

Hybrid Intelligent Parsimony Search in Small High-dimensional Datasets (SHDD)

18th International Conference on Hybrid Artificial Intelligent Systems,
HAIS 2023, Salamanca (September, 2023)

J. Divasón, A. Pernia, A. Romero and F.J. Martinez-de-Pison*

University of La Rioja, Spain



Index

- 1 Introduction
- 2 Objective of this Work
- 3 Performance Analysis of HYB-PARSIMONY
- 4 Strategy for Working with SHDD
- 5 Conclusions



Introduction

Parsimonious Model Selection (PMS) on High Dimensional Tabular Data

- Nowadays, Autogluon is one of the best AutoML package for tabular problems. It constructs an ensemble with deep neural networks and tree-based algorithms such LightGBM, XGBoost, CatBoost, RandomForest, and so on.
- However, complex ensemble models that make use of advanced methods **may contain biases that are difficult to detect**.
- **That's why companies are progressively seeking interpretable models featuring a limited set of input variables, even if their accuracy doesn't match that of ensemble models.**
- A linear model or a decision tree might **prove more valuable in numerous decision-making scenarios due they are easier to maintain and understand; and could be more robust against perturbations or noise.**
- Even a black box model that boasts better accuracy, and has been developed using a narrower range of the original features, **can be more effectively analyzed using modern methods like ELI5 and SHAP.**

Introduction

Evolution: GA-PARSIMONY -> PSO-PARSIMONY -> HYB-PARSIMONY

- Our work over the years has focused on the development of methods on Parsimonious Model Selection (PMS).
- GA-PARSIMONY was proposed to search for parsimonious solutions with genetic algorithms (GA) by performing HO and FS. However, in this kind of problems where each solution has a high computational cost, it is not possible to evaluate a large number of individuals in each iteration. **This makes GA not as efficient as other optimization techniques where hundreds or thousands of individuals are evaluated.**
- PSO-PARSIMONY used PSO combined with a parsimony criterion to find parsimonious and accurate machine learning models. The comparison between both methods **showed that PSO always improved accuracy over GA, but GA found solutions approximately 10% less complex on datasets with a low number of features.**
- HYB-PARSIMONY **improved the search of parsimonious ML models against the other methodologies** by combining GA operations (selection, crossover and mutation) and PSO optimization. (available for Python at <https://github.com/jodivaso/HYBparsimony>).

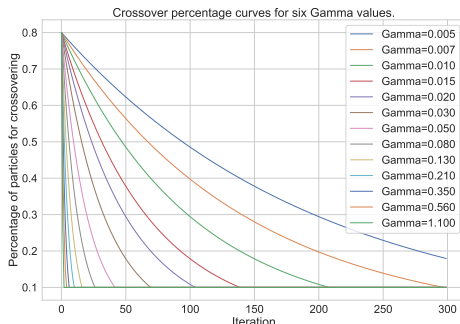


UNIVERSIDAD
DE LA RIOJA

Introduction

How HYB-PARSIMONY works?

- HYB-PARSIMONY uses $pcrossover = \max(0.80 \cdot e^{(-\Gamma \cdot t)}, 0.10)$ to calculate the **percentage of particles to be substituted by GA crossover in each iteration t .**
- In the first iterations, **the hybrid method performs the substitution by crossing a high percentage of particles.** As the optimization process progresses, the number of substituted particles **is reduced exponentially** until it ends up fixed at a percentage of 10%. Thus, **the hybrid method begins by facilitating the search for parsimonious models using GA-based mechanisms and ends up using more PSO optimization to improve the accuracy search.**



Index

- 1 Introduction
- 2 Objective of this Work
- 3 Performance Analysis of HYB-PARSIMONY
- 4 Strategy for Working with SHDD
- 5 Conclusions



