U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Artificial Intelligence and Society

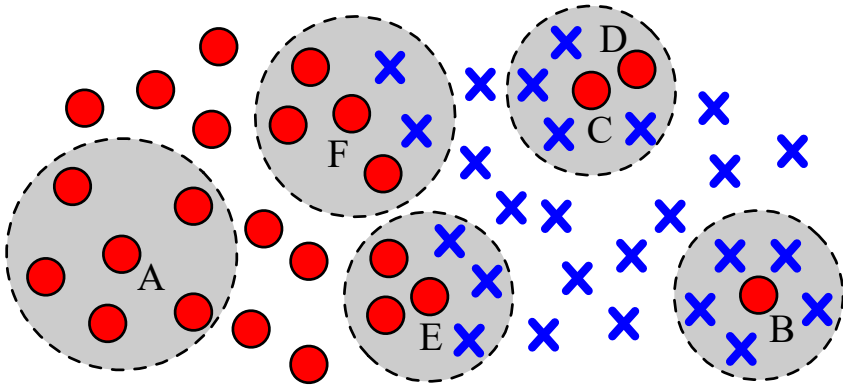
Module 02: Data Complexity & Meta-Learning

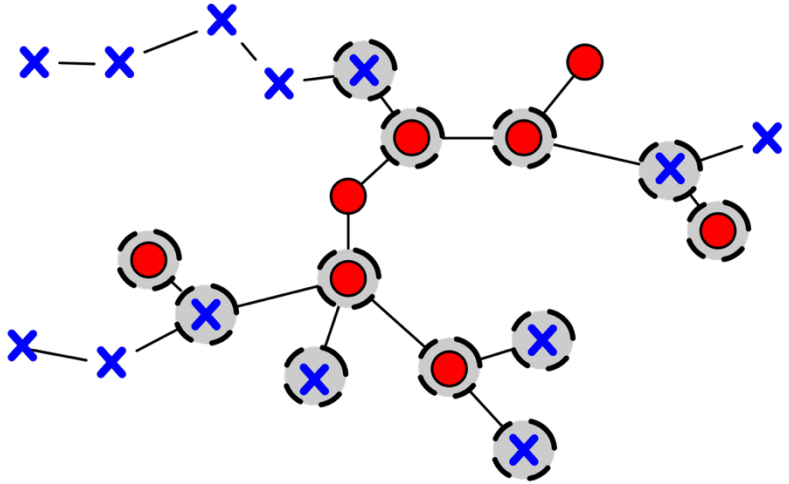
Miriam Seoane Santos
LIAAD, INESC TEC, FCUP, University of Porto
miriam.santos@fc.up.pt

Previously…

Other Data Intrinsic Characteristics

- Studies along this line discuss the estimation of the inherent complexity of the dataset, namely through the quantification of **borderline examples** and **instance hardness** measures.



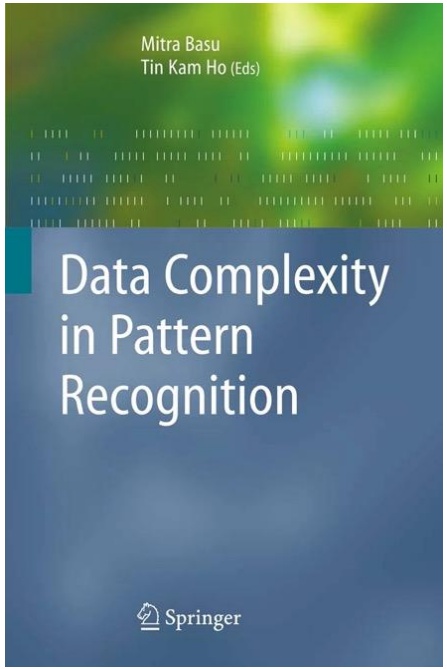


Data Complexity

Characterizing data complexity and classification behaviour

Data Complexity

- "When classifiers are not perfect, is it a deficiency of the algorithms by design, or is it a difficulty intrinsic to the classification task?" *(Ho and Basu, 2006)*



Measures of Geometrical Complexity in Classification Problems

Tin Kam Ho, Mitra Basu, and Martin Hiu Chung Law

Summary. When popular classifiers fail to perform to perfect accuracy in a practical application, possible causes can be deficiencies in the algorithms, intrinsic difficulties in the data, and a mismatch between methods and problems. We propose to address this mystery by developing measures of geometrical and topological characteristics of point sets in high-dimensional spaces. Such measures provide a basis for analyzing classifier behavior beyond estimates of error rates. We discuss several measures useful for this characterization, and their utility in analyzing data sets with known or controlled complexity. Our observations confirm their effectiveness and suggest several future directions.

Data Complexity

- Given data from a new problem, can we determine whether there exists a clean decision boundary between the classes?
- Are the classes intrinsically distinguishable?
- To what extent can this boundary be inferred by the learning algorithms?
- Which classifiers can do the best job?

These questions are about the intrinsic complexity of a classification problem, and the match of a classifier's capability to a problem's intrinsic complexity.

- Factors affecting performance can be:
  - The **shape** of the classes and thus the shape of the decision boundary
  - The amount of **overlap** bewteen the classes
  - The **proximity** of two classes
  - The number of **informative** samples available for training
  - (...)

Data Complexity: Learning Paradigms and Classifier Footprints

- Several real-world applications suffer from distinct (and often combined) data irregularities.
- Classifiers respond to different complexity factors **in their unique ways**:

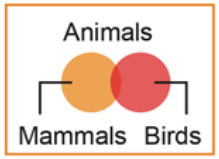
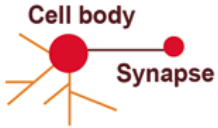
Table 1
Examples of data irregularities in real-world applications.

Scenario	Type of data irregularity
Credit card fraud detection	Class imbalances, class skew [18]
Breast cancer diagnosis	Class imbalance, class skew, small disjuncts [19,20]
Market segmentation	Class imbalance, class skew [21]
Facial and emotion recognition	Small disjuncts [22]
Survey data	Unstructured missingness [23]
Phylogeny problem	Unstructured missingness [24]
Gene expression data	Unstructured missingness [25]
Visual object recognition	Structural missingness or absent features [17]
Software effort prediction	Unstructured and structural missingness [26]

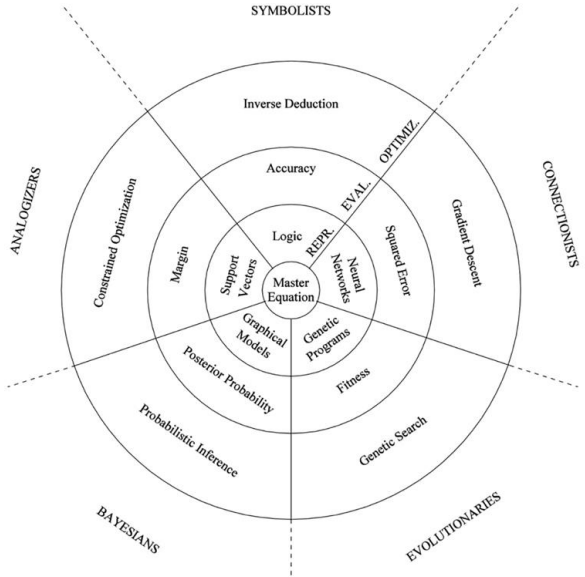
- Max-margin Classifiers – sensitive to class imbalance, small disjuncts, class distribution skew, absent features, missing features.
- Neural Networks – sensitive to class imbalance, small disjuncts, absent features, missing features.
- *k*-Nearest Neighbours (*k*-NN) – sensitive to class imbalance, small disjuncts, absent features, missing features; immune to class distribution skew as it does not make any assumptions regarding the class-conditional distributions.
- Bayesian Inference – sensitive to class imbalance, small disjuncts, class distribution skew, absent features, missing features.
- Decision Trees – sensitive to class imbalance, small disjuncts, class distribution skew; inherently immune to feature missingness as branching is based only on the observed features.

