Bottom-Up and Top-Down Workflows for Hypercube- and Clustering-based Knowledge Extractors

Federico Sabbatini[*] Roberta Calegari[†]

[*]Department of Pure and Applied Sciences (DiSPeA) – University of Urbino
[†]Alma Mater Research Institute for Human-Centered Artificial Intelligence – Alma Mater Studiorum—University of Bologna

f.sabbatini@unibo.it, roberta.calegari@unibo.it

5th International Workshop on EXplainable and TRAnsparent AI and Multi-Agent Systems (EXTRAAMAS 2023)
May 29, 2023, London, UK

TAILOR

This work has been partially supported by the EU ICT-48 2020 project TAILOR (No. 952215).

Next in Line...

1 Context & Motivation

2 Hypercube- and Clustering-Based Knowledge Extraction

3 Conclusions

Context

Black-box predictors

- currently adopted in almost every field [Rocha et al., 2012]

  e.g. pattern detection, image and speech recognition
- detect patterns and relationships buried in data

✓ knowledge acquired and reused in similar applications

  ✓ impressive predictive capabilities

✗ knowledge sub-symbolically represented (internal parameters)

  ✗ not suitable for human comprehension (BB behaviour [Lipton, 2018])

  ✗ no explanations provided for predictions

→ unreliability in critical applications, e.g., healthcare, finance, ...

☞ interpretable predictors [Rudin, 2019]

☞ symbolic knowledge-extraction techniques [Kenny et al., 2021]

Motivation

Symbolic knowledge-extraction techniques

The literature offers a growing amount of extraction techniques:

✓ each method offers peculiar advantages

✗ each method is subject to some limitations

- typical design choices are identifiable in the literature
  e.g. hypercubic input space partitioning for regression tasks
  e.g. *first-order logic* rules as interpretable outputs

✓ hypercubic partitioning is human-interpretable

✗ but it may present several criticalities
  e.g. symmetric top-down partitioning on asymmetric data sets
  e.g. bottom-up partitioning without region relevance awareness

Contribution

Discuss knowledge extractors based on hypercubes and clustering

- analyse recurrent designs adopted to explain opaque regressors

  i.e. create rules associated with hypercubic input space regions

- identify of possible issues deriving from this approach

  e.g. slow convergence

  e.g. human-interpretability hindrances

- propose two different knowledge-extraction workflows
  - involving clustering approaches
  - possibly overcoming the identified drawbacks

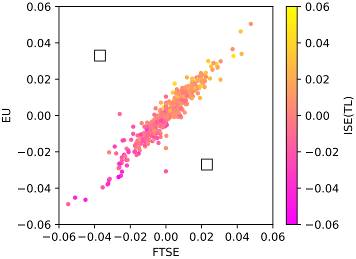
Next in Line. . .

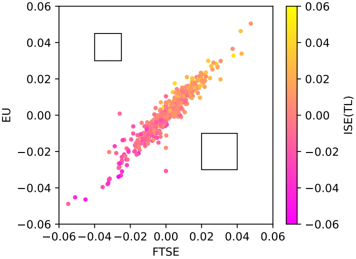
1 Context & Motivation

2 Hypercube- and Clustering-Based Knowledge Extraction

3 Conclusions

ITER [Huysmans et al., 2006]





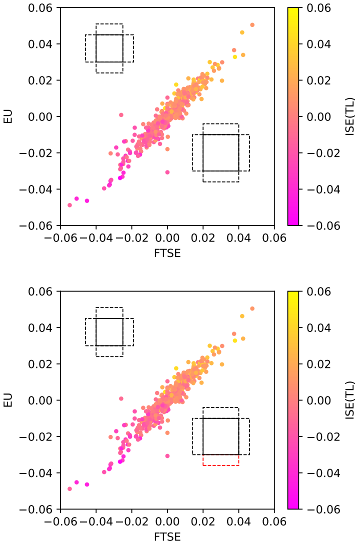
- pedagogical technique for BB regressors
- bottom-up strategy
- it induces a hypercubic partitioning of the input feature space

- starting cubes: random points in the multidimensional space
- cubes are expanded until the final hypercubic output regions

☞ Istanbul Stock Exchange Data Set [Akbilgic et al., 2014]

- FTSE: UK stock market return index
- EU: MSCI European index

ITER [Huysmans et al., 2006] II



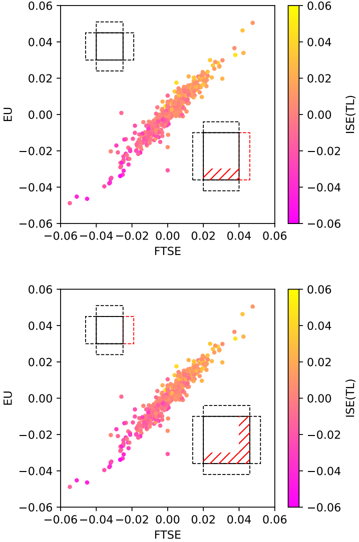
1. build adjacent temporary cubes around the existing ones
   - i.e. 2 temporary cubes per input dimension per existing cube
2. select the best temporary cube
   - i.e. the most similar w.r.t. the adjacent existing cube
3. merge the best temporary cube with the existing one
4. repeat for every successive iteration

! at each iteration only one cube is expanded towards one direction

→ waste of time and resources

☞ non-exhaustivity issue

ITER [Huysmans et al., 2006] III



- stopping criteria
  - maximum number of allowed iterations
  - coverage of the whole input space
  - no possibility of further expand cubes

- non-exhaustivity is particularly evident with high-dimensional domains

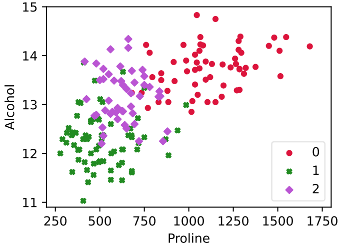
☞ lacking of focus on relevant input space subregions

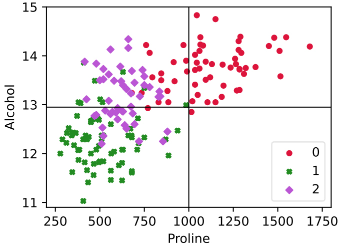
- possible relevance estimate through
  - amount of contained data samples

  e.g. outliers are in low-density regions

  → low-density regions are negligible

GridEx [Sabbatini et al., 2021] and GridREx [Sabbatini and Calegari, 2022] I



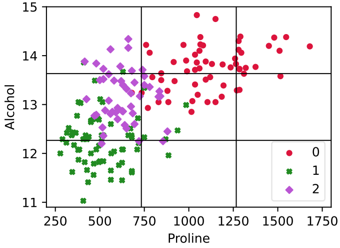


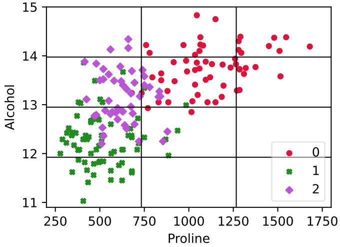
- pedagogical techniques for BB regressors
- top-down strategy
- they induce a hypercubic partitioning of the input feature space

- symmetric, recursive partitioning
- cubes are split until the final hypercubic output regions

☞ Wine Quality Data Set [Forina et al., 1988]
  - x-axis: Proline input feature
  - y-axis: Alcohol input feature

GridEx [Sabbatini et al., 2021] and GridREx [Sabbatini and Calegari, