
PyHard: a novel tool for generating hardness embeddings to support data-centric analysis

Pedro Yuri Arbs Paiva

Instituto Tecnológico de Aeronáutica (ITA)
Praça Marechal Eduardo Gomes, 50
São José dos Campos, SP, Brazil
paiva@ita.br

Kate Smith-Miles

School of Mathematics and Statistics
The University of Melbourne
Parkville VIC 3010, Australia
smith-miles@unimelb.edu.au

Maria G. Valeriano

Instituto Tecnológico de Aeronáutica (ITA)
Praça Marechal Eduardo Gomes, 50
São José dos Campos, SP, Brazil
maria.valeriano@ita.br

Ana Carolina Lorena

Instituto Tecnológico de Aeronáutica (ITA)
Praça Marechal Eduardo Gomes, 50
São José dos Campos, SP, Brazil
aclorena@ita.br

Abstract

For building successful Machine Learning (ML) systems, it is imperative to have high quality data and well tuned learning models. But how can one assess the quality of a given dataset? And how can the strengths and weaknesses of a model on a dataset be revealed? Our new tool PyHard employs a methodology known as Instance Space Analysis (ISA) to produce a hardness embedding of a dataset relating the predictive performance of multiple ML models to estimated instance hardness meta-features. This space is built so that observations are distributed linearly regarding how hard they are to classify. The user can visually interact with this embedding in multiple ways and obtain useful insights about data and algorithmic performance along the individual observations of the dataset. We show in a COVID prognosis dataset how this analysis supported the identification of pockets of hard observations that challenge ML models and are therefore worth closer inspection, and the delineation of regions of strengths and weaknesses of ML models.

1 Introduction

A recent Meta-Learning (MtL) (Vilalta and Drissi, 2002; Vanschoren, 2019) methodology known as Instance Space Analysis (ISA) has demonstrated success in gaining insights through the visual analysis of the relationships between the performance of ML classification techniques and data meta-features across multiple datasets (Muñoz et al., 2018). Using ISA, the performance of multiple ML classification techniques can be predicted and understood in terms of a set of meta-features describing popular classification datasets from public repositories. A 2-D projection relating meta-features values and algorithmic performance is then created, presenting linear trends that reveal pockets of hard and easy datasets, that is, datasets that are harder or easier for the algorithms to classify, revealing the strengths and weaknesses of each classifier.

We adapt here such a framework for a fine-grained analysis of a single classification dataset, rather than a collection of datasets, so that we can obtain similar insights about the difficulties of individual observations within a dataset. Given a dataset of interest, multiple measures of the hardness level of its observations are extracted and related to the algorithmic performance of popular ML models. As a result, a 2-D hardness embedding of the original dataset is obtained, which can be scrutinized to

inspect data quality and understand classifier behaviors. We present herein *PyHard*, an analytical tool designed to assist the understanding of the difficulty level of the individual observations in a dataset, and also to trace regions where each algorithm is expected to perform well. We show in a case study with a dataset for COVID-19 prognosis how this tool supported the identification of problematic observations and the understanding of the behavior of different classifiers across the dataset.

2 Instance Space Analysis for a single dataset

As originally proposed, the Instance Space Analysis (ISA) builds upon the Algorithm Selection Problem (ASP) (Rice, 1976; Smith-Miles, 2009), whose objective is to automate the process of selecting good candidate algorithms and their hyperparameters for solving new problems, based on knowledge gathered from similar problems they solved in the past. The ISA framework goes further and extends the ASP analysis to give insights into why some instances are harder to solve than others, combining the information of meta-features and algorithm performance in a new embedding that can be visually inspected. To this end, an optimization problem is solved to find the mapping from the high dimensional meta-features space into a 2-D space, such that the distribution of algorithm performance metrics and meta-feature values across instances in the 2-D space displays as much of a linear trend as possible to assist the interpretation of hardness directions. Footprints of algorithm performance representing regions where each algorithm performs well can also be delineated, allowing to characterize their domains of competence. Objective measures of algorithmic performance along the IS can be extracted from the footprints too, such as area, density and purity. Details on the construction of the instance space and calculation of algorithm footprints can be found at (Muñoz et al., 2018; Muñoz et al., 2021).

Here we recast the framework for the analysis of a single classification dataset, as outlined in Figure 1. Given a dataset \mathcal{D} with n labeled observations \mathbf{x}_i , a set of *hardness measures* (HM) \mathcal{F} assessing their difficulty (aka hardness level) is extracted. A pool of ML classifiers of distinct biases are applied to the dataset and compose the algorithm space \mathcal{A} , namely: Bagging, Gradient Boosting, Support Vector Machine (both linear and RBF kernels), Logistic Regression, Multilayer Perceptron and Random Forest (but alternative algorithms can be added). The performance space \mathcal{Y} records the performance obtained by each algorithm in \mathcal{A} for each instance $\mathbf{x}_i \in \mathcal{D}$. A meta-dataset \mathcal{M} is built by joining a subset of \mathcal{F} and the set \mathcal{Y} . The former subset contains meta-features which relates more to the performance space \mathcal{Y} , which are screened in a feature selection step. ISA is then applied for obtaining the 2-D hardness embedding (instance space). Some of the most specific components of our framework are described next.

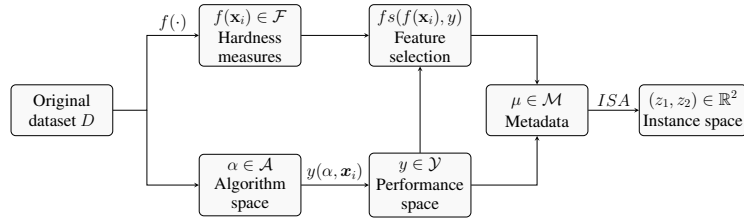


Figure 1: ISA framework for a single dataset.

Hardness measures. One important aspect when performing ISA is using a set of informative meta-features able to reveal the capabilities of the algorithms and the level of difficulty each individual instance poses. Here we revisit the definition of *Instance Hardness* (IH) proposed by Smith et al. (2014) as a property that indicates the likelihood that an instance will be misclassified. IH is measured by Equation 1, where $p(c_i|\mathbf{x}_i, \alpha_j)$ gives the probability a learning model α_j induced from \mathcal{D} assigns \mathbf{x}_i to its expected class c_i . According to this formulation, observations that are frequently misclassified by a pool \mathcal{A} of diverse learning algorithms are considered hard, whilst easy instances are likely to be correctly classified by any of the considered algorithms.

$$IH_{\mathcal{A}}(\mathbf{x}_i, c_i) = 1 - \frac{1}{|\mathcal{A}|} \sum_{j=1}^{|\mathcal{A}|} p(c_i|\mathbf{x}_i, \alpha_j) \quad (1)$$

Smith et al. (2014) also define a set of *hardness measures* (HM) intended to explain why some instances are often misclassified. These are the measures employed as meta-features in \mathcal{F} . Table 1

Table 1: Hardness measures employed as meta-features in this work.

Measure	Acron.	Min	Max	Reference
k-Disagreeing Neighbors	kDN	0	1	(Smith et al., 2014)
Disjunct Class Percentage	DCP	0	1	(Smith et al., 2014)
Tree Depth (pruned)	TD_P	0	1	(Smith et al., 2014)
Tree Depth (unpruned)	TD_U	0	1	(Smith et al., 2014)
Class Likelihood	CL	0	1	(Smith et al., 2014)
Class Likelihood Difference	CLD	0	1	(Smith et al., 2014)
Frac. features in overlapping areas	$F1$	0	1	(Arruda et al., 2020)
Frac. nearby instances of different class	$N1$	0	1	(Arruda et al., 2020)
Ratio of intra-extra class distances	$N2$	0	≈ 1	(Arruda et al., 2020)
Local set cardinality	LSC	0	1	(Arruda et al., 2020)
Local set radius	LSR	0	1	(Arruda et al., 2020)
Usefulness	U	≈ 0	1	(Arruda et al., 2020)
Harmfulness	H	0	≈ 1	(Arruda et al., 2020)

summarizes the HM employed in this work, with their names, acronyms, minimum and maximum values achievable and references from where they are extracted. We introduced modifications into some of the measures in order to limit and standardize their values. Consequently, for all measures higher values are registered for instances that are harder to classify.