Objective of this Work

Improve HYB-PARSIMONY with SHDD

- Creating accurate models with small high-dimensional datasets (SHDD) is a current challenge. If the dataset has hundreds or a few thousand instances, and the dimension is high (several tens or hundreds of features), the search for models that correctly generalize **the problem will face two fundamental problems: the curse of dimensionality and an excessive over-fitting in the optimization process.**
- HYB-PARSIMONY is such an intensive search method that when working with SHDD the method may find parsimonious solutions that are **too specific to that set of instances.** Thus, the chosen hyperparameters and the selected features may be the most appropriate for that sample **but not be sufficient to create a model that will generalize correctly in the future.**
- The objective of this study has been to develop a method **for analyzing the parsimony performance of HYB-PARSIMONY**, as well as to **establish a strategy for creating robust and concise models within the SHDD framework.**

Index

- 1 Introduction
- 2 Objective of this Work
- 3 Performance Analysis of HYB-PARSIMONY**
- 4 Strategy for Working with SHDD
- 5 Conclusions

Performance Analysis of HYB-PARSIMONY

Experiment's Parameters

- *method*: HYB vs previous methods (PSO or GA).
- *nruns*: number of runs with different random seeds. Value: 10.
- Γ (only for the hybrid method). Values: 0.005, 0.007, 0.010, 0.015, 0.020, 0.030, 0.050, 0.080, 0.130, 0.210, 0.350, 0.560, 1.100, 1.170.
- P : population size. Values: $[5, 5 + 1 \cdot 5, 5 + 2 \cdot 5, \dots, 40]$.
- *#feats*: dimension of the hypothetical data set. Values: 50, 150, 250, 350.
- i_{dim} : intrinsic dimension that refers to the features, F_{selec} , with relevant information present in a dataset. That is, the number of input features of the hypothetical model that explains an hypothetical target. Values: $5, 5 + 1 \cdot 20, 5 + 2 \cdot 20, \dots, \lfloor 0.90 \cdot \#feats \rfloor$.
- β : value which balances the weight between recall and precision in the F_{beta} score used to evaluate each individual (see below). Values: $[0.20, 0.20 + 1 \cdot 0.06, 0.20 + 2 \cdot 0.06, \dots, 1.68]$.

Performance Analysis of HYB-PARSIMONY

Evaluation of the Efficiency of HYB-PARSIMONY in Finding the Correct Features (I)

For each combination of experiment's parameters, F_{selec} were randomly selected according to i_{dim} . F_{selec} corresponded to i_{dim} random feature positions selected within the range $[0, \#feats - 1]$. Total = 115170 experiments.

To evaluate each solution, F_{beta} score was used. F_{beta} is the weighted harmonic mean of precision and recall. $\beta < 1$ gives more weight to precision, while $\beta > 1$ favors recall. F_{beta} is equal to $F1$ score with $\beta = 1.0$ and to precision with $\beta = 0.0$.

It is defined as:

$$F_{beta} = \frac{(1 + \beta^2)(precision \cdot recall)}{\beta^2 \cdot precision + recall} \quad (1)$$

and:

$$precision = \frac{TP}{TP + FP}; recall = \frac{TP}{TP + FN} \quad (2)$$

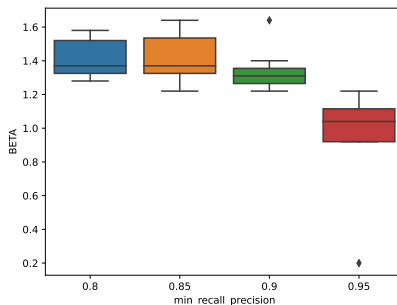
where TP are the correctly chosen features belonging to F_{selec} , TN the features not chosen and not belonging to F_{selec} , FP the features chosen but not belonging to F_{selec} , and FN the features not chosen but belonging to F_{selec} .



