synthetic data for a better understanding of models and algorithms:

GANs for stress testing and other methods

carlos soares

csoares@fe.up.pt

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

LIACC

Fraunhofer
PORTUGAL

plan

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- current ML research vs understanding algorithm behavior
- the promise of metalearning
- ... but data is not enough
- the promise of (semi-)synthetic data
  - algorithms
  - models

what the goal of an (empirical) science of ML should be

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

4. Conclusions

The results reported in this paper show that the predictive accuracy of induced decision trees, both pruned and unpruned, is not sensitive to the goodness of split measure. This confirms Breiman et al.'s (1984) results. All of the methods tried are quite sophisticated and make good use of the information

instead…

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- yada yada yada
- my method is the best
- yada yada yada
- no clue why
- yada yada yada
- but I used a super-computer
- yada yada yada
- learned 10 gazillion parameters
- yada yada yada
- overfitting?!! nah…
- yada yada yada
- I'm so cool
- yada yada yada

	boring dataset 1	useless dataset 2
my method	**90.00%**	**92.00%**
method which was previously considered sota	89.00%	91.00%

is this one of those “kids these days” things?

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

4. Conclusions

The results reported in this pa
induced decision trees, both pru
goodness of split measure. This co
the methods tried are quite sophist

Machine Learning 3: 319–342, 1989

An Empirical Comparison of Selection Measures for Decision-Tree Induction

JOHN MINGERS (BSRCD@CU.WARWICK.AC.UK)
School of Industrial and Business Studies, University of Warwick, Coventry CV4 7AL, U.K.

Neuro-s
AutoML

3.2 Data sets

The experiments drew on four data sets, three from natural domains and one constructed artificially.

Profiles of B.A. Business Studies degree students (BABS). These data relate various attributes of each student, on entry to the course, to the final class of degree achieved. There are 186 observations with seven attributes – age (years), type of entry qualification (A-level,[3] BTEC Ordinary National Diploma, or some other), sex (male/female), number of O-levels, number of points at A-level (0–20), grade of maths O-level (A, B, C, FAIL), and full-time employment before the course (yes/no). There are four possible classes of degree – first, upper second, lower second, or third. Three of the attributes are integer and four symbolic. There is no known noise, but many other factors affecting the results have not been (and probably could not be) measured, giving high residual variation. This is an example of a prediction task.

The recurrence of breast cancer (Cancer). These data, containing 286 examples, are derived from those used in Bratko and Kononenko (1986) and concern the recurrence of breast cancer . There are two classes (recur or not recur) and nine attributes, of which four are integer. These include age, tumor size, number of nodes, malignant (yes/no), age of menopause (< 60, ≥ 60, not occurred), breast (left, right), radiation treatment (yes/no), and quadrant of breast (left, right, top, bottom, center). There are both missing data and residual variation. It is another example of a prediction task.

Classifying types of Iris (Iris). Kendall and Stewart (1976, p. 331) use these data as a test of discriminant analysis. There are 150 examples of three different varieties of Iris, with roughly equal numbers of each. The four integer attributes are measurements such as petal length and petal width, from which the examples can be classified. There is little noise or residual variation.

Recognizing LCD display digits (Digits). This is an artificial domain suggested by Breiman et al. (1984). A digit in a calculator display consists of seven lines, each of which may be on or off. Thus, there are ten classes (one for each digit) and seven binary-valued attributes (one for each line). Residual variation is introduced by assuming that a malfunction leads to a 10% chance of a line being incorrect. Such errors affect the attributes but not the class. Note that the chance of an example being completely correct is $0.9^7 = 0.48$. Three hundred cases were randomly generated. This is another example of a classification task.

repositories were created!

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

We currently maintain 160 data sets as a service to the machine learning community.

Machine learning, better, together

… and (empirical) ML research was better!

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

classifiers available today. We use **121 data sets**, which represent **the whole UCI** data base (excluding the large-scale problems) and other own real problems, in order to achieve significant conclusions about the classifier behavior, not dependent on the data set col-

Journal of Machine Learning Research 15 (2014) 3133-3181 Submitted 11/13; Revised 4/14; Published 10/14

Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?

Manuel Fernández-Delgado MANUEL.FERNANDEZ.DELGADO@USC.ES
Eva Cernadas EVA.CERNADAS@USC.ES
Senén Barro SENEN.BARRO@USC.ES
CITIUS: Centro de Investigación en Tecnoloxías da Información da USC
University of Santiago de Compostela
Campus Vida, 15872, Santiago de Compostela, Spain

Dinani Amorim DINANIAMORIM@GMAIL.COM
Departamento de Tecnologia e Ciências Sociais- DTCS
Universidade do Estado da Bahia
Av. Edgard Chastinet S/N - São Geraldo - Juazeiro-BA, CEP: 48.305-680, Brasil

or maybe not...

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

classifiers available today. We use **121 data sets**, which represent **the whole UCI** data base (excluding the large-scale problems) and other own real problems, in order to achieve significant conclusions about the classifier behavior, not dependent on the data set collection. **The classifiers most likely to be the bests are the random forest** (RF) versions, the best of which (implemented in R and accessed via caret) achieves 94.1% of the maximum accuracy overcoming 90% in the 84.3% of the data sets. However, the dif-

Journal of Machine Learning Research 15 (2014) 3133-3181 Submitted 11/13; Revised 4/14; Published 10/14

Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?

Manuel Fernández-Delgado MANUEL.FERNANDEZ.DELGADO@USC.ES
Eva Cernadas EVA.CERNADAS@USC.ES
Senén Barro SENEN.BARRO@USC.ES
CITIUS: Centro de Investigación en Tecnoloxías da Información da USC
University of Santiago de Compostela
Campus Vida, 15872, Santiago de Compostela, Spain

Dinani Amorim DINANIAMORIM@GMAIL.COM
Departamento de Tecnologia e Ciências Sociais- DTCS
Universidade do Estado da Bahia
Av. Edgard Chastinet S/N - São Geraldo - Juazeiro-BA, CEP: 48.305-680, Brasil

… and they can be overfitted

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

4.1 Are Commercial Tool Developers Overfitting the UCI?

In Section 3.7 we observed that C5.0 rules, C5.0 tree, MLP and RBFN seem to be overfitting the UCI-R. However, the Kolmogorov-Smirnov did not detect any