Algorithm performance assessment. Given a dataset D , we first split it into five folds according to the cross-validation (CV) strategy, such that each instance belongs to one of the test sets. For each instance, we measure the *log-loss error rate* achieved in the predictions, which takes into account the probabilities associated to each class. Furthermore, a hyper-parameter optimization step was added, acting as an inner loop for each of the training sets of the outer CV. Within this inner loop, a candidate set of parameters is evaluated through cross-validation upon the training data from the outer loop.

Feature selection. According to Muñoz et al. (2018), it is advisable to maintain just the most informative meta-features in the meta-dataset \mathcal{M} before the IS projection is generated. We performed a supervised meta-feature selection in \mathcal{M} , based on the continuous response value $y(\alpha_k, \mathbf{x}_i)$, that is, the log-loss performance of the classifiers for the instances in \mathcal{D} . Since there are seven classification algorithms in the pool, a ranking of meta-features for each one of them is obtained. Next, a rank aggregation method is employed to merge these subsets, as suggested in (Prati, 2012).

IS representation and footprints. PyHard embeds a Python package named PyISpace implementing the main ISA functionalities, including obtaining the projections and the algorithms’ footprints. In order to assist the visual interpretation of the 2-D projections, we introduce a rotation step in order to ensure that hard instances are always placed towards the upper left corner of the IS, whilst the easier instances are placed towards the bottom right corner of the space. We also introduce an *instance easiness footprint*, which are regions of the IS which encompass observations with IH values below a threshold of 0.4. These are the easiest instances of the dataset for most of the algorithms.

3 Case study: inspecting a COVID prognosis dataset

Here we present an analysis of a dataset containing anonymized data from citizens positively diagnosed for COVID-19, collected from March 1st, 2020 to April 15th, 2021¹. The dataset involves predicting whether an individual will require future hospitalization or not taking as input the following attributes routinely supplied during COVID scanning: age, sex, initial symptoms (fever, cough, sore throat, dyspnea, respiratory distress, low saturation, diarrhea, vomit and other symptoms) and comorbidities (chronic cardiovascular disease, immunodeficiency-immunodepression, chronic kidney disease, diabetes mellitus, obesity, chronic respiratory diseases and other risks). The dataset has data from 5,156 citizens, half of which were hospitalized.

Our objective is to analyze the hardness profile of this dataset and to extract some insights from the visualization and interaction with its IS². Figure 2 presents the ISA of the hereafter called hospitalization dataset. Each point corresponds to an observation of the dataset, that is, a confirmed

¹The data comes from a partnership with the health department of the city of São José dos Campos - SP, Brazil. Part of such data was formatted for some predictive analysis to support public health decision making.

²The complete experiment with anonymizations can be found at: <https://gitlab.com/ita-ml/pyhard/-/tree/master/experiments/hospitalization>. Instructions can be found in ‘experiments’ folder.

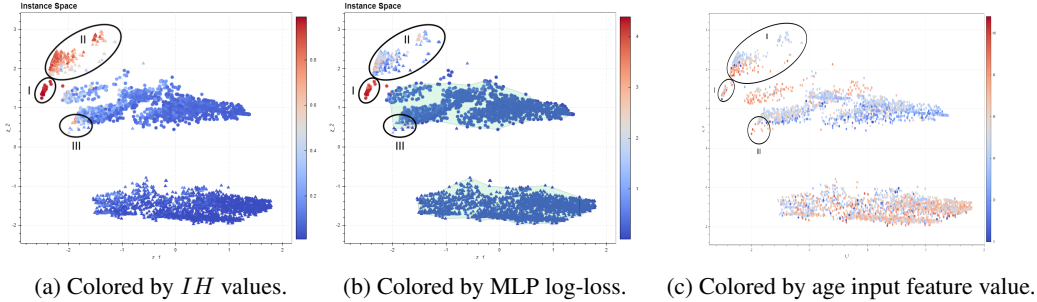


Figure 2: ISA for the hospitalization dataset.

COVID case. In Figure 2a, the observations are colored according to their IH values, with harder observations colored in red and easier observations colored in blue, hospitalized individuals plotted as triangles and non-hospitalized observations represented by circles. Most of the observations in this dataset are easy to classify correctly by the algorithm portfolio and the instance easiness footprint has a large area.