Performance Analysis of HYB-PARSIMONY

Evaluation of the Efficiency of HYB-PARSIMONY in Finding the Correct Features (II)

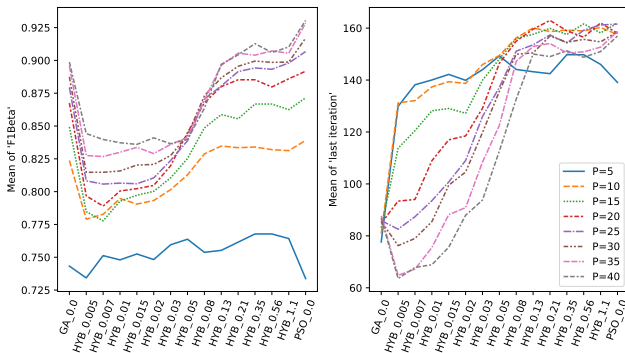
Figure shows the number of experiments that successfully met the objectives of overcoming a minimum precision and recall defined by a threshold, thr_{pr} , and by each β and P values. At low thr_{pr} values, the median of the best β for each combination of $[method, \Gamma, P, \#feats, i_{dim}]$ is about 1.3. This indicates that **precision tends to be prioritized over recall**. Only at $thr_{pr} = 0.95$ it is observed that the median of the best β is close to 1.0, so the relationship between precision and recall is balanced **when the level of demand is very high**.



Performance Analysis of HYB-PARSIMONY

Evaluation of the Efficiency of HYB-PARSIMONY in Finding the Correct Features (III)

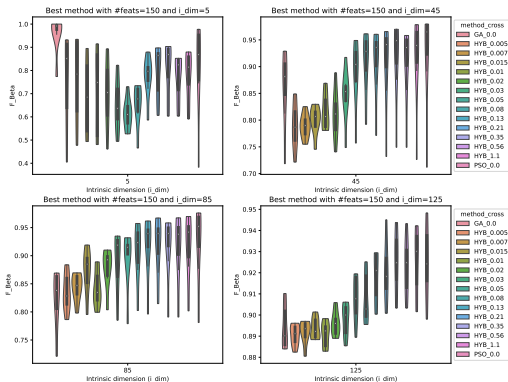
Figure shows respectively the mean of the last iteration ($last_{iter}$) (right) and the average of F_{beta} (left) with $\beta = 1.34$ and for each method and P . **GA and HYB with low Γ converge, on average, faster than PSO and HYB with high Γ values, as they reach twice the number of final iterations.** With respect to F_{beta} , **the highest averages are obtained with PSO and HYB with Γ values greater than 0.08, but GA has similar performance to HYB_0.08.**



Performance Analysis of HYB-PARSIMONY

Evaluation of the Efficiency of HYB-PARSIMONY in Finding the Correct Features (IV)

Results are average values that may be different for each method, $\#feats$ and i_{dim} . Figure shows the distribution of F_{beta} for each method with $\#feats = 150$, $\beta = 1.34$ and two four different i_{dim} values: 5, 45, 85 and 125. At very low values of i_{dim} , GA is competitive with the hybrid and PSO methods. However, as the intrinsic dimension is closer to the real dimension of the dataset, PSO and hybrid models with high Γ obtain better accuracy. The problem is that for a particular dataset the intrinsic dimension of the data will be unknown, so it will be necessary to realize an estimation of i_{dim} in order to select an appropriate method.



Modeling Performance Analysis of HYB-PARSIMONY

Ridge Models for Estimating F_{beta} and $last_{iter}$

- Ridge models were trained with the previously obtained dataset but eliminating instances corresponding to GA and PSO, and **selecting only those cases with a β within the range $[0.92, 1.64]$ where the methodology was most successful.**
- Equations correspond to the best Ridge models selected with a 10-fold cross-validation RMSE error of 0.0815 and 57.36 with values of the α Ridge hyperparameter equal to 4.0 and 2.0, respectively.
- $\hat{F}_{beta} = -0.0462 \cdot \Gamma - 0.0027 \cdot P + 0.0012 \cdot \#feats - 0.0011 \cdot i_{dim} - 0.0108 \cdot \beta - 0.88$
- $\hat{last}_{iter} = 28.391 \cdot \Gamma - 0.8883 \cdot P + 0.2963 \cdot \#feats - 0.38 \cdot i_{dim} + 36.517 \cdot \beta + 72.71$

