-780, 782-808, 810-821, 823- 855, 857-871, 873- 882, 884-947, 949-1023, 1025, 1026, 1037-1042, 1044-1050, 1053-1057, 1059-1069, 1071, 1073, 1075, 1077-1088, 1100-1102, 1104, 1106, 1107, 1109-1117, 1119-1167, 1169, 1178-1183, 1185, 1186, 1205, 1209, 1211, 1212, 1214, 1218-1220, 1222, 1233, 1235-1238, 1240-1242, 1351-1410, 1413, 1441-1444, 1446, 1447, 1451-1453, 1455, 1457-1468, 1471-1473, 1475-1504, 1506-1520, 1523-1549, 1551-1560, 1562- 1569, 1590, 1596, 1597, 4134, 4135, 4153, 4154, 4329, 4340, 4534, 4538, 4552, 6332, 23380, 23381, 23499, 23512, 23517, 40474-40478, 40496-40499, 40514-40520, 40536, 40646-40648, 40650, 40660, 40663-40666, 40668-40671, 40677, 40678, 40680, 40681-40683, 40685-40687, 40690, 40691, 40693, 40700-40702, 40704-40711, 40713, 40714, 40900, 40910, 40923, 40926, 40927, 40966, 40971, 40975, 40978, 40979, 40981, 41168, 41169, 41496, 41526, 41671]. The IDs from OpenML platform of the flows which serve as the pipelines in this benchmark are: [56-60, 61, 62, 64-67, 70, 72, 74-87, 90-99, 101, 103, 105, 106, 108, 121, 124, 126, 130, 131, 133, 139, 144, 148, 150, 151, 156, 180, 182-184, 193, 198, 199, 204, 206, 208, 209, 212, 213, 364, 365, 375, 376, 378, 380, 384, 385, 387, 389-394, 396-398, 404-407, 411, 413, 417-420, 422, 423, 441, 471, 482, 506, 522-524, 527-535, 563, 582-585, 589, 591-593, 595-599, 611, 613, 615, 616, 622, 675, 677, 708, 710, 1068-1071, 1073-1080, 1082, 1084, 1087-1091, 1094-1096, 1098-1101, 1103-1106, 1108, 1111, 1112, 1114-1117, 1120, 1122-1125, 1127, 1129, 1130, 1132, 1133, 1135-1138, 1143, 1145, 1148, 1154, 1155, 1160, 1163, 1165, 1166, 1168, 1172, 1174, 1177-1180, 1182, 1183, 1185-1188, 1190-1197, 1199, 1200, 1244, 1349, 1350, 1716, 1718-1721, 1724-1730, 1745, 1750, 1789, 1805, 1817-1823, 1880, 1944, 1965, 1970, 2010, 2032, 2034, 2048, 2054, 2058, 2070, 2072, 2074, 2094, 2096, 2136, 2140, 2151, 2183, 2228, 2230, 2236, 2238, 2242-2245, 2247, 2250, 2254-2259, 2261-2270, 2272, 2277, 2278, 2283, 2291-2324, 2326-2328, 2330-2338, 2390-2393, 2408-2411, 2459, 2517, 2539, 2540, 2553-2555, 2560, 2561, 2563, 2565-2588, 2590-2599, 2601-2608, 2687-2690, 2697-2699, 2722, 2724, 2726, 2728, 2749-2751, 2753, 2754, 2762, 2763, 2775, 2777-2779, 2791, 2793, 3284, 3287, 3326, 3332, 3353, 3354, 3357-3364, 3416, 3418-3421, 3448, 3456-3461, 3463-3467, 3469-3481, 3548, 3554, 3558, 3564, 3568-3571, 3903, 3905, 3910, 3914, 3916, 3918, 3920, 3932, 3934, 3935, 3939, 3947, 3949, 3951, 3957, 3960, 3963, 3971, 3985, 4002, 4006, 4016, 4019, 4024, 4027, 4028, 4030, 4283, 4289, 4295, 4326, 4693, 4793, 4798-4812, 4814, 4821, 4822, 4825-4827, 4829, 4830, 4833-4835, 5434, 5528, 5531, 5533-5539, 5541, 5546, 5548, 5551, 5552, 5706, 5707, 5710, 5711, 5713-5715, 5721, 5724, 5725, 5728, 5909, 5910, 5978, 6023, 6840, 6946, 6952, 6969, 6970, 7026, 7089, 7096, 7116, 7122, 7170, 7694, 7707, 7722, 7725, 7729, 7754, 7756, 7777, 7778, 7781, 7782, 7784, 7786, 7787, 7789-7794, 7798-7801, 7835, 7836, 7838-7840, 7842-7845, 7847, 7849, 7850, 8299, 8308, 8309, 8311, 8312, 8315, 8317, 8330, 8351, 8353, 8365, 8399, 8455, 8456, 8673, 8690, 8692, 8693, 8695, 8774, 8786, 8788, 8789, 8793, 8795-8797, 8815, 8817, 8834, 8844, 8885, 8890, 8908, 8918, 9666, 9767, 12736, 12738, 13013, 13293, 13295, 15083, 16345, 16360, 17311, 17369, 17371, 17373, 17374, 17401, 17411, 17413, 17419, 17420, 17429, 17431, 17433, 17434, 17436, 17438, 17440, 17442, 17444, 17475, 17476, 17488, 17640, 17642, 18594].