There is a large group of hospitalized citizens with an easier profile placed in the bottom of the IS. Non-hospitalized observations had mostly an intermediate hardness level, with low to medium IH . But some particular clusters of observations hard to classify, highlighted in circles numbered as I, II and III have caught our attention. Group I contains observations from the non-hospitalized class, whilst groups II and III have mostly instances from the hospitalized class. The observations from these groups have: a low likelihood of belonging to their classes (measured by CL meta-feature); are placed in disjuncts with elements which do not share their labels (measured by DCP); are close to elements from the opposite class (measured by $N1$); and have a high proportion of nearest neighbors with labels which differ from their own (measured by kDN). Overall, this indicates these groups probably contain noisy or outlier instances with input characteristics similar to that of other class. Plotting the values of the raw input attributes along the IS allowed us to observe some interesting features. Figure 2c shows one of them: age (red colors are attributed to elderly and youngsters are colored in blue), a well known feature with influences on COVID severity, impacting hospitalization. Indeed, the easiest cases of non-hospitalized citizens tend to be younger. But some observations from groups I and II have conflicting age patterns to what is commonly expected. Contradicting patterns are also observed for other input features along the observations from groups I, II and III. Inspecting the original records from these groups of citizens, we were able to confirm that group I has either individuals wrongly labeled as non-hospitalized and some atypical cases, such as people seeking for hospitalization too late. Group III has individuals with few and mild symptoms who were hospitalized and it is possible that their forms have missing information on some of the symptoms. Group I has also atypical cases which may contain missing information.

Concerning algorithmic performance, Figure 2b presents the ISA, colored according to the log-loss performance of the MLP classifier which attained the largest footprint area (shown within a green polygon). The MLP showed a good predictive performance for most of the instances except from those of the groups I and II. The PyHard tool allows to save and to inspect the characteristics of the instances from a selected footprint for gain further insights into an algorithm’s strengths and weaknesses, and the explanations for such based on the characteristics of these instances.

4 Conclusion

In this paper we have presented *PyHard*, a new analytical tool intended to address a gap in meta-learning studies about instance hardness for a single dataset, distributed publicly at PyPI³. The tool allows for the construction, visualization and interaction with a hardness embedding of a dataset, allowing to obtain useful insights to support data-centric analysis. Our case study demonstrates the utility of the tool for identifying anomalous and noisy observations within a dataset and to highlight observations for which a given classifier shows a consistently good predictive performance and where it may be biased.

³<https://pypi.org/project/pyhard/>

Acknowledgements

To the São José dos Campos municipal health secretariat and to the research agencies CAPES - Finance Code 001 (main grant 88887.507037/2020-00), CNPq (grant 307892/2020-4) and the Australian Research Council (grant FL140100012).

References

- Arruda, J. L., Prudêncio, R. B., and Lorena, A. C. (2020). Measuring instance hardness using data complexity measures. In *Brazilian Conference on Intelligent Systems*, pages 483–497. Springer.
- Muñoz, M. A., Villanova, L., Baatar, D., and Smith-Miles, K. (2018). Instance spaces for machine learning classification. *Machine Learning*, 107(1):109–147.
- Muñoz, M. A., Yan, T., Leal, M. R., Smith-Miles, K., Lorena, A. C., Pappa, G. L., and Rodrigues, R. M. (2021). An instance space analysis of regression problems. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(2):1–25.
- Prati, R. C. (2012). Combining feature ranking algorithms through rank aggregation. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Rice, J. R. (1976). The algorithm selection problem. In *Advances in Computers*, volume 15, pages 65 – 118. Elsevier.
- Smith, M. R., Martinez, T., and Giraud-Carrier, C. (2014). An instance level analysis of data complexity. *Machine Learning*, 95(2):225–256.
- Smith-Miles, K. A. (2009). Cross-Disciplinary Perspectives on Meta-Learning for Algorithm Selection. *ACM Computing Surveys*, 41(1).
- Vanschoren, J. (2019). Meta-learning. In *Automated Machine Learning*, pages 35–61. Springer.
- Vilalta, R. and Drissi, Y. (2002). A perspective view and survey of meta-learning. *Artificial Intelligence Review*, 18(2):77–95.