Data Complexity: Learning Paradigms and Classifier Footprints

- According to Pedro Domingos, there are 5 tribes in Machine Learning, looking into different understanding of logic and reasoning and methodology: *Symbolists, Bayesians, Connectionists, Evolutionaires, and Analogizers.* **Each of them has their strengths and weaknesses.**

Symbolists	Bayesians	Connectionists	Evolutionaries	Analogizers
				
Use symbols, rules, and logic to represent knowledge and draw logical inference	Assess the likelihood of occurrence for probabilistic inference	Recognize and generalize patterns dynamically with matrices of probabilistic, weighted neurons	Generate variations and then assess the fitness of each for a given purpose	Optimize a function in light of constraints ("going as high as you can while staying on the road")
Favored algorithm Rules and decision trees	**Favored algorithm** Naive Bayes or Markov	**Favored algorithm** Neural networks	**Favored algorithm** Genetic programs	**Favored algorithm** Support vectors

Source: Pedro Domingos, *The Master Algorithm*, 2015



Data Complexity Measures

- Various aspects of data complexity can affect the behaviour of different families in different ways.
- A prerequisite for setting **proper expectations on classification performance** is to understand complexity of a specific data set arising from an application.
- To understand data complexity is to find out **whether, or to what extent, patterns exist in the data.**
- **It is also to obtain guidance on selecting specific classification techniques** (or transformation, or resampling, or cleaning, or the devise of specialized strategies)!
- ***Enters the field of Meta-Learning…***

Data Complexity: Sources of Difficulty in Classification

- What makes classification difficult? Traditionally:

(1) Class Ambiguity

(2) Boundary Complexity

(3) Sample Sparsity and Feature Space Dimensionality

Data Complexity: Class Ambiguity

- Cases in a classification problem cannot be distinguished using the given features by *any* classification algorithm. This is often a consequence of the problem formulation and can either arise due to **poorly defined class concepts (instrinsically inseparable)** or to **insufficient features** to distinguish them.

(a) The shapes of the lower case letter "l" and the numeral "1" are the same in many fonts. They cannot be distinguished by shape alone. Which class a sample belongs to depends on context.

(b) There may be sufficient features for classifying the shells by shape, but not for classifying by the time of the day when they were collected, or by which hand they were picked up.

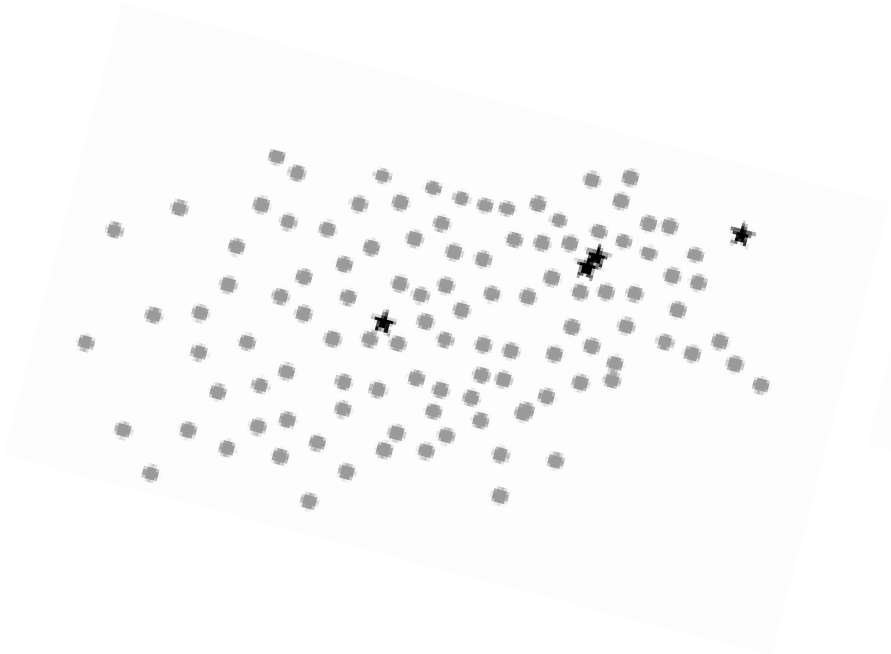
Data Complexity: Boundary Complexity

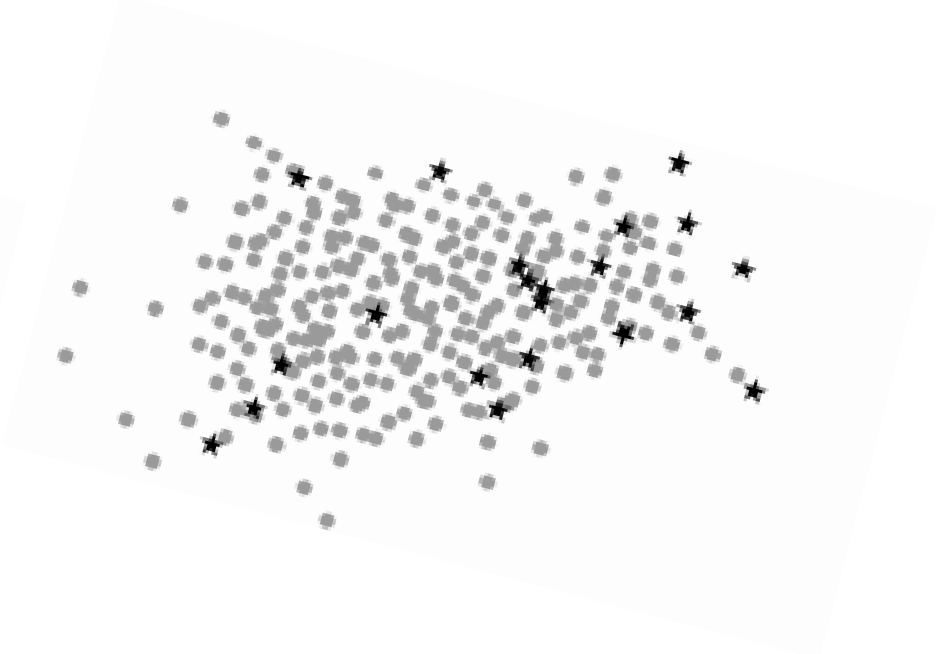
- A class boundary is complex if it requires a long description, possibly indicating a list of all the points together with their class labels.
- Associated with the concept of *geometrical complexity* of a dataset (most classifiers are characterized by geometrical descriptions of their decision regions).
- **Related to the *class overlap* problem.**

Fig. 1.2. Classification problems of different geometrical complexity: (a) linearly separable problem with wide margins and compact classes; (b) linearly separable problem with narrow margins and extended classes; (c) problem with nonlinear class boundary; (d) heavily interleaved classes following a checker board layout.

Data Complexity: Sparsity and Dimensionality

- Incomplete or sparse samples hinder the generalization of classifiers. This is especially difficult in high-dimensional spaces.
- **Related to the problem of Lack of Data or Lack of Density.**





Data Complexity Measures

- In practical applications, often a problem becomes difficult because of a **mixture of boundary complexity and sample sparsity effects.**
- **Data Complexity Measures started being organized into groups or categories:**

Ho and Basu *(2002)*

(1) Overlap of Individual Feature Values
(2) Separability of Classes
(3) Geometry, Topology, Density of Manifolds

Sotoca *et al.* *(2005)*

(1) Overlap
(2) Class Separability
(3) Geometry and Density

Lorena *et al.* *(2019)*

(1) Feature-Based Measures
(2) Linearity Measures
(3) Neighbourhood Measures
(4) Network Measures
(5) Dimensionality Measures
(6) Class Imbalance Measures

Data Complexity Measures: Feature-based measures

- **Feature-based measures:** Characterise how informative the available features are to separate the classes.