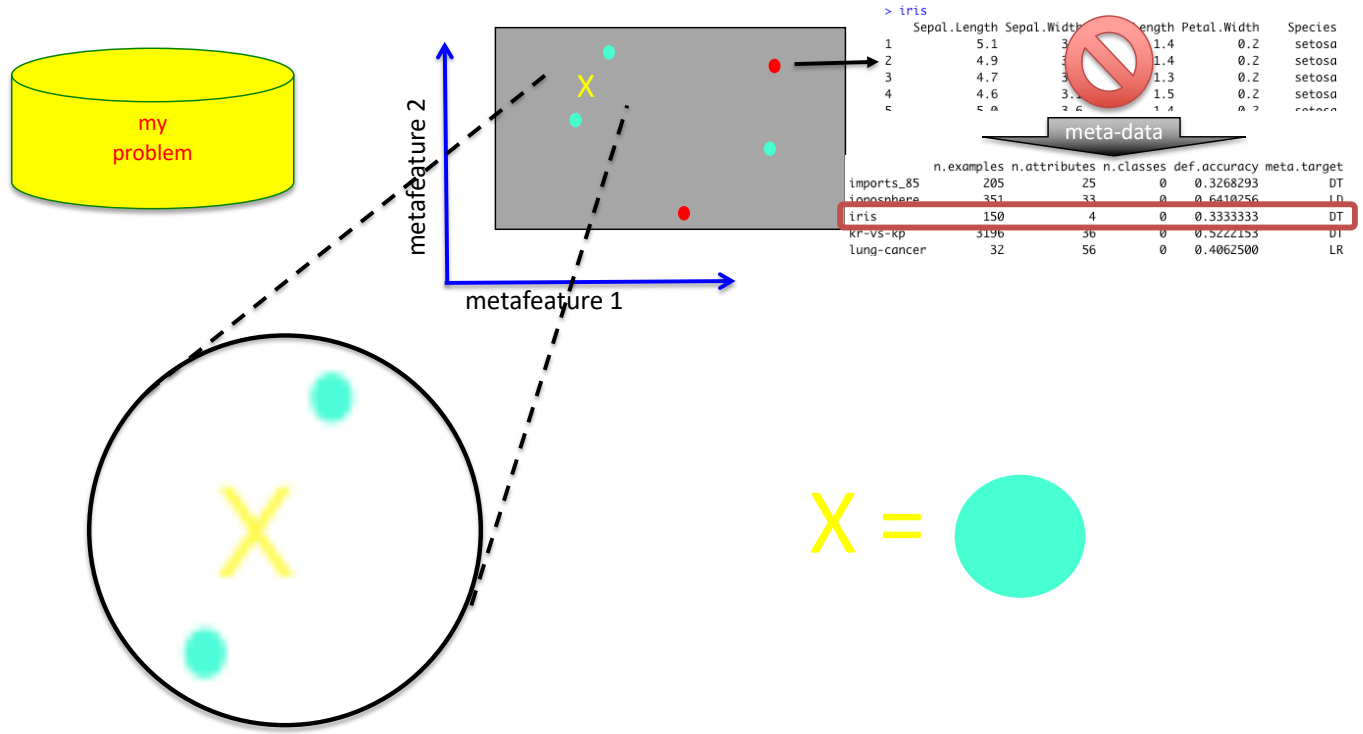
Soares, C. (2003). Is the UCI repository useful for data mining? In F. M. Pires & S. Abreu (Eds.), *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 2902, pp. 209–223). Springer-Verlag. https://doi.org/10.1007/978-3-540-24580-3_28

and then there was metalearning

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

metalearning for algorithm selection (2/2)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



Brazdil P, Soares C, Costa J. Ranking Learning Algorithms: Using {IBL} and Meta-Learning on Accuracy and Time Results. Machine Learning. 2003;50(3):251-277.

… and it promised to deliver metaknowledge!

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Machine Learning, 54, 275–312, 2004

On Data and Algorithms: Understanding Inductive Performance

ALEXANDROS KALOUSIS kalousis@cui.unige.ch
University of Geneva, Computer Science Department, 24, rue du General Dufour, CH-1211 Geneva 4, Switzerland

JOÃO GAMA jgama@liacc.up.pt
LIACC, FEP—University of Porto, Rua Campo Alegre 823, 4150 Porto, Portugal

MELANIE HILARIO hilario@cui.unige.ch
University of Geneva, Computer Science Department, 24, rue du General Dufour, CH-1211 Geneva 4, Switzerland

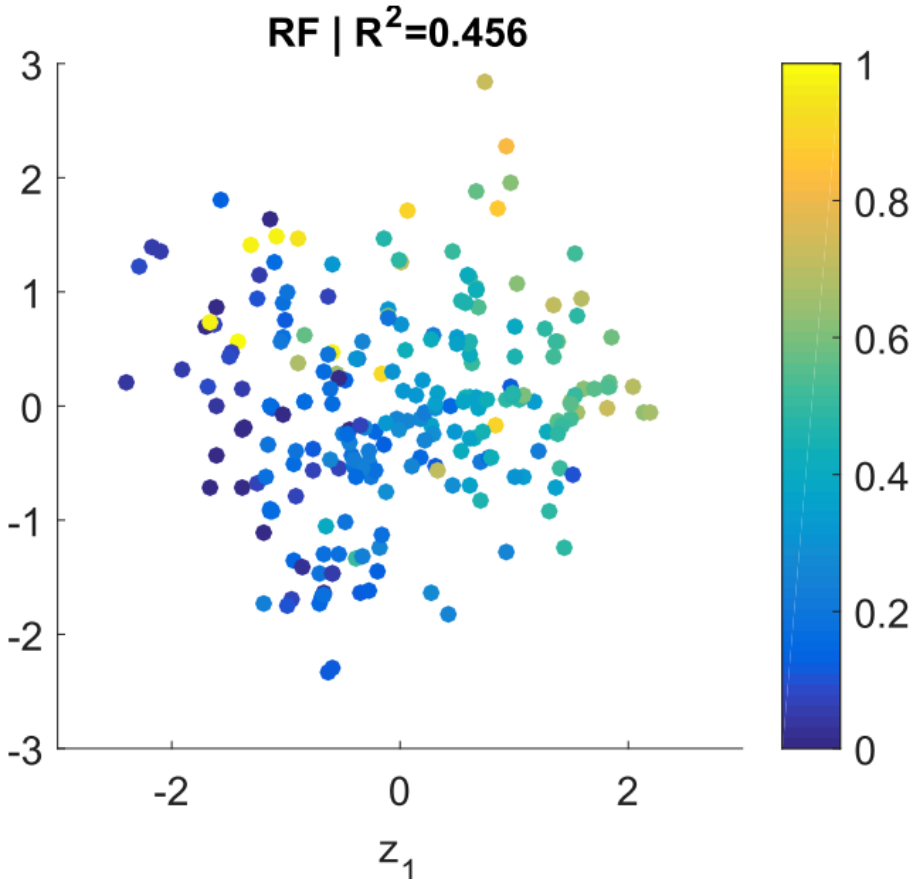
For the High Error Correlation group the Class Entropy is strongly peaked and concentrated to the low values of the scale compared to a more uniform distribution within the Low Error Correlation group. The low values of *CE* can be due to two factors, large

… meaning,

it doesn't matter which algorithm you use

when class entropy is low

… in different formats



Muñoz, M. A., Villanova, L., Baatar, D., & Smith-Miles, K. (2018). Instance spaces for machine learning classification. *Machine Learning*, *107*(1), 109–147. https://doi.org/10.1007/s10994-017-5629-5

… but there weren't enough datasets!

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

classifiers available today. We use **121 data sets**, which represent **the whole UCI** data base (excluding the large-scale problems) and other own real problems, in order to achieve significant conclusions about the classifier behavior, not dependent on the data set col-

maybe we can use artificial data?