Index

- 1 Introduction
- 2 Objective of this Work
- 3 Performance Analysis of HYB-PARSIMONY
- 4 Strategy for Working with SHDD**
- 5 Conclusions



Strategy for Working with SHDD

Challenge in Working with Small High Dimensional Datasets (SHDD)

- Creating accurate models with SHDD is a current challenge. The search for models that correctly generalize the problem **will face two fundamental problems: the curse of dimensionality and an excessive over-fitting in the optimization process.**
- Although there are algorithms, such as trees and neural networks, that may be less affected by the curse of dimensionality **is highly recommended to use feature selection or dimensional reduction with SHDD.**
- HYB-PARSIMONY greatly facilitates both aspects since it seeks to reduce as much as possible the number of features of the selected model, as well as its internal complexity.
- However, **HYB-PARSIMONY is such an intensive search method that when working with SHDD the method may find parsimonious solutions that are too specific to that set of instances.** Thus, the chosen hyperparameters and the selected features **may be the most appropriate for that sample but not be sufficient to create a model that will generalize correctly in the future.**

Strategy for Working with SHDD

Methodology for Working with SHDD (HYB-PARSIMONY-SHDD)

To reduce this over-fitting and to find a feature selection that can be used to create a robust model that generalizes correctly in this kind of problems, we propose the HYB-PARSIMONY-SHDD methodology:

- 1 Repeat n runs with different random seeds the search for the best model with HYB-PARSIMONY and hold-out validation. In each repetition, extract the feature probability vector of the best individual.
- 2 Average the probabilities for each feature and select those that have a value greater than a given threshold, thr_{fs} .
- 3 Performs hyperparameter tuning with BO and the features selected in the previous point.
- 4 Repeat points 2 and 3 with different thr_{fs} .
- 5 Select the model that obtains the best error J with another test dataset.

HYB-PARSIMONY-SHDD

Results obtained with BO (J) vs HYB-PARSIMONY-SHDD with $thr_{fs} = 0.50$

Results for 7 datasets obtained with BO (J) versus HYB-PARSIMONY-SHDD with $thr_{fs} = 0.50$ followed by BO. First, 25 runs of HYB-PARSIMONY were performed with $\Gamma = 0.50$, $nruns = 200$, $P = 15$, $early_stopping = 35$, hold-out validation with a 20%, $ReRank = 0.001$, and *KernelRidge* as ML algorithm. Next, BO was used with the features with mean of $probs \geq 0.50$ ($thr_{fs} = 0.50$). Columns in table indicate: the last iteration ($last_{iter}$) of HYB-PARSIMONY, RMSE ($J0.5$) and the number of features used ($Fs0.5$) in the final model. Results correspond to the average values and the standard deviation (in parentheses) of 5 runs of the whole methodology with different random seeds.

Dataset				BO	HYB-PARSIMONY-SHDD		
name	$train_{size}$	$test_{size}$	#feats	J_{BO}	$last_{iter}$	$J0.5$	$Fs0.5$
slice	2000	23000	378	.1414	165.8 (42.4)	.1449 (.0022)	148.0 (3.74)
blog	2000	50397	276	.8216	178.6 (45.6)	.8154 (.0170)	67.0 (17.42)
crime	1107	1108	127	.6373	199.0 (0.0)	.6379 (.0076)	28.8 (4.32)
aileron	2000	11750	40	.3984	131.2 (17.9)	.3982 (.0016)	13.2 (2.05)
bank	2000	6192	32	.6745	160.0 (32.1)	.6726 (.0023)	15.6 (1.52)
puma	2000	6192	32	.8762	106.2 (16.3)	.2006 (.0250)	3.6 (0.89)
pol	2000	13000	26	.3165	148.0 (32.2)	.2413 (.0034)	7.2 (0.45)