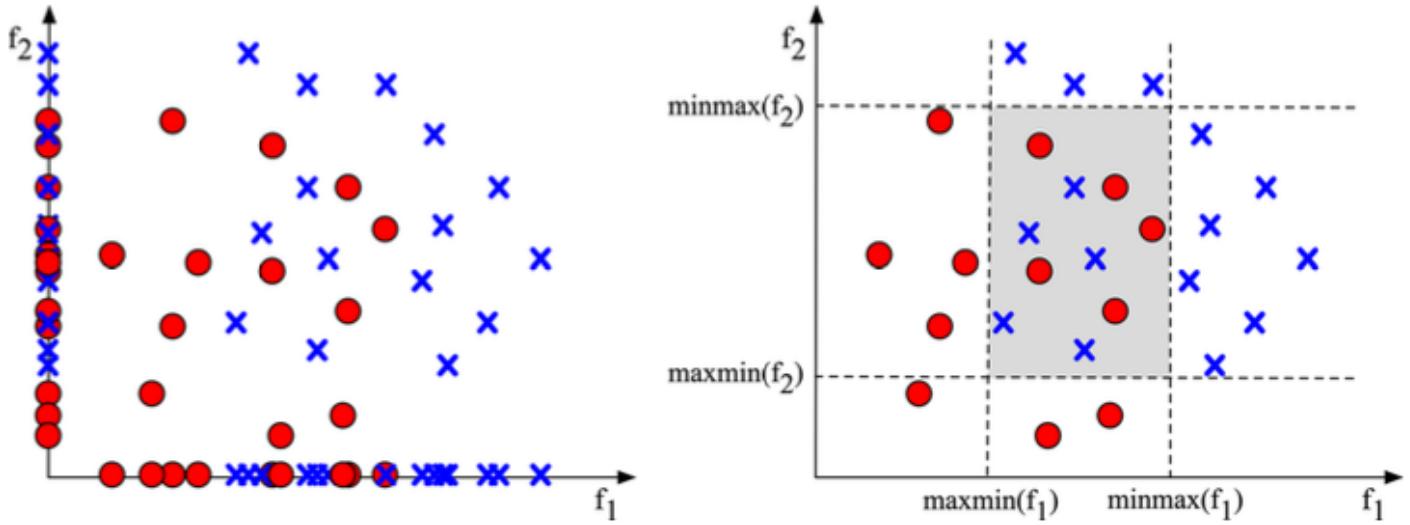
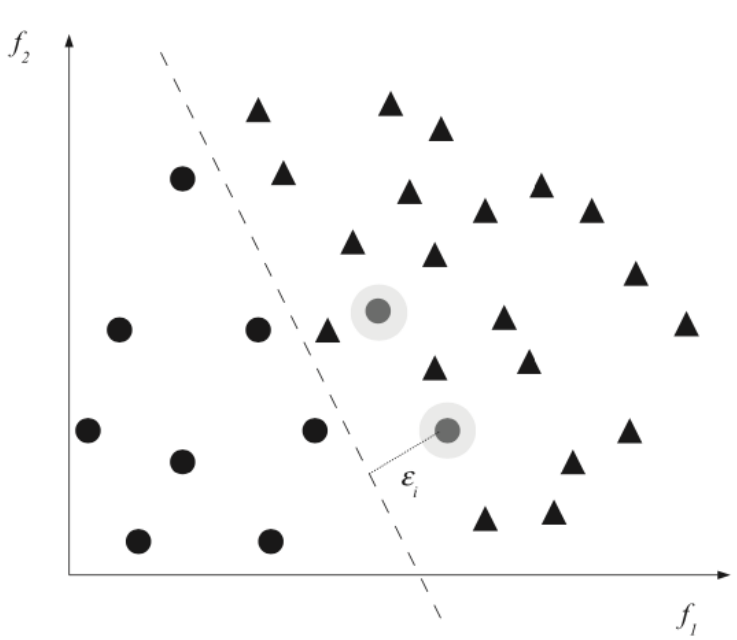


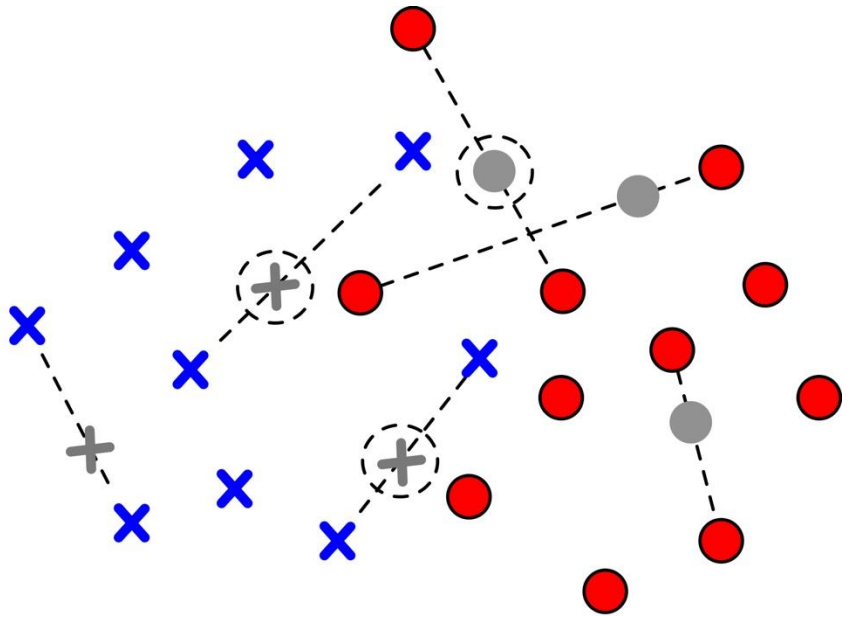
Fig. 5 Representations of F1 (leftside) and F2 (rightside) measures for the same dataset. Note how F1 projects data onto the axis to establish the amount of overlap, where f_1 is the feature with highest discriminative power, i.e., lowest overlap. In turn, F2 considers both features to define a region where classes coexist

Data Complexity Measures: Linearity measures

- **Linearity measures:** Quantify whether classes can be linearly separated.



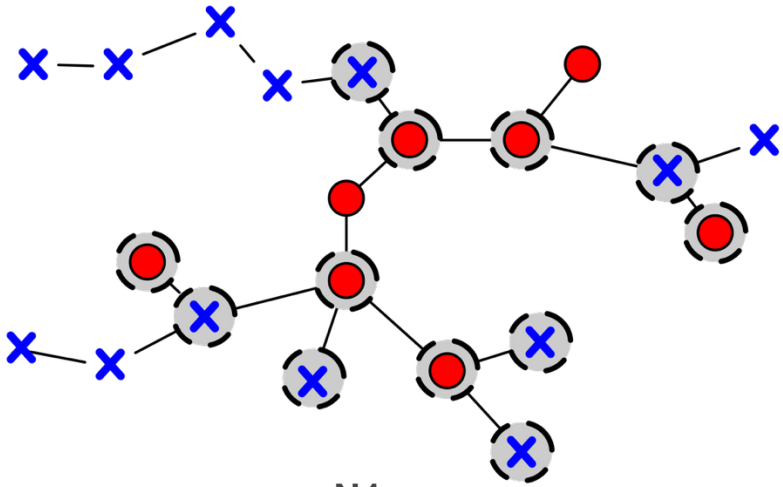
L1



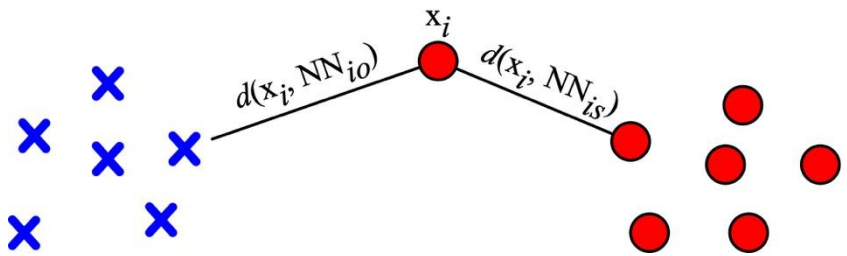
L3

Data Complexity Measures: Neighborhood measures

- **Neighborhood measures:** Characterize the presence and density of same or different classes in local neighborhoods.



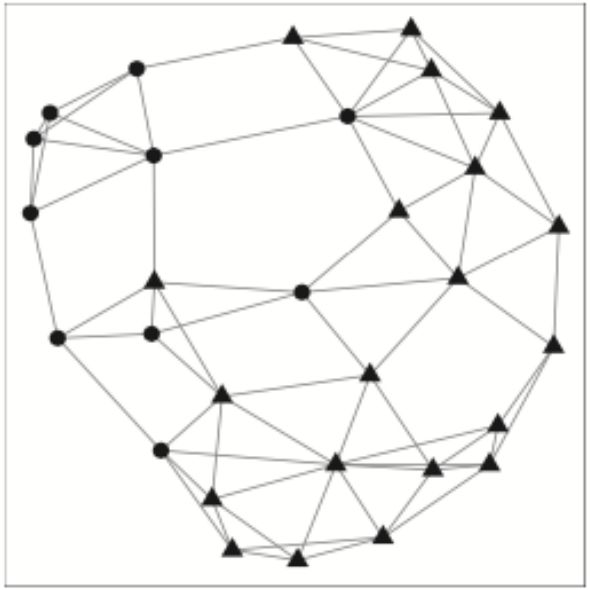
N1

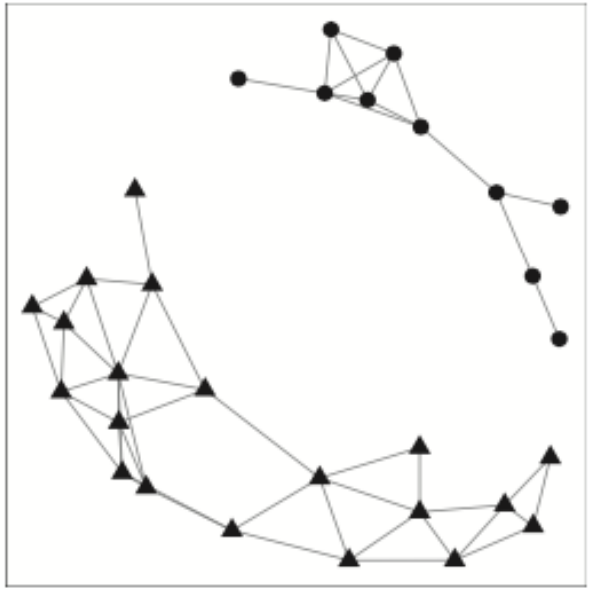


N2

Data Complexity Measures: Network measures

- **Network measures:** Extract structural information from the dataset by modeling it as a graph.