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Mach Learn (2018) 107:109–147
https://doi.org/10.1007/s10994-017-5629-5

Instance spaces for machine learning classification

Mario A. Muñoz[1] · **Laura Villanova**[1] · **Davaatseren Baatar**[1] · **Kate Smith-Miles**[1]

To generate instances with a desired target vector of features, $\mathbf{f}_T$, we tune a Gaussian Mixture Model (GMM) until the Mean Squared Error (MSE) between $\mathbf{f}_T$ and the feature vector of a sample from the GMM, $\mathbf{f}_S$, is zero, assuming that the GMM is sampled using a fixed seed to guarantee some level of repeatability. Let us define

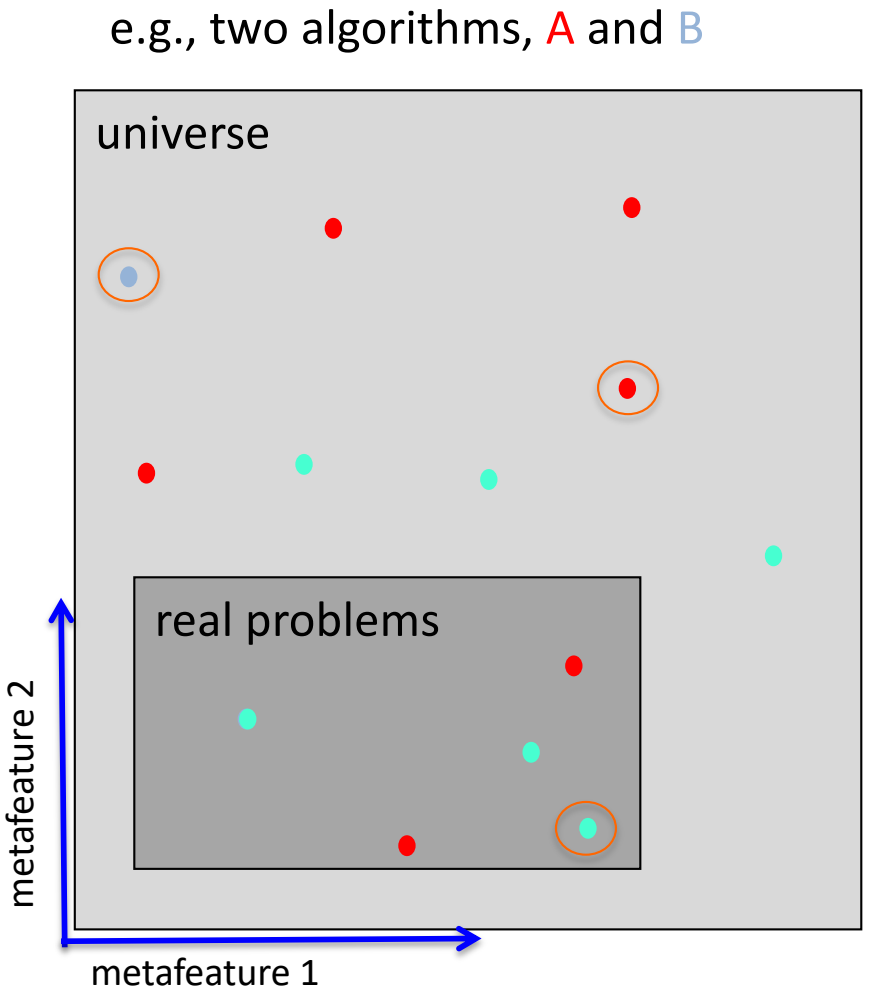
… meaning,

we learn a multivariate distribution that generates datasets with the desired metafeature values

or maybe not

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

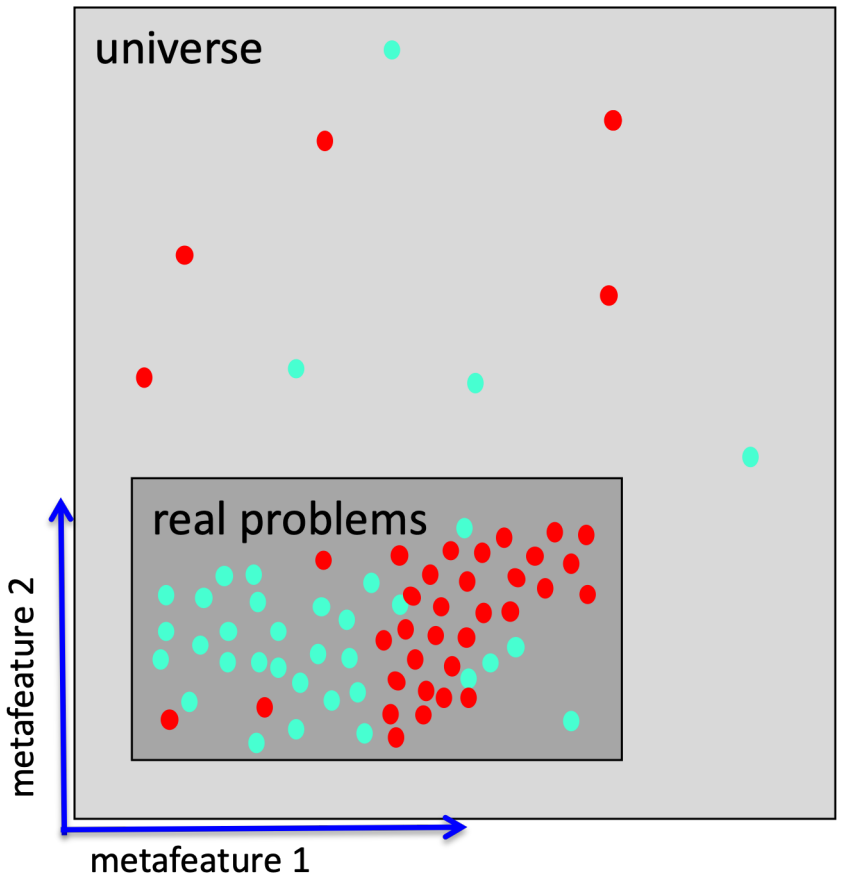
random datasets are
probably not very realistic



the promise of semi-synthetic data

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

e.g., two algorithms, A and B



mom, are we there yet?

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- current ML research vs understanding algorithm behavior
- the promise of metalearning
- ... but data is not enough
- the promise of (semi-)synthetic data
  - algorithms
    - data manipulation
    - dataset morphing
    - datasetoids
  - models

maybe we can manipulate existing data?

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Information Sciences 261 (2014) 237–262

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Towards UCI+: A mindful repository design

Núria Macià *, Ester Bernadó-Mansilla

Grup de Recerca en Sistemes Intel·ligents, La Salle – Universitat Ramon Llull, 08022 Barcelona, Spain

5.3. Comparison with an artificial sample

In order to verify whether the UCI repository could be enhanced with an artificial sample of problems, we used a set of 80,000 artificial data sets, whose details are explained below.