HYB-PARSIMONY-SHDD

HYB-PARSIMONY-SHDD with different thr_{fs}

- Previous results with $thr_{fs} = 0.50$ could be improved by using different thr_{fs} values.
- The proposed methodology **obtained more accurate models in all datasets**.
- But **the most outstanding results were observed in the significant reduction of the number of average features**.

dataset	#feats	J_{BO}	$J0.2$	$Fs0.2$	$J0.3$	$Fs0.3$	$J0.4$	$Fs0.4$	$J0.5$	$Fs0.5$	$J0.6$	$Fs0.6$	$J0.7$	$Fs0.7$
slice	378	.1414	.1392	346.4	.1374	294.2	.1370	228.2	.1449	148.0	.1583	85.8	.2267	40.2
blog	276	.8216	.8265	247.0	.8288	196.8	.8234	128.8	.8154	67.0	1.014	29.4	.9462	10.8
crime	127	.6373	.6367	108.8	.6371	85.0	.6386	56.4	.6379	28.8	.6461	12.8	.7867	3.8
aileron	40	.3984	.3986	33.2	.3994	25.4	.3993	17.8	.3982	13.2	.4338	9.8	.5043	5.6
bank	32	.6745	.6729	29.4	.6756	24.6	.6744	20.4	.6726	15.6	.6791	11.4	.6889	9.0
puma	32	.8762	.3030	12.2	.2308	7.6	.2096	4.6	.2006	3.6	.2007	3.6	.2230	2.8
pol	26	.3165	.2736	13.2	.2584	10.4	.2461	8.2	.2413	7.2	.2413	7.2	.2387	6.4

HYB-PARSIMONY-SHDD

HYB-PARSIMONY-SHDD vs. PSO-PARSIMONY

- The results show that the new methodology improved J in the four higher dimensionality datasets in conjunction with a considerable reduction in the number of features.
- HYB-PARSIMONY-SHDD obtained better J in bank, puma and pol, but with worse parsimony in the first two.
- However, other higher Γ values can be used in order to approximate to the PSO-PARSIMONY method.

dataset	HYB-PARSIMONY-SHDD				PSO-PARSIMONY			
	$last_{iter}$	thr_{fs}	J	Fs	$last_{iter}$	thr_{fs}	J	Fs
slice	165.8	.4	.1370(.003)	228.2(6.0)	182.8	.5	.1372(.002)	239.0(6.9)
blog	178.6	.5	.8154(.017)	67.0(17.4)	191.6	.0	.8215(.000)	276.0(0.0)
crime	199.0	.2	.6367(.002)	108.8(1.3)	175.4	.1	.6371(.000)	124.8(1.3)
aileron	131.2	.5	.3982(.002)	13.2(2.1)	154.8	.5	.3984(.002)	16.0(3.5)
bank	160.0	.5	.6725(.002)	15.6(1.5)	149.8	.5	.6724(.002)	16.8(1.6)
puma	106.2	.5	.2006(.025)	3.6(0.9)	104.4	.4	.1894(.000)	4.0(0.0)
pol	148.0	.7	.2387(.004)	6.4(0.6)	127.2	.7	.2374(.003)	6.2(0.5)

HYB-PARSIMONY-SHDD

HYB-PARSIMONY-SHDD vs. SKLEARN-GENETICS-OPT

- Table shows a comparison of the proposed methodology versus the use of the `sklearn-genetic-opt` library for feature selection using 10 runs with different random seeds.
- With this package the procedure consisted of three steps: perform the hyperparameter adjustment with GA, select with GA the best features and finally perform again the hyperparameter tuning with GA but using only the selected variables.
- The new methodology obtained better results with statistical significance in all datasets (p -value < 0.05 in Wilcoxon–Mann–Whitney test), although in some cases the feature reduction was lower.