Data Complexity Measures: Dimensionality Measures

- **Dimensionality Measures:** Evaluate data sparsity based on the number of samples relative to the data dimensionality.

  - **Average Number of Features per Dimension (T2):**

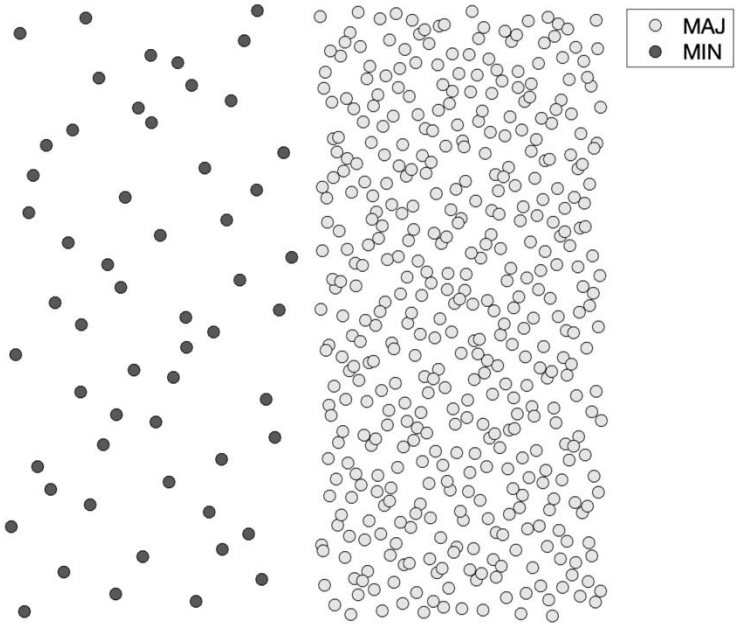
$$T2 = \frac{m}{n}$$

  - **Average Number of PCA Dimensions per Points (T3):**

$$T3 = \frac{m'}{n}$$

Data Complexity Measures: Class Imbalance Measures

- **Class Imbalance Measures:** Consider the ratio of the number of examples between classes.



- **Entropy of Class Proportions (C1):**

$$C1 = -\frac{1}{\log(n_c)} \sum_{i=1}^{n_c} p_{c_i} \log(p_{c_i})$$

- **Imbalance Ratio (C2):**

$$C2 = 1 - \frac{1}{IR}$$

$$IR = \frac{n_c - 1}{n_c} \sum_{i=1}^{n_c} \frac{n_{c_i}}{n - n_{c_i}}$$

Data Complexity Measures: Class Overlap Measures

- **Class Overlap Measures:** Consider class overlap as a concept comprising multiple sources of complexity (feature-level, instance-level overlap, structural overlap, multiresolution overlap).

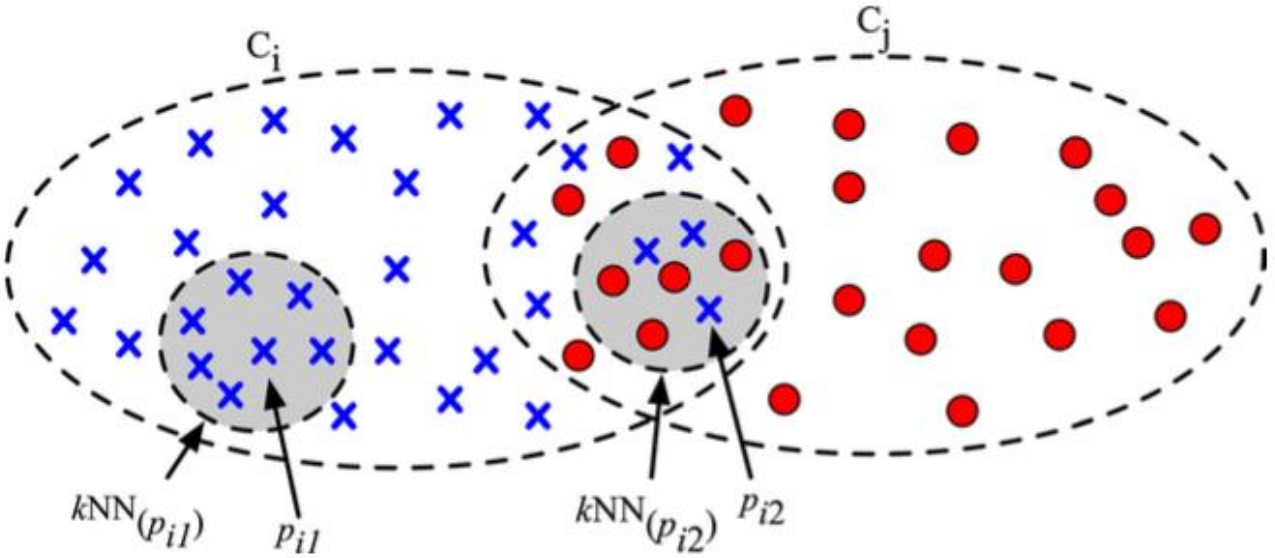
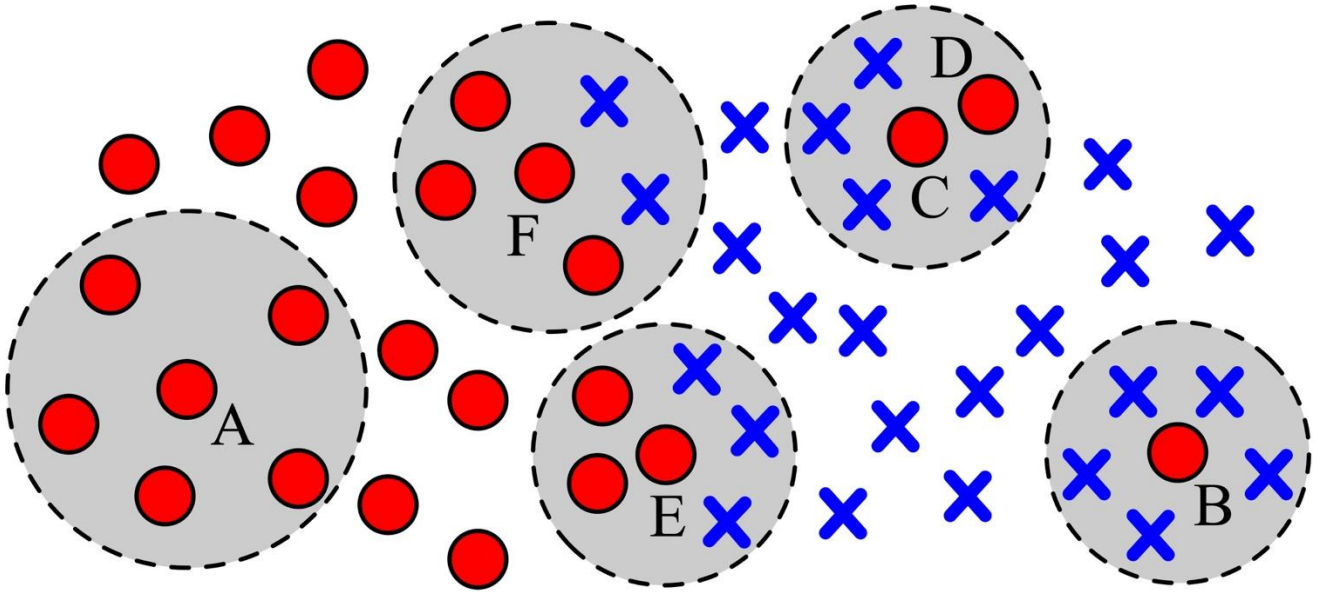


Fig. 18 Basic concepts for R-value computation. Note how $|\text{kNN}(p_{i1}, C_j)| = 0$ and $|\text{kNN}(p_{i2}, C_j)| = 4$, for $k = 6$. Adapted from Oh (2011)

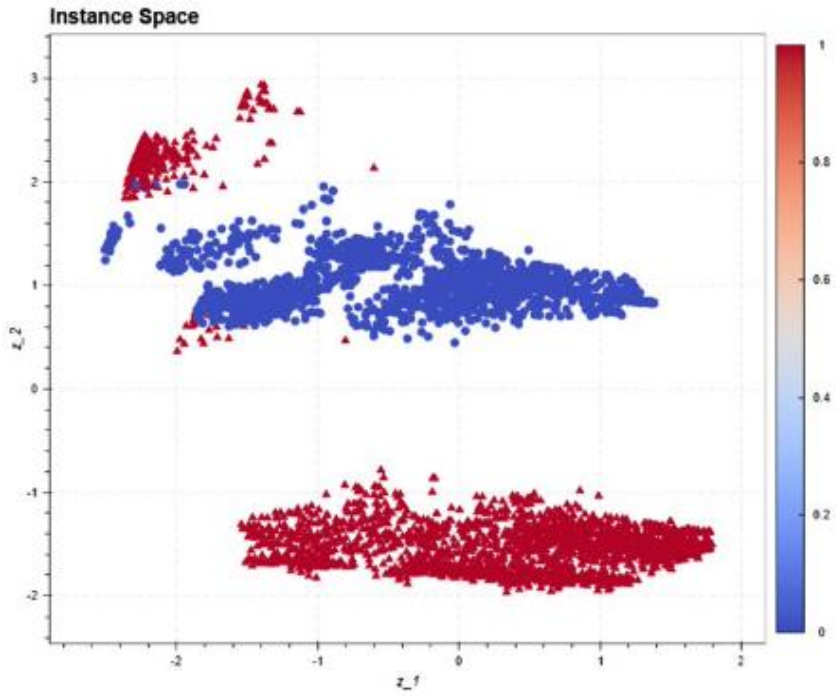
Data Complexity Measures: Data Typology

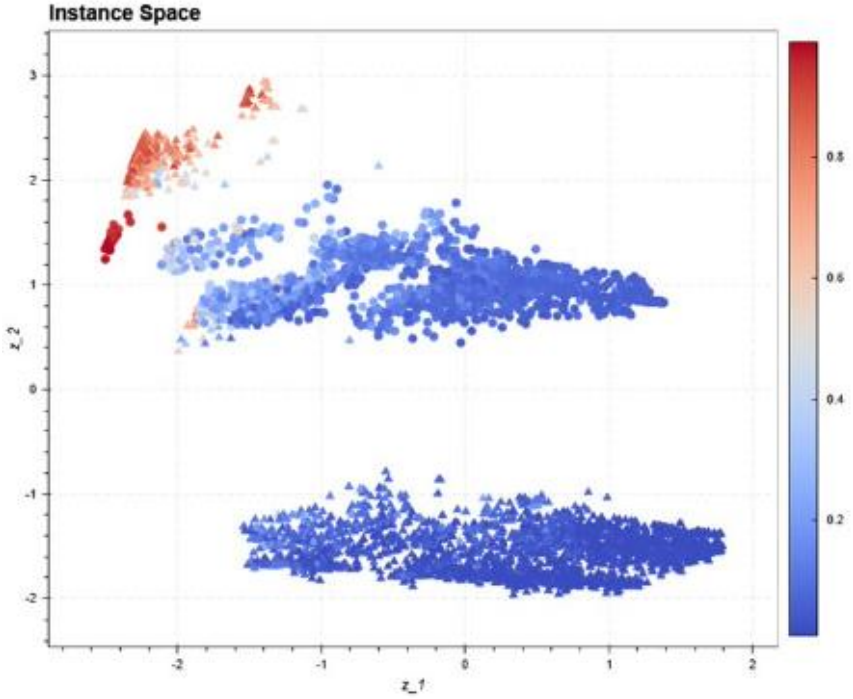
- **Data Typology:** Data complexity is mapped according to the types of examples in data – ***Safe*** (A), ***Borderline*** (E, F), ***Rare*** (C, D), and ***Outlier*** (B).



Data Complexity Measures: Instance Hardness

- **Instance Hardness or Instance-Level Complexity**





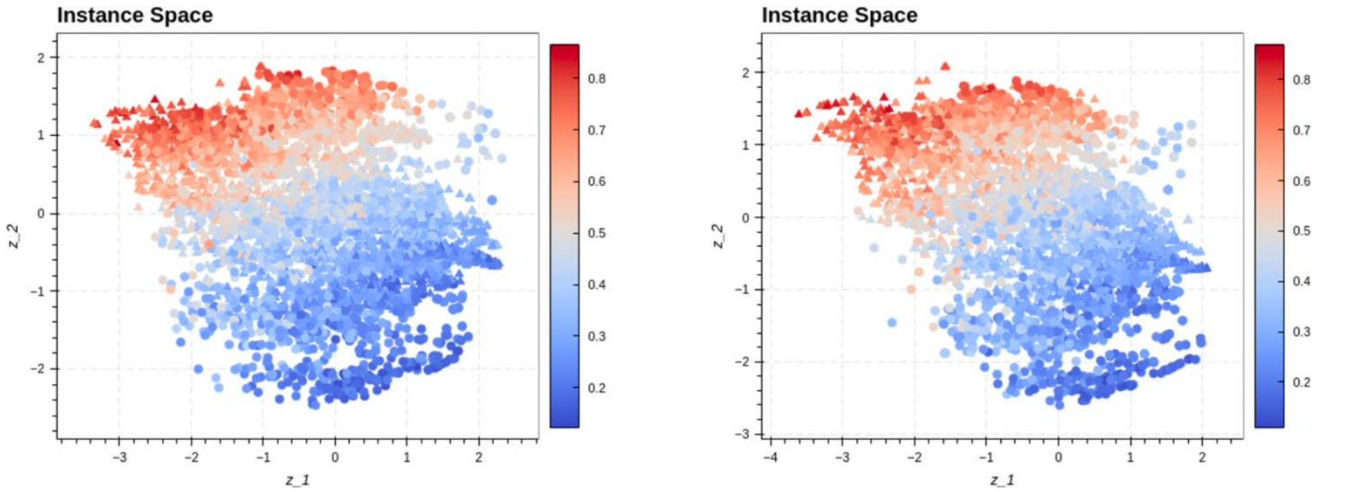
Data Complexity: Applications

- **For a particular application:**
  - Determine the existence of any **learnable structure**
  - **Set expectations** on potential gains by learning algorithms
  - **Compare** different problem formulations, including *alternative class definitions, noise conditions, sampling strategies, choices of features, feature transformation*
  - **Selection of classifiers** and classifier combination schemes, or control classifier training

- **For research in classification methods:**
  - Determine if a dataset **is suitable for evaluating different learning algorithms**
  - Tailor the **collection of benchmarks** to cover a large range of the complexity space
  - Outline the **domains of competence** of classifiers
  - Guide the **design of new classifiers** covering “bling spots” (regions where no known classifiers can do well).

Data Complexity: Applications

- **For a particular application:**



(a) COMPAS dataset ISA projections with `race` as an input attribute.

(b) COMPAS dataset ISA projections without `race` as an input attribute.

Fig. 8 COMPAS dataset ISA projections with and without `race` as an input attribute, colored according to the IH value for each instance

Data Complexity: Applications

- **For a particular application:**

Table 2 Summary statistics for high IH instances of both analyses of the COMPAS dataset, with and without `race` as an input attribute

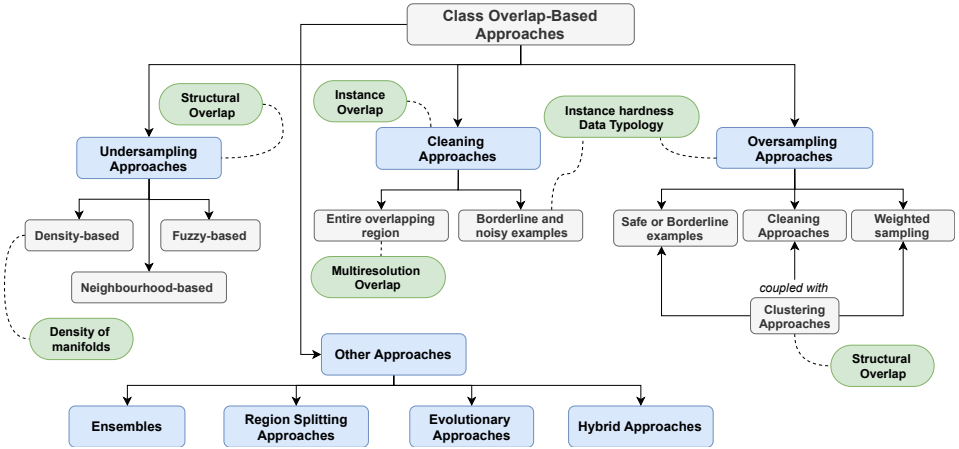
Class	Recidivist stats	With `race`	Without `race`
Recidivist	Caucasian (%)	52.7	50.6
	Men (%)	77.3	78.8
	Avg Age	37.6	37.1
	Avg Prior Offenses	1.3	1.4
Non-recidivist	African American (%)	80.4	71.2
	Men (%)	92.4	90.2
	Avg Age	30.1	30.0
	Avg Prior Offenses	6.9	6.9

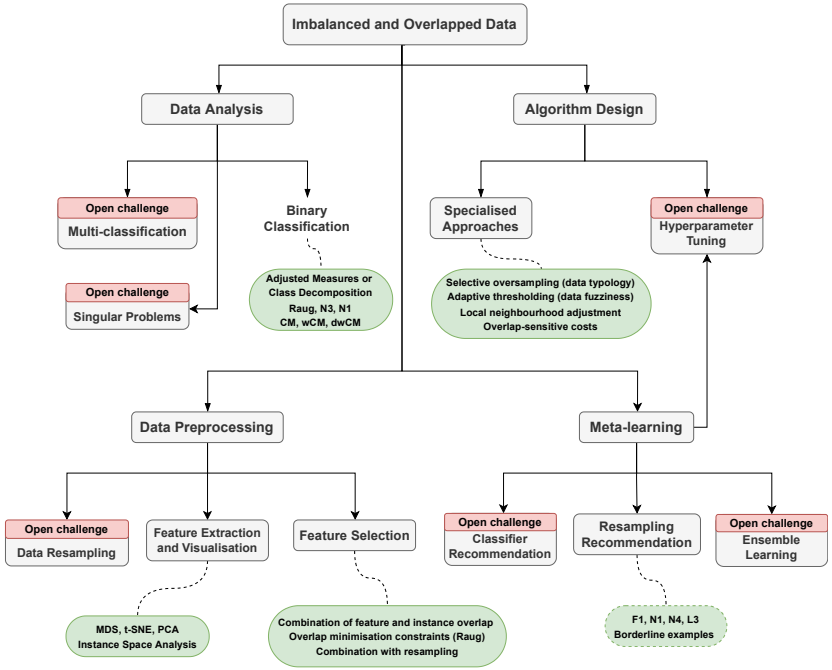
- Non-recidivists with high values of IH are a surrogate for FP. In this group, instances represent, on average, **younger male African-Americans with a higher average of prior offenses than the average of the non-recidivists in the entire dataset**, which explain why they are harder to classify.

- While FN is most likely for individuais on the Caucasian race, **FP has occurred more frequently for the African-American individuals.**

Data Complexity: Applications

- **For research in classification methods:**

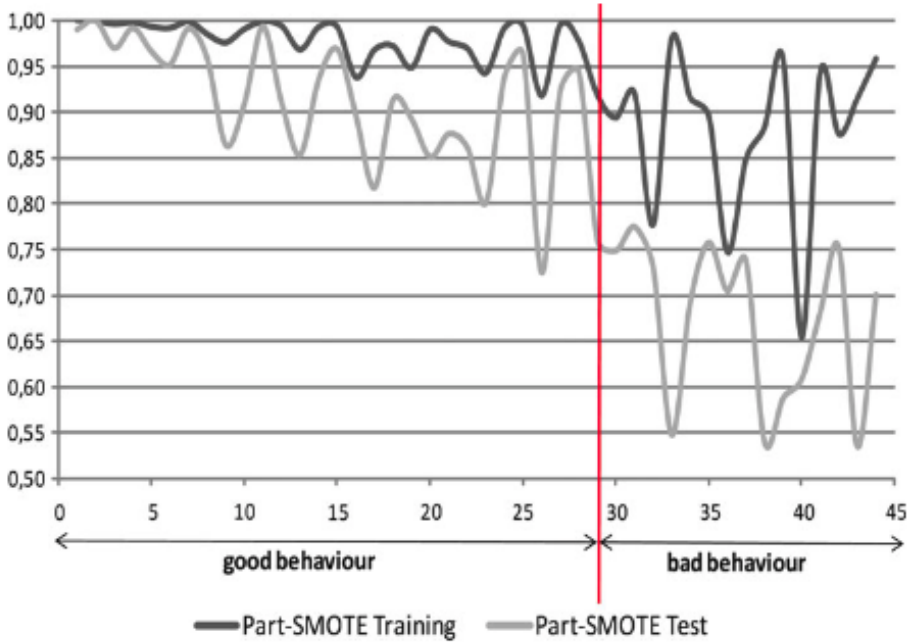




Data Complexity: Applications

- **For applications and research alike:**

Fig. 16 PART with SMOTE AUC in Training/Test sorted by N4



Meta-Learning

Learning (how) to learn

Meta-Learning (MtL): Definition

1. Metalearning studies how learning systems can increase in efficiency through experience; the goal is to understand how learning itself can become flexible according to the domain or task under study (Vilalta and Drissi 2002a).
2. The primary goal of metalearning is the understanding of the interaction between the mechanism of learning and the concrete contexts in which that mechanism is applicable (Giraud-Carrier 2008).
3. Metalearning is the study of principled methods that exploit metaknowledge to obtain efficient models and solutions by adapting machine learning and data mining processes (Brazdil et al. 2009).
4. Metalearning monitors the automatic learning process itself, in the context of the learning problems it encounters, and tries to adapt its behaviour to perform better (Vanschoren 2010).

Lemke et al. (2015), Metalearning: A survey of trends and technologies

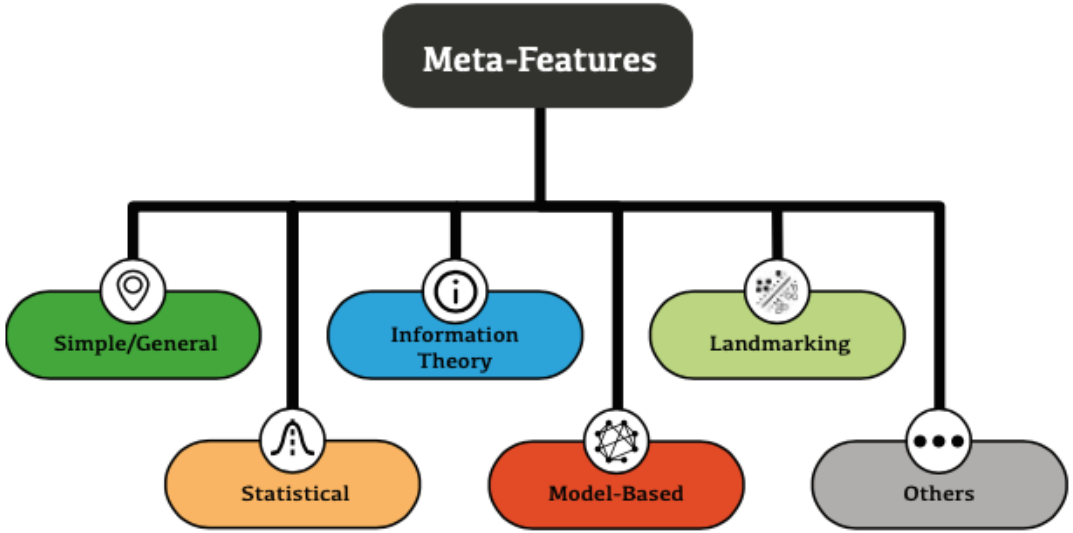
Meta-Learning: Definition

- **What algorithm to use on a given dataset? What parameters to select? How to preprocess the data?**
- **No Free Lunch Theorem:** No algorithm will perform best on all problems. Algorithms need to be tailored to the problem at hand.
- By **characterizing datasets in terms of meta-features**, we can compare and discuss different datasets and relate them to algorithm performance.
- This avoids extensive **trial-and-error** procedures, **brute force** searches for parametrization, saves time and may derive new insights.

Gemert (2017), On the influence of dataset characteristics on classifier performance.

Meta-Learning: Meta-features

- In a learning task, an algorithm learns from several features of a dataset, which, **in the case of meta-learning, are meta-properties, meta-knowkedge, meta-information, metadata** of the dataset. These dataset characteristics are called **meta-features.**



Rivolli et al. (2018). Towards reproducible empirical research in meta-learning. arXiv preprint arXiv:1808.10406, 32-52.