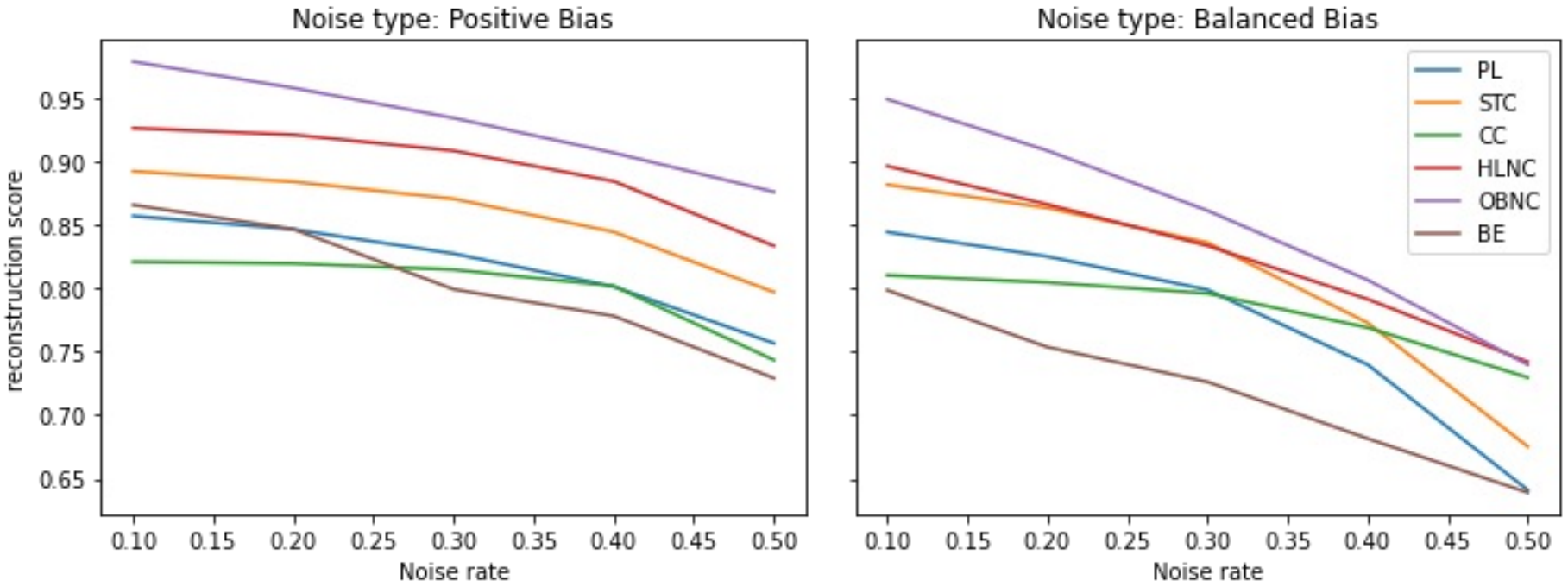
These data sets were designed to analyse whether they could cover the regions of the complexity measurement space where the UCI has not any representative. The extrinsic characteristics can be easily controlled in an artificial sample, as well as the introduction of noise or missing values. What represents a major challenge is to obtain a collection of data sets whose target concepts are spread across a wide range of geometrical complexities. For this purpose, we relied on the aforementioned data complexity measures and designed an algorithm which was able to synthetically generate data satisfying a given constraint of complexity. This was achieved by means of an evolutionary multi-objective algorithm—Non-dominated Sorting Genetic Algorithm II (NSGA-II) [6]—, which was configured to simultaneously optimise a set of complexity measures, such as the maximisation of N1 and the minimisation of F1. The synthesising approach finds the vector $\mathbf{t} = [\mathbf{z}_1, \ldots, \mathbf{z}_k]^T$, $2 \leqslant k \leqslant n$, where n is the total number of instances and $\mathbf{z}_i$ is instance i, ($\mathbf{t}$ is a sub-set of $[\mathbf{z}_1, \ldots, \mathbf{z}_n]^T$), which optimises (minimises or maximises) $f(\mathbf{t}) = [F1v(\mathbf{t}), F1(\mathbf{t}), F2(\mathbf{t}), F3(\mathbf{t}), F4(\mathbf{t}), L1(\mathbf{t}), L2(\mathbf{t}), L3(\mathbf{t}), N1(\mathbf{t}), N2(\mathbf{t}), N3(\mathbf{t}), N4(\mathbf{t}), T1(\mathbf{t}), T2(\mathbf{t})]$ and subject to the following constraints: (1) $k \geqslant k_{min}$, where k_{min} is the minimum number of instances specified by the user, (2) class imbalance specified by the user, and (3) no duplicate instances.

… meaning,

we sample from an existing dataset to generate datasets with the desired metafeature values

illustrative example: stress testing label correction methods

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



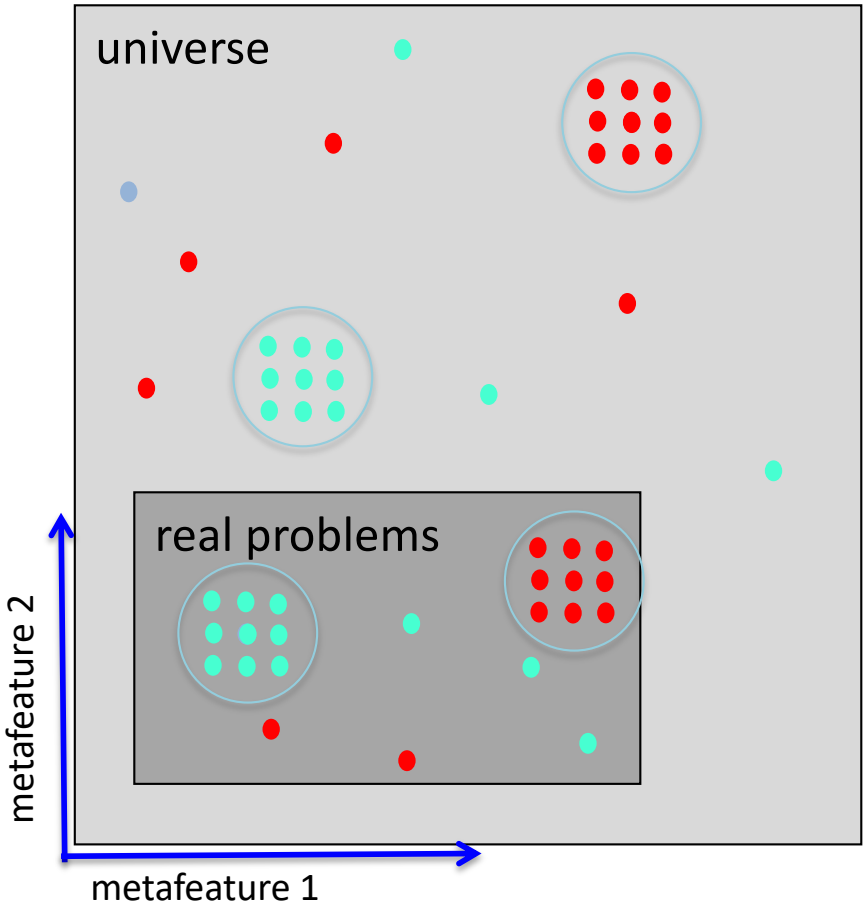
I. Oliveira E Silva, C. Soares, I. Sousa and R. Ghani (2023), Systematic analysis of the impact of label noise correction on ML Fairness, *accepted for publication at IJCAI 2023*

the promise of semi-synthetic data: manipulation of real data

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- removing
  - e.g. sampling
- adding
  - e.g. new random attributes
- changing
  - e.g. adding noise
- stress testing algorithms