dataset	J_{hyb}	J_{gen}	p -value	Fs_{hyb}	Fs_{gen}	p -value
slice	0.131430	0.142033	0.000330	183.3	202.4	0.000489
blog	0.823028	0.841876	0.002827	275.8	143.6	0.000085
crime	0.630145	0.633698	0.011330	70.7	67.7	0.148308
aileron	0.398032	0.398548	0.025748	40.0	22.9	0.000062
bank	0.672656	0.676080	0.021134	24.7	19.8	0.000143
puma	0.189436	0.258673	0.000183	4.0	11.0	0.000048
pol	0.236990	0.276652	0.000183	6.0	15.7	0.000058

HYB-PARSIMONY-SHDD

HYB-PARSIMONY-SHDD with 19 OpenML Binary Datasets

- HYB-PARSIMONY on its own performs quite effectively, leading to a significant reduction in NFs.
- HYB-PARSIMONY-SHDD it achieves excellent NFs reduction, particularly when $n_{cols} \geq 80$. At times, the NFs reduction can be truly remarkable in scenarios such as 'scene,' 'clean2,' and 'national'.
- **Recommendation** for $n_{cols} > 50$: First, employ BO with all features. Second, use HYB-PARSIMONY by itself. Third, employ HYB-PARSIMONY-SHDD. Then, select the most suitable model in accuracy and parsimony.

				BO	HYB-PARSIMONY			HYB-PARSIMONY-SHDD				GA-SKLEARN-OPT	
name_file	n_rows	n_cols	trnval_size	bo_TST	hyb_NFS	hyb_TST	thr_feats	NFS_SHDD	TST_SHDD	ga_NFS	ga_TST		
philippine.csv	5832	308	2916	0.557	62.6	0.542	0.4	103.8	0.557	151.6	0.552		
scene.csv	2407	304	1203	0.200	49.2	0.061	0.6	10.8	0.043	142.6	0.108		
jasmine.csv	2984	280	1492	0.445	71.4	0.448	0.4	92.8	0.452	141	0.446		
madeline.csv	3140	259	1570	0.670	54	0.680	0.5	43.8	0.696	126.8	0.672		
clean2.csv	6598	168	3299	0.000	36	0.000	0.4	29.6	0.000	73.8	0.000		
COIL2000.csv	5822	150	2911	0.210	35.6	0.213	0.4	43.6	0.215	77.2	0.211		
ECG5000.csv	4998	140	2499	0.044	29.4	0.050	0.4	42	0.048	68.6	0.044		
yeast.csv	2417	129	1208	0.067	38	0.114	0.5	32.5	0.090	66.5	0.072		
mushroom.csv	8124	125	4062	0.000	59.4	0.000	0.3	116.8	0.000	70.4	0.000		
sylvia.csv	14395	108	7197	0.018	55.8	0.020	0.5	45.8	0.019	61.2	0.019		
ada.csv	4562	102	2281	0.353	43.4	0.356	0.5	37.8	0.355	56.6	0.353		
coil2000.csv	9822	85	4911	0.204	23	0.206	0.4	31.8	0.206	42.2	0.204		
national.csv	4908	80	2454	0.019	13.4	0.009	0.6	2.6	0.006	28.8	0.009		
kr.csv	3196	74	1598	0.101	33.6	0.099	0.4	64.4	0.102	42	0.102		
PhishingWebsites.csv	11055	68	5527	0.143	30.2	0.143	0.4	47.8	0.143	40.4	0.143		
test.csv	15547	60	7773	0.455	43.4	0.458	0.5	46.8	0.456	43	0.459		
spambase.csv	4601	57	2300	0.235	31	0.240	0.5	38.6	0.246	34.8	0.250		
dis.csv	3772	54	1886	0.069	19.2	0.066	0.4	28.8	0.068	30.8	0.067		
sick.csv	3772	52	1886	0.108	10.2	0.109	0.4	17	0.109	21.8	0.109		

Index

- 1 Introduction
- 2 Objective of this Work
- 3 Performance Analysis of HYB-PARSIMONY
- 4 Strategy for Working with SHDD
- 5 Conclusions**