Meta-features: Simple/General

- **Simple meta-features:** Comprise basic descriptive information about the dataset, such as the number of examples or class proportions**.** They can be **easily extracted** from data with **low computational resources**.

Acronym	Task	Extraction	Argument	Domain	Hyperp.	Range	Card.	Exception
attrToInst	Any	Direct	*P	Both	No	$[0, \overline{d}]$	1	No
catToNum	Any	Direct	*P	Both	No	$[0, \overline{d}]$	1	Yes
freqClass	Classif.	Direct	T	Categ.	No	$[0, 1]$	q	No
instToAttr	Any	Direct	*P	Both	No	$[0, \overline{n}]$	1	No
nrAttr	Any	Direct	*P	Both	No	$[1, +\infty]$	1	No
nrBin	Any	Direct	*P	Both	No	$[0, d]$	1	No
nrCat	Any	Direct	*P	Both	No	$[0, d]$	1	No
nrClass	Classif.	Direct	T	Categ.	No	$[2, \overline{n}]$	1	No
nrInst	Any	Direct	*P	Both	No	$[q, +\infty]$	1	No
nrNum	Any	Direct	*P	Both	No	$[0, d]$	1	No
numToCat	Any	Direct	*P	Both	No	$[0, \overline{d}]$	1	Yes

Meta-features: Statistical

- **Statistical meta-features:** Capture the **statistical properties** of data, such as data distribution indicators (average, standard deviation, skewness, kurtosis). They characterize **only numerical features.**

Acronym	Task	Extraction
canCor	Classif.	Indirect
gravity	Classif.	Indirect
cor	Any	Direct
cov	Any	Direct
nrDisc	Classif.	Indirect
eigenvalues	Any	Indirect
gMean	Any	Direct
hMean	Any	Direct
iqRange	Any	Direct
kurtosis	Any	Direct
mad	Any	Direct
max	Any	Direct
mean	Any	Direct
median	Any	Direct
min	Any	Direct
nrCorAttr	Any	Direct
nrNorm	Any	Direct
nrOutliers	Any	Direct
range	Any	Direct
sd	Any	Direct
sdRatio	Classif.	Indirect
skewness	Any	Direct
sparsity	Any	Direct
tMean	Any	Direct
var	Any	Direct
wLambda	Classif.	Indirect

Meta-features: Information Theory

- **Information-theoretic meta-features:** Captures measures from the **information theory**, mostly based on **entropy** measures. They can be used to characterize **discrete attributes.**

Acronym	Task	Extraction	Argument	Domain	Hyperp.	Range	Card.	Exception
attrConc	Any	Direct	2P	Categ.	No	$[0, 1]$	$\overline{d^2}$	No
attrEnt	Any	Direct	1P	Categ.	No	$[0, \log_2(n)]$	d	No
classConc	Classif.	Direct	1P+T	Categ.	No	$[0, 1]$	d	No
classEnt	Classif.	Direct	T	Categ.	No	$[0, \log_2(q)]$	1	No
eqNumAttr	Classif.	Direct	$*$P+T	Categ.	No	$[0, \infty]$	1	No
jointEnt	Classif.	Direct	1P+T	Categ.	No	$[0, \log_2(n)]$	d	No
mutInf	Classif.	Direct	1P+T	Categ.	No	$[0, \log_2(n)]$	d	No
nsRatio	Classif.	Direct	$*$P+T	Categ.	No	$[0, \infty]$	1	No

Meta-features: Model-Based

- **Model-based meta-features:** Extracted from a **model induced** from the training data.

Acronym	Task	Extraction	Argument	Domain	Hyperp.	Range	Card.	Exception
leaves	Supervised	Indirect	*P+T	Both	Yes	$[q, \overline{n}]$	1	No
leavesBranch	Supervised	Indirect	*P+T	Both	Yes	$[1, \overline{n}]$	$\overline{n}$	No
leavesCorrob	Supervised	Indirect	*P+T	Both	Yes	$[0, 1]$	$\overline{n}$	No
leavesHomo	Supervised	Indirect	*P+T	Both	Yes	$[q, +\infty]$	$\overline{n}$	No
leavesPerClass	Classification	Indirect	*P+T	Both	Yes	$[0, 1]$	q	No
nodes	Supervised	Indirect	*P+T	Both	Yes	$[q, \overline{n}]$	1	No
nodesPerAttr	Supervised	Indirect	*P+T	Both	Yes	$[0, \overline{n}]$	1	No
nodesPerInst	Supervised	Indirect	*P+T	Both	Yes	$[0, 1]$	1	No
nodesPerLevel	Supervised	Indirect	*P+T	Both	Yes	$[1, \overline{n}]$	$\overline{n}$	No
nodesRepeated	Supervised	Indirect	*P+T	Both	Yes	$[0, \overline{n}]$	$\overline{d}$	No
treeDepth	Supervised	Indirect	*P+T	Both	Yes	$[1, \overline{n}]$	$\overline{n}$	No
treeImbalance	Supervised	Indirect	*P+T	Both	Yes	$[0, 1]$	$\overline{n}$	No
treeShape	Supervised	Indirect	*P+T	Both	Yes	$[0, 0.5]$	$\overline{n}$	No
varImportance	Supervised	Indirect	*P+T	Both	Yes	$[0, 1]$	$\overline{d}$	No

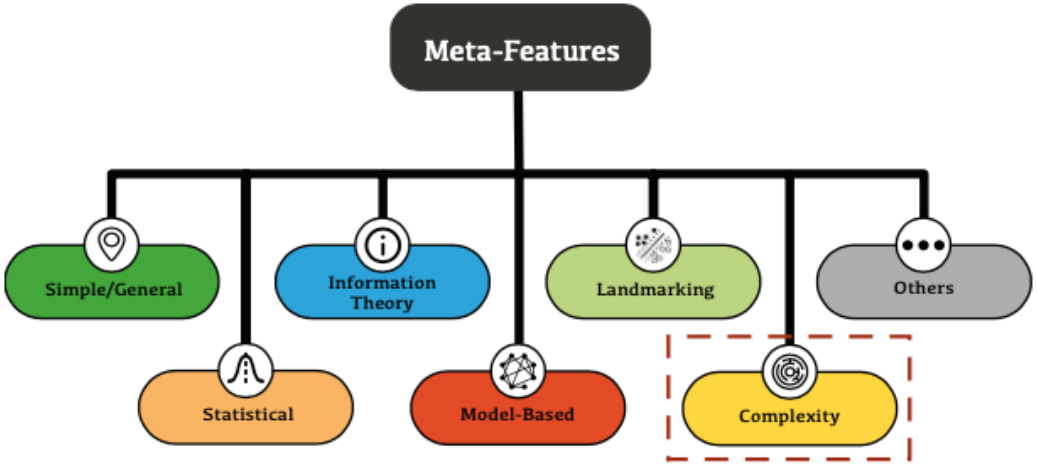
Meta-features: Landmarking

- **Landmarking meta-features:** Use the performance of **simple and fast learning algorithms** to characterize datasets. Algorithms should have **different biases and low computational cost**.

Acronym	Task	Extraction	Argument	Domain	Hyperp.	Range	Card.	Exception
bestNode	Supervised	Indirect	∗P+T	Both	Yes	[0, 1]	*user-defined*	No
eliteNN	Supervised	Indirect	∗P+T	Both	Yes	[0, 1]	*user-defined*	No
linearDiscr	Supervised	Indirect	∗P+T	Num.	Yes	[0, 1]	*user-defined*	Yes
naiveBayes	Supervised	Indirect	∗P+T	Both	Yes	[0, 1]	*user-defined*	No
oneNN	Supervised	Indirect	∗P+T	Both	Yes	[0, 1]	*user-defined*	No
randomNode	Supervised	Indirect	∗P+T	Both	Yes	[0, 1]	*user-defined*	No
worstNode	Supervised	Indirect	∗P+T	Both	Yes	[0, 1]	*user-defined*	No

Meta-features: Others

- "Others" includes:
  - Clustering and Distance-Based
  - **Complexity**
  - Miscellaneous