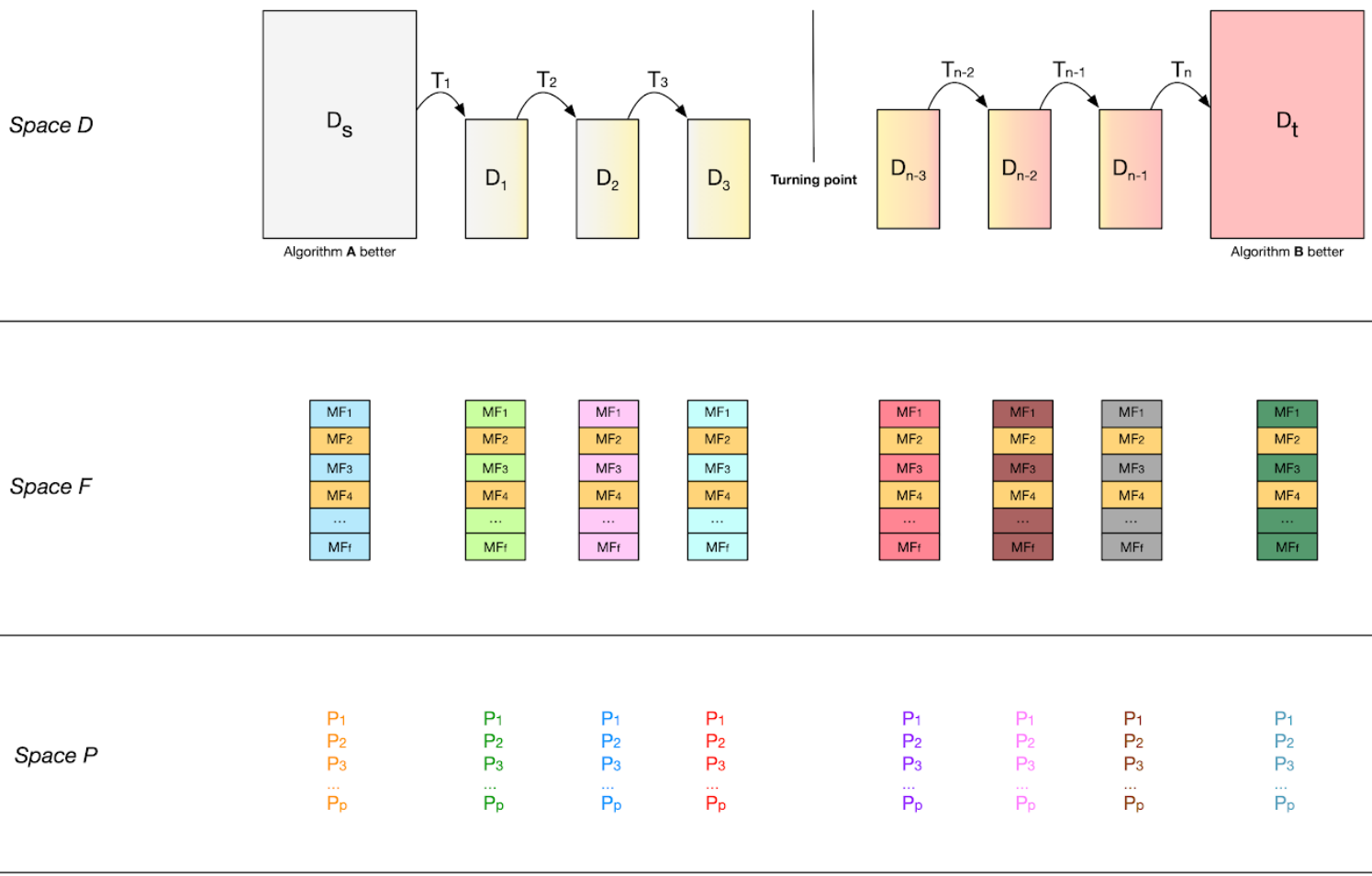
e.g., two algorithms, A and B



Hilario, Kalousis. Quantifying the Resilience of Inductive Classification Algorithms. In Proc. 4th European Conf. on Principles of Data Mining and Knowledge Discovery. Springer-Verlag; 2000:106-115.

dataset morphing to understand algorithm behavior

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



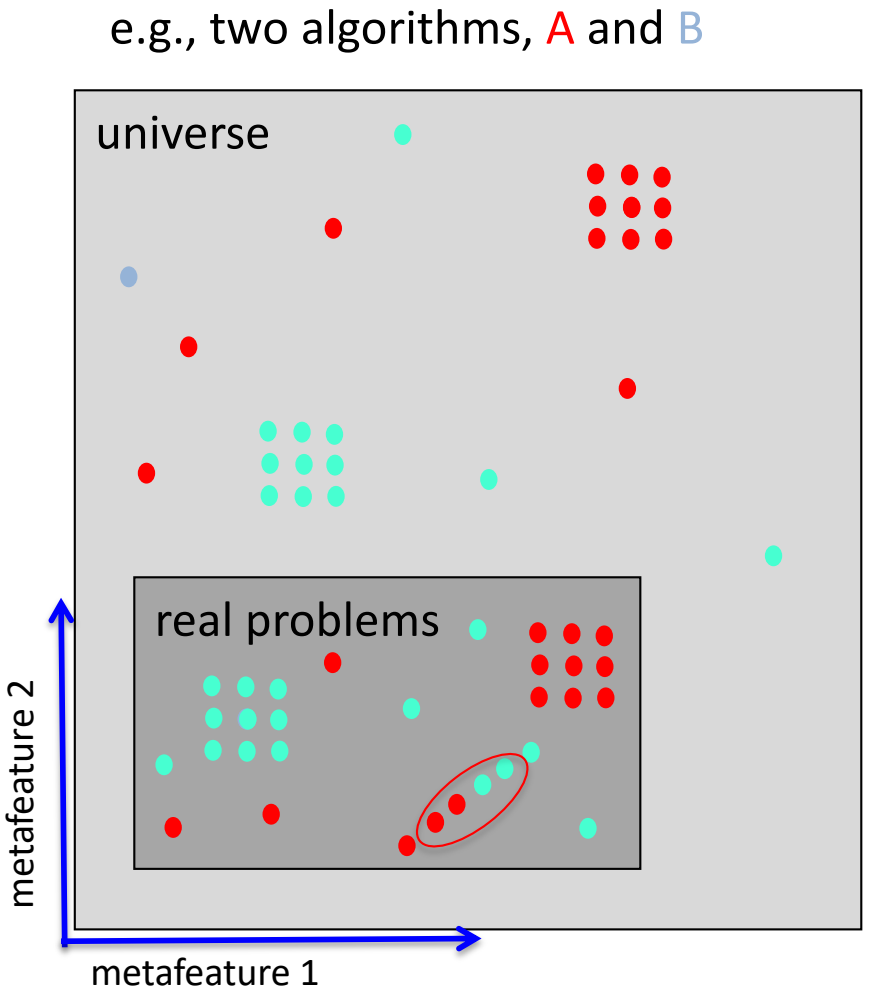
Correia, A., Soares, C., & Jorge, A. (2019). Dataset Morphing to Analyze the Performance of Collaborative Filtering. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 11828 LNAI* (pp. 29–39). https://doi.org/10.1007/978-3-030-33778-0_3