REFERENCES

[1] E. Alcobaça, F. Siqueira, A. Rivolli, L. P. F. Garcia, J. T. Oliva, A. C. de Carvalho *et al.*, "Mfe: Towards reproducible meta-feature extraction." *The Journal of Machine Learning Research (JMLR)*, vol. 21, pp. 111–1, 2020.

[2] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The Journal of Machine Learning Research (JMLR)*, vol. 15, no. 1, pp. 3133–3181, 2014.

TABLE II: 86 meta-features for the learning problems in *110-classifiers*.

feature name	process	description
		general features
attr_to_inst	/	ratio between number of attributes.
freq_class	mean, sd	relative frequency of each distinct class.
inst_to_attr	/	ratio between number of instances and attributes.
nr_attr	/	total number of attributes.
nr_bin	/	number of binary attributes.
nr_class	/	number of distinct classes.
nr_inst	/	number of instances (rows) in dataset.
nr_num	/	number of numeric features.
		statistical
can_cor	mean	Compute canonical correlations of data.
cor	mean, sd	absolute value of correlation of distinct dataset column pairs.
cov	mean, sd	absolute value of covariance of distinct dataset attribute pairs.
eigenvalues	mean, sd	eigenvalues of covariance matrix from dataset.
gravity	/	distance between minority and majority classes center of mass.
iq_range	mean, sd	interquartile range (IQR) of each attribute.
kurtosis	mean, sd	kurtosis of each attribute.
mad	mean, sd	Median Absolute Deviation (MAD) adjusted by a factor.
max	mean, sd	maximum value from each attribute.
mean	mean, sd	mean value of each attribute.
median	mean, sd	median value from each attribute.
min	mean, sd	minimum value from each attribute.
nr_cor_attr	/	number of distinct highly correlated pair of attributes.
nr_disc	/	number of canonical correlation between each attribute and class.
nr_norm	/	number of attributes normally distributed based in a given method.
nr_outliers	/	number of attributes with at least one outlier value.
p_trace	/	Pillai's trace.
range	mean, sd	range (max - min) of each attribute.
sd	mean, sd	standard deviation of each attribute.
skewness	mean, sd	skewness for each attribute.
sparsity	mean, sd	Compute (possibly normalized) sparsity metric for each attribute.
t_mean	mean, sd	trimmed mean of each attribute.
var	mean, sd	variance of each attribute.
w_lambda	/	Wilks' Lambda value.
		info-theory
attr_conc	mean, sd	Compute concentration coef. of each pair of distinct attributes.
attr_ent	mean, sd	Compute Shannon's entropy for each predictive attribute.
class_conc	mean, sd	Compute concentration coefficient between each attribute and class.
class_ent	/	Compute target attribute Shannon's entropy.
eq_num_attr	/	number of attributes equivalent for a predictive task.
joint_ent	mean, sd	joint entropy between each attribute and class.
mut_inf	mean, sd	mutual information between each attribute and target.
ns_ratio	/	noisiness of attributes.
		model-based
leaves	/	number of leaf nodes in DT model.
leaves_branch	mean, sd	size of branches in DT model.
leaves_corrob	mean, sd	leaves corroboration of DT model.
leaves_homo	mean, sd	DT model Homogeneity for every leaf node.
leaves_per_class	mean, sd	proportion of leaves per class in DT model.
nodes	/	number of non-leaf nodes in DT model.
nodes_per_attr	/	ratio of nodes per number of attributes in DT model.
nodes_per_inst	/	ratio of non-leaf nodes per number of instances in DT model.
nodes_per_level	mean, sd	ratio of number of nodes per tree level in DT model.
nodes_repeated	mean, sd	number of repeated nodes in DT model.
tree_depth	mean, sd	depth of every node in DT model.
tree_imbalance	mean, sd	tree imbalance for each leaf node.
tree_shape	mean, sd	tree shape for every leaf node.
var_importance	mean, sd	features importance of DT model for each attribute.