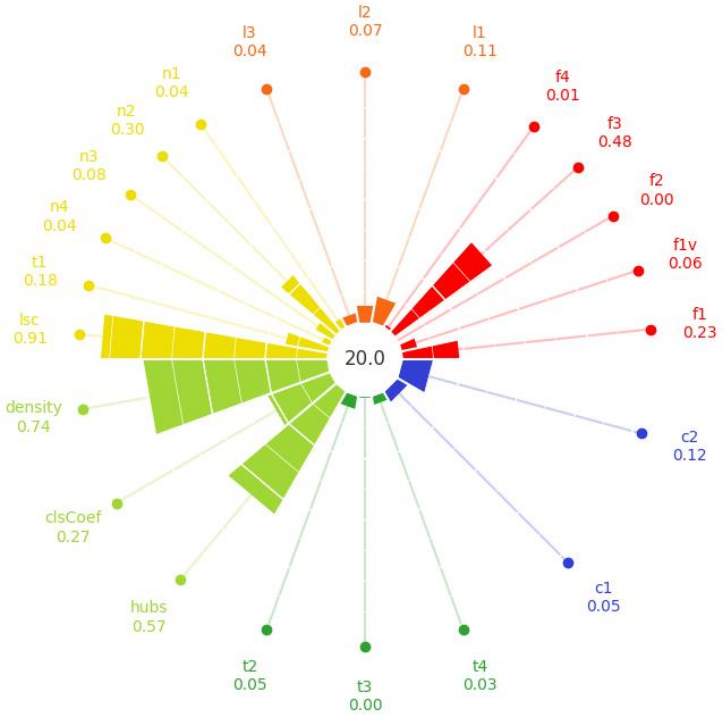


Acronym	Task	Extract	Argument	Domain
clsCoef	Classif.	Indirect	∗P+T	Num.
graphDensity	Classif.	Indirect	∗P+T	Num.
F1	Classif.	Direct	1P+T	Both
F1v	Classif.	Indirect	∗P+T	Both
F2	Classif.	Direct	1P+T	Num.
F3	Classif.	Direct	1P+T	Num.
F4	Classif.	Direct	∗P+T	Num.
Hubs	Classif.	Indirect	∗P+T	Num.
LSC	Classif.	Direct	∗P+T	Num.
L1	Classif.	Indirect	∗P+T	Num.
L2	Classif.	Indirect	∗P+T	Num.
L3	Classif.	Indirect	∗P+T	Num.
N1	Classif.	Indirect	∗P+T	Num.
N2	Classif.	Direct	∗P+T	Both
N3	Classif.	Direct	∗P+T	Both
N4	Classif.	Direct	∗P+T	Both
T1	Classif.	Direct	∗P+T	Num.
T2	Any	Direct	∗P	Both
T3	Any	Indirect	∗P	Num.
T4	Any	Indirect	∗P	Num.

Data Complexity & Meta-Learning

Practice with Python

Problexity: Problem Complexity Assessment



- Implements the data complexity measures as described in *Lorena et al. (2019).*

```
# Loading benchmark dataset from scikit-learn
from sklearn.datasets import load_breast_cancer
X, y = load_breast_cancer(return_X_y=True)

# Initialize ComplexityCalculator with default parametrization
cc = px.ComplexityCalculator()

# Fit model with data
cc.fit(X,y)
```

https://problexity.readthedocs.io/en/latest/

pip install problexity

Komorniczak and Ksieniewicz (2023), Problexity – an open-source Python library for binary classification problem complexity assessment. Neurocomputing.

PyMFE: Python Meta-Feature Extractor

```python
# Load a dataset
from sklearn.datasets import load_iris
from pymfe.mfe import MFE

data = load_iris()
y = data.target
X = data.data

# Extract default measures
mfe = MFE()
mfe.fit(X, y)
ft = mfe.extract()
print(ft)

# Extract general, statistical and information-theoretic measures
mfe = MFE(groups=["general", "statistical", "info-theory"])
mfe.fit(X, y)
ft = mfe.extract()
print(ft)
```

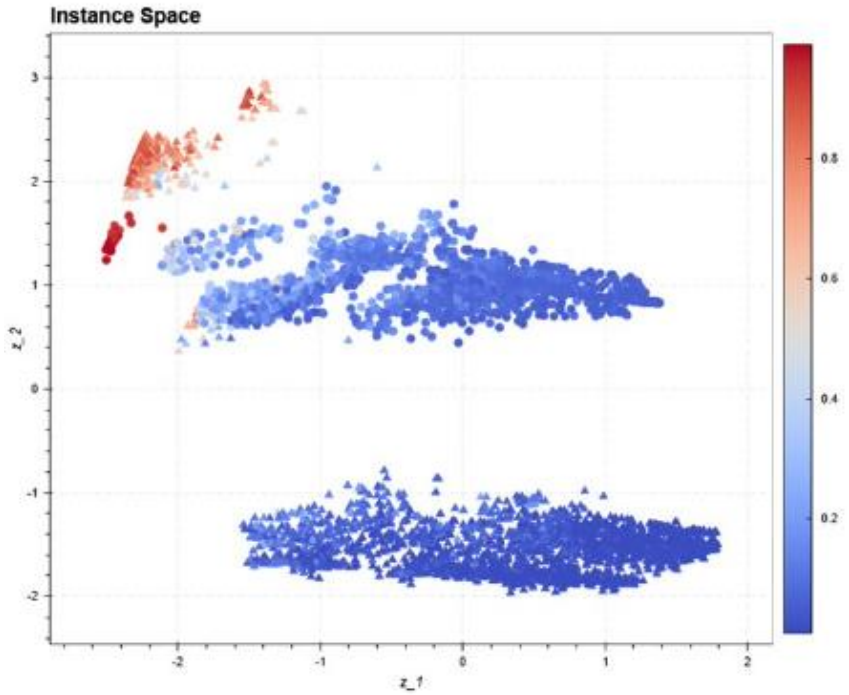
- Comprehensive suite of meta-features
- Different families of meta-features
- Several summarization functions

https://problexity.readthedocs.io/en/latest/

pip install pymfe

Rivolli et al. (2018). Towards reproducible empirical research in meta-learning. arXiv

PyHard: Instance Hardness Analysis in Machine Learning



- Uses Instance Space Analysis (ISA) to produce a hardness embedding of a dataset relating the performance of ML models to estimated instance hardness meta-features.

https://ita-ml.gitlab.io/pyhard/

```
pip install pyhard
```

Paiva et al. (2021), PyHard: a novel tool for generating hardness embeddings to support data-centric analysis

References and Further Reading

- Ho, T. K., & Basu, M. (2002). Complexity measures of supervised classification problems. IEEE transactions on pattern analysis and machine intelligence, 24(3), 289-300.
- Lorena, A. C., Garcia, L. P., Lehmann, J., Souto, M. C., & Ho, T. K. (2019). How complex is your classification problem? a survey on measuring classification complexity. ACM Computing Surveys (CSUR), 52(5), 1-34.
- Komorniczak, J., & Ksieniewicz, P. (2023). problexity: An open-source Python library for supervised learning problem complexity assessment. Neurocomputing, 521, 126-136.
- Alcobaça, E., Siqueira, F., Rivolli, A., Garcia, L. P., Oliva, J. T., & De Carvalho, A. C. (2020). MFE: Towards reproducible meta-feature extraction. Journal of Machine Learning Research, 21(111), 1-5.
- Rivolli, A., Garcia, L. P., Soares, C., Vanschoren, J., & de Carvalho, A. C. (2018). Towards reproducible empirical research in meta-learning. arXiv preprint arXiv:1808.10406, 32-52.
- Paiva, P. Y. A., Smith-Miles, K., Valeriano, M. G., & Lorena, A. C. (2021). PyHard: a novel tool for generating hardness embeddings to support data-centric analysis. arXiv preprint arXiv:2109.14430.
- Paiva, P. Y. A., Moreno, C. C., Smith-Miles, K., Valeriano, M. G., & Lorena, A. C. (2022). Relating instance hardness to classification performance in a dataset: a visual approach. Machine Learning, 111(8), 3085-3123.
- Pascual-Triana, J. D., Fernández, A., Novais, P., & Herrera, F. (2024). Fair Overlap Number of Balls (Fair-ONB): A Data-Morphology-based Undersampling Method for Bias Reduction. arXiv preprint arXiv:2407.14210.
- Lorena, A. C., Paiva, P. Y., & Prudêncio, R. B. (2024). Trusting my predictions: on the value of Instance-Level analysis. ACM Computing Surveys, 56(7), 1-28.
- Santos, M. S., Abreu, P. H., Japkowicz, N., Fernández, A., & Santos, J. (2023). A unifying view of class overlap and imbalance: Key concepts, multi-view panorama, and open avenues for research. Information Fusion, 89, 228-253.

Tutorial

T02: Data Complexity & Meta-Learning

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Artificial Intelligence and Society

Module 02: Data Complexity & Meta-Learning

Miriam Seoane Santos
LIAAD, INESC TEC, FCUP, University of Porto
miriam.santos@fc.up.pt