Conclusions (I)

- GA-PARSIMONY, PSO-PARSIMONY and HYB-PARSIMONY are methodologies that have been developed for the search of accurate but low complexity ML models. However, **an intensive search with these methods in SHDD can lead to overfitted models.**

Conclusions (I)

- GA-PARSIMONY, PSO-PARSIMONY and HYB-PARSIMONY are methodologies that have been developed for the search of accurate but low complexity ML models. However, **an intensive search with these methods in SHDD can lead to overfitted models.**
- The proposed HYB-PARSIMONY-SHDD methodology is based **on bootstrapping by repeating HYB-PARSIMONY with different random seeds and by using hold-out validation.** In this way, at each run the search for the best model is validated with a different part of the dataset.

Conclusions (I)

- GA-PARSIMONY, PSO-PARSIMONY and HYB-PARSIMONY are methodologies that have been developed for the search of accurate but low complexity ML models. However, **an intensive search with these methods in SHDD can lead to overfitted models.**
- The proposed HYB-PARSIMONY-SHDD methodology is based **on bootstrapping by repeating HYB-PARSIMONY with different random seeds and by using hold-out validation.** In this way, at each run the search for the best model is validated with a different part of the dataset.
- By averaging the feature probability vectors, **a more robust final feature selection can be achieved.** After selecting these features through experimentation with different thresholds, Bayesian Optimization (BO) is employed to fine-tune the model's hyperparameters



Conclusions (and II)

- Results demonstrated that it **is possible to obtain more accurate models with a significant reduction in the number of features against other methodologies.**

Conclusions (and II)

- Results demonstrated that it **is possible to obtain more accurate models with a significant reduction in the number of features against other methodologies.**
- In SHDD problems, when sufficient computational time is at hand, **HYB-PARSIMONY-SHDD can be employed with various values for Γ , *Rerank*, and thr_{fs} to yield the most robust and concise model possible.**

Conclusions (and II)

- Results demonstrated that it is **possible to obtain more accurate models with a significant reduction in the number of features against other methodologies.**
- In SHDD problems, when sufficient computational time is at hand, **HYB-PARSIMONY-SHDD can be employed with various values for Γ , $Rerank$, and thr_{fs} to yield the most robust and concise model possible.**
- **Recommendation for $n_{cols} > 50$:** First, employ BO with all features. Second, use HYB-PARSIMONY by itself. Third, employ HYB-PARSIMONY-SHDD. Then, compare and select the most suitable model in accuracy and parsimony.

References

- Divasón, J., Pernia-Espinoza, A., Martinez-de-Pison, F.J. New Hybrid Methodology Based on Particle Swarm Optimization with Genetic Algorithms to Improve the Search of Parsimonious Models in High-Dimensional Databases. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 13469 LNAI, pp. 335-347 (2022).
- Divasón, J., Cenicerros, J.F., Sanz-Garcia, A., Pernia-Espinoza, A., Martinez-de-Pison, F.J. PSO-PARSIMONY: A method for finding parsimonious and accurate machine learning models with particle swarm optimization. Application for predicting force?displacement curves in T-stub steel connections. Neurocomputing 548, 1 – 12 (2023).
- Martinez-de Pison, F.J., Ferreiro, J., Fraile, E., Pernia-Espinoza, A.: A comparative study of six model complexity metrics to search for parsimonious models with GAParsimony R Package. Neurocomputing 452, 317 – 332 (2021).
- Urraca, R., Sodupe-Ortega, E., Antonanzas, J., Antonanzas-Torres, F., de Pison, F.M.: Evaluation of a novel GA-based methodology for model structure selection: The GA-PARSIMONY. Neurocomputing 271(Supplement C) 9 – 17 (2018).



THANK YOU VERY MUCH
FOR YOUR ATTENTION



**UNIVERSIDAD
DE LA RIOJA**

Questions?