TABLE III: 69 meta-features for the learning problems in *OpenML*.

feature name	description
	general
Dimensionality	The dimension of figure dataset.
NumberOfBinaryFeatures	The number of binary features.
NumberOfClasses	The number of classes.
NumberOfFeatures	The number of features.
NumberOfInstances	The number of instances.
NumberOfInstancesWithMissingValues	The number of instances with has missing values.
NumberOfMissingValues	The number of missing values.
NumberOfNumericFeatures	The number of numeric features.
NumberOfSymbolicFeatures	The number of symbolic features.
	statistical
AutoCorrelation	Compute correlations of data.
MajorityClassPercentage	The percentage of majority class.
MajorityClassSize	The size of majority class.
MaxNominalAttDistinctValues	The max value of distinct attributes.
MeanNominalAttDistinctValues	The mean value of distinct attributes.
MinNominalAttDistinctValues	The min value of distinct attributes.
MinorityClassPercentage	The percentage of minority class.
MinorityClassSize	The size of minority class.
PercentageOfBinaryFeatures	The percentage of binary features.
PercentageOfInstancesWithMissingValues	The percentage of instances with has missing values.
PercentageOfMissingValues	The percentage of missing values.
PercentageOfNumericFeatures	The percentage of numeric features.
PercentageOfSymbolicFeatures	The percentage of symbolic features.
StdvNominalAttDistinctValues	The standard deviation of each distinct attribute.
	info-theory
ClassEntropy	Compute target attribute Shannon's entropy.
	landmarking
CfsSubsetEval_DecisionStump	The performance of DecisionStrump with CfsSubsetEval.
CfsSubsetEval_NaiveBayes	The performance of NavieBayes with CfsSubsetEval.
CfsSubsetEval_kNN1N	The performance of 1-NN with CfsSubsetEval.
DecisionStump	The performance of DecisionStrump.
J48.00001	The performance of C4.5 DT with 0.00001 confidence factor.
J48.0001	The performance of C4.5 DT with 0.0001 confidence factor.
J48.001	The performance of C4.5 DT with 0.001 confidence factor.
NaiveBayes	The performance of NavieBayes.
REPTreeDepth1	The performance of 1-depth REP tree.
REPTreeDepth2	The performance of 2-depth REP tree.
REPTreeDepth3	The performance of 3-depth REP tree.
RandomTreeDepth1	The performance of 1-depth random tree.
RandomTreeDepth2	The performance of 2-depth random tree.
RandomTreeDepth3	The performance of 3-depth random tree.
kNN1N	The performance of 1-NN.