the promise of semi-synthetic data: dataset morphing

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

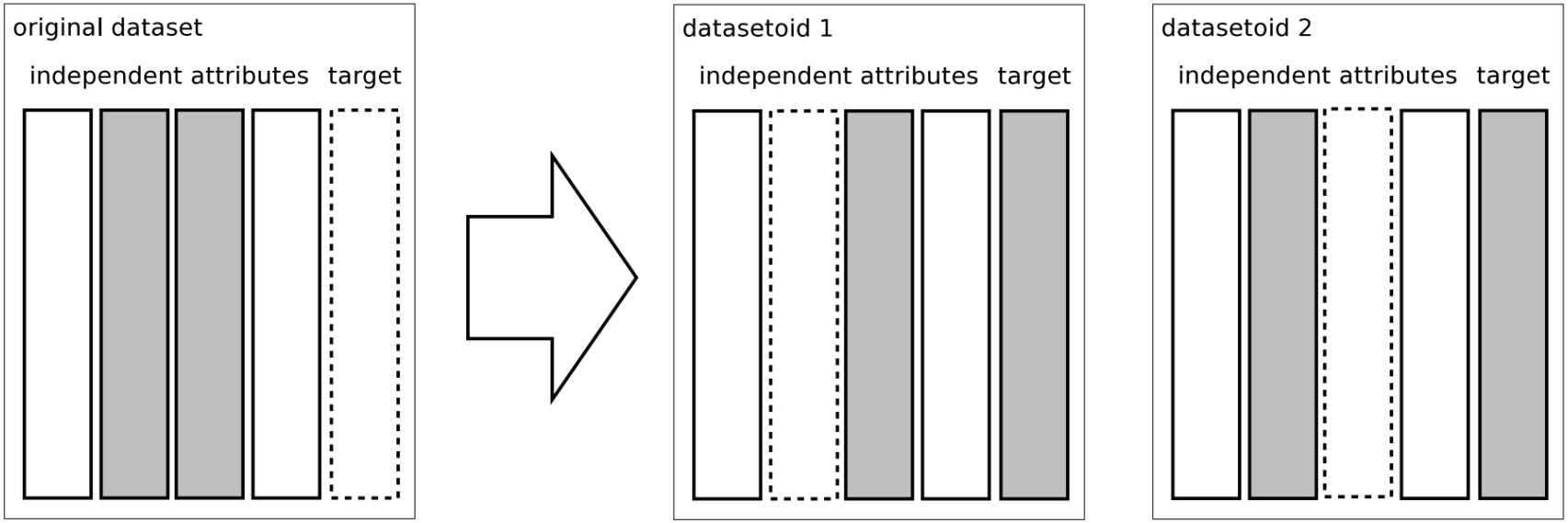
- gradually transforming one dataset into another one

- comparison between two algorithms
- ... or when an algorithm performs very well and very badly



datasetoids to understand algorithm behavior

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

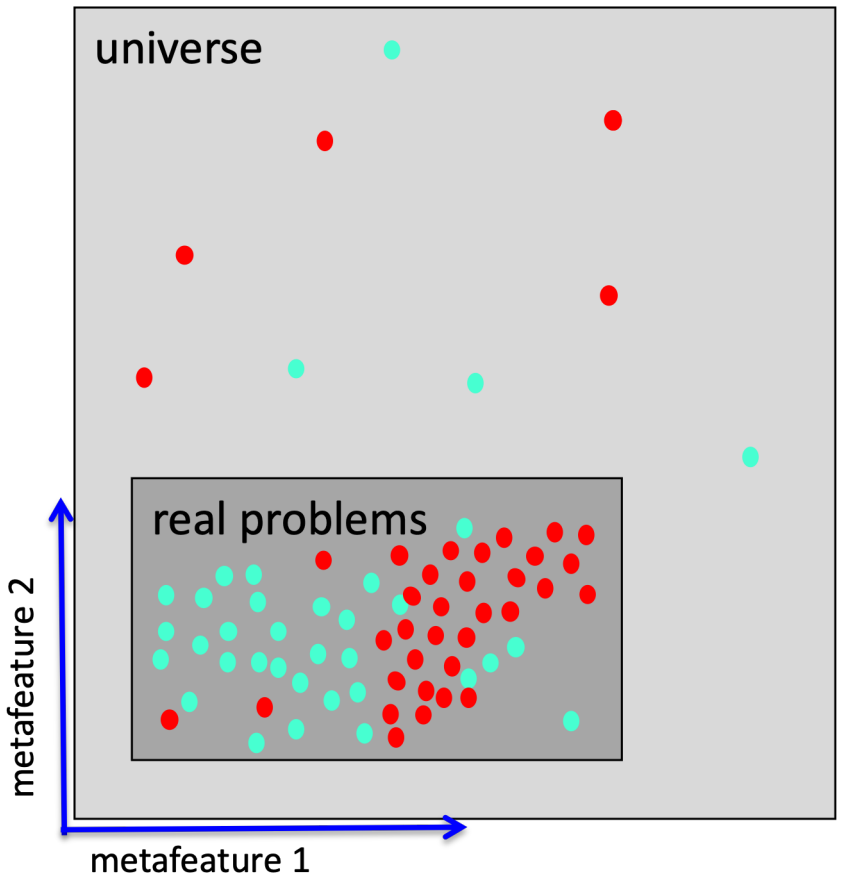


Soares, C. (2009). UCI++: Improved Support for Algorithm Selection Using Datasetoids. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 5476 LNAI* (pp. 499–506). Springer-Verlag. https://doi.org/10.1007/978-3-642-01307-2_46

the promise of semi-synthetic data: datasetoids (hopefully…)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

e.g., two algorithms, A and B



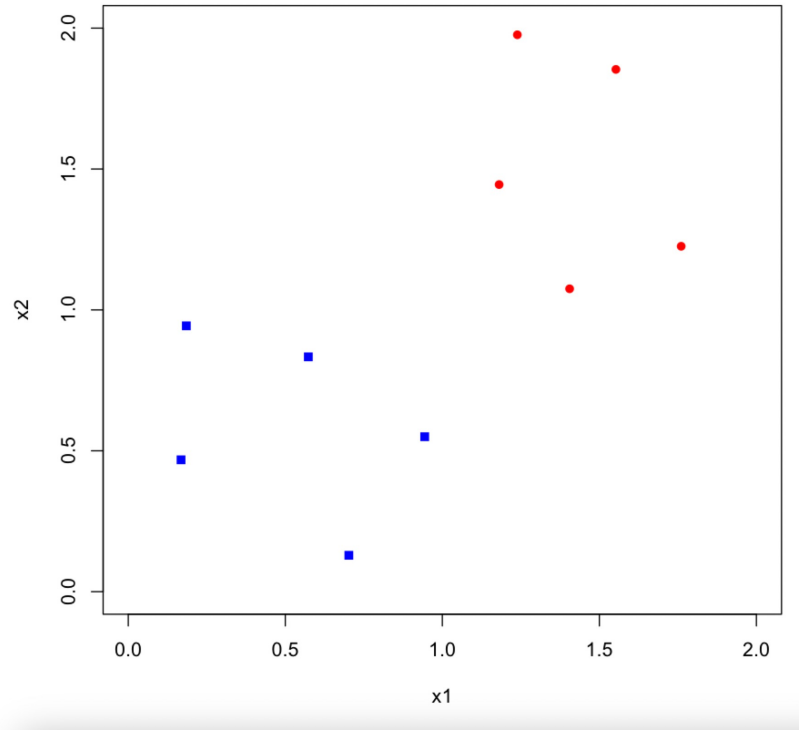
mom, are we there yet?

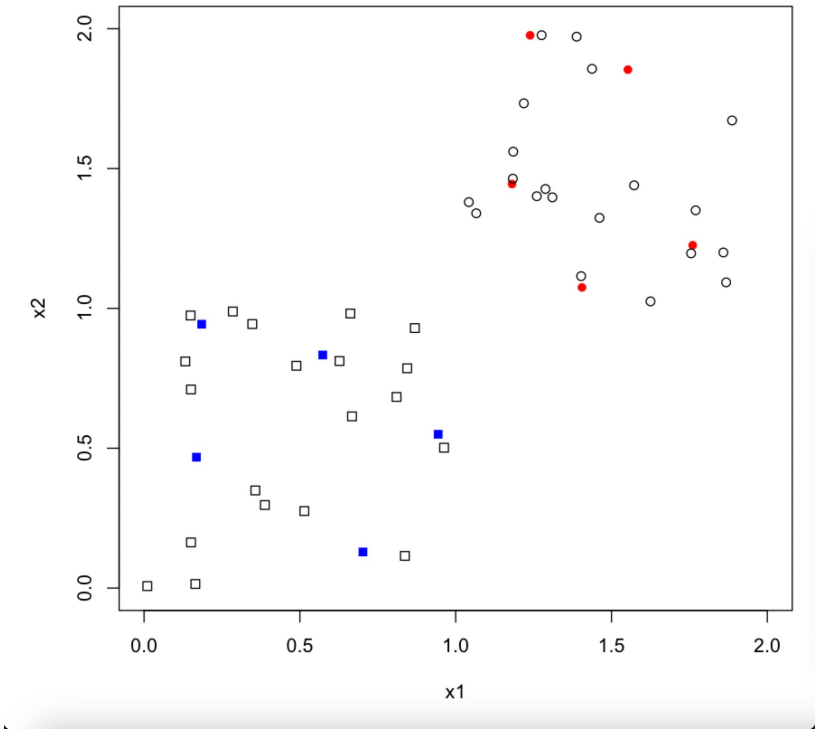
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- current ML research vs understanding algorithm behavior
- the promise of metalearning
- … but data is not enough
- the promise of (semi-)synthetic data
  - algorithms
  - models
    - GASTEN

semi-synthetic data for better models

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO





realistic data for better models: 1 vs 7 in MNIST

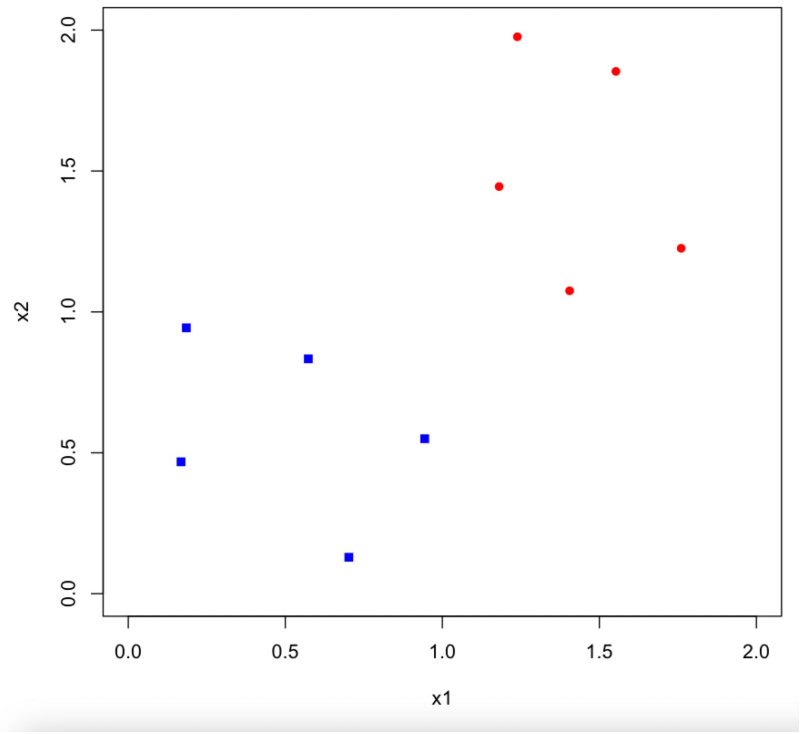
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

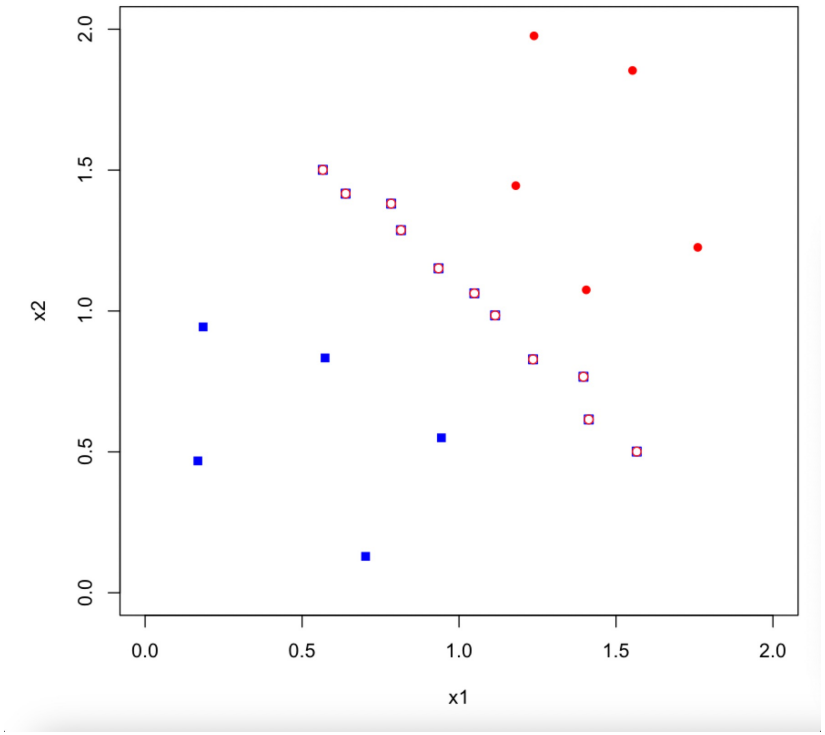
original data

generated data

semi-synthetic data for stress testing models

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

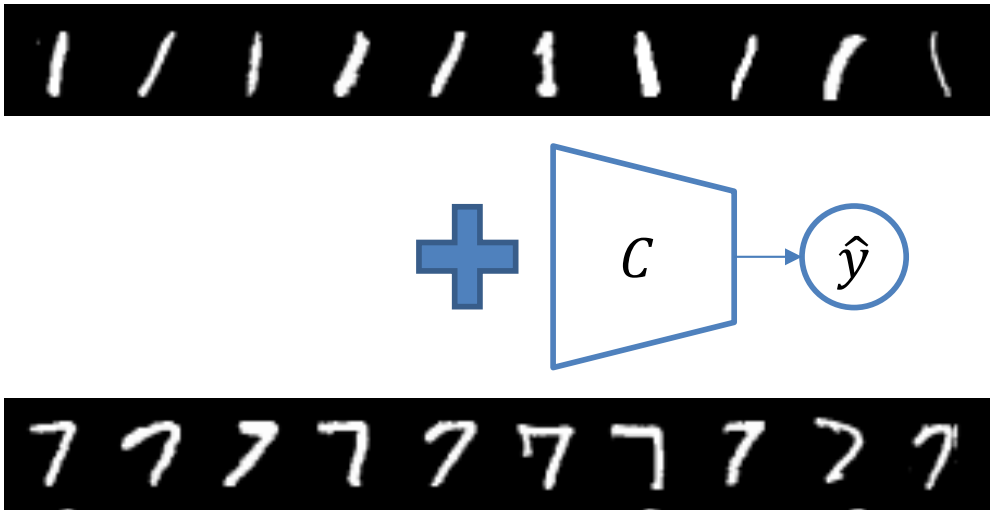




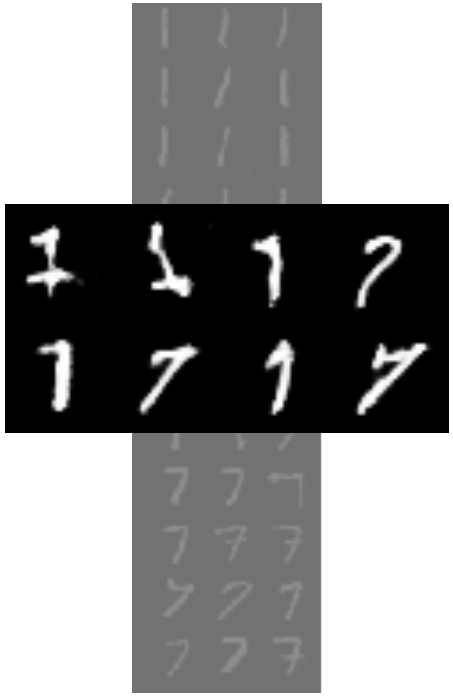
Realistic data for stress testing models: 1 vs 7 in MNIST

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

original data



generated data

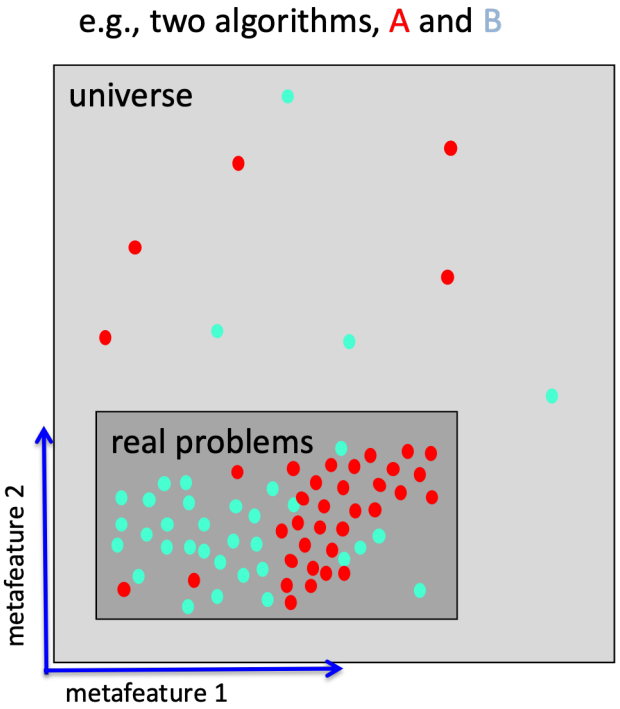


Luís Cunha, Carlos Soares, André Restivo, and Luís F. Teixeira. 2023. GASTeN: Generative Adversarial Stress Test Networks. In Advances in Intelligent Data Analysis XXI: 21st International Symposium on Intelligent Data Analysis, IDA 2023, Louvain-la-Neuve, Belgium, April 12–14, 2023, Proceedings. Springer-Verlag, Berlin, Heidelberg, 91–102. https://doi.org/10.1007/978-3-031-30047-9_8

conclusions

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- semi-synthetic data is the way towards a better understanding of models and algorithms
- … towards an empirical science of ML



- promising approaches
  - data manipulation
  - dataset morphing
  - datasetoids
  - GANs & friends

current work on stress testing algorithms

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- label correction methods
  - with I. Oliveira e Silva, I. Sousa, R. Ghani
- time series forecasting and anomaly detection
  - with R. Andrade, N. Vasconcelos, Y. Baghoussi, V. Cerqueira, J. Mendes-Moreira
- hierarchical time series forecasting
  - with L. Roque, L. Torgo

current work on dataset morphing

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- time series forecasting
  - with M. Santos, A. Carvalho
- recommender systems
  - with A. Correia and A. Jorge

current work on datasetoids

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- classification
  - with G. Freire
- regression
  - with L. Viegas, G. Freire
- graphs
  - with R. Andrade, P. Ribeiro

current work on stress testing models

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- CV
  - with L. Cunha, I. Gomes, L.F. Teixeira, A. Restivo
- NLP
  - with D. Prêda, I. Gomes, H.L. Cardoso
- time series forecasting
  - with A. Monteiro, V. Ribeiro, Y. Baghoussi, V. Cerqueira, A.P. Serra
- hierarchical time series forecasting
  - with L. Roque, L. Torgo

funding

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Center for
Responsible AI

AISYM4MED

PRR
Plano de Recuperação
e Resiliência

REPÚBLICA
PORTUGUESA

Financiado pela
União Europeia
NextGenerationEU

Funded by the European Union under Grant Agreement N.210810650. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.

AISYM4MED

UK Research
and Innovation

Project funded by

https://centerforresponsible.ai/

https://aisym4med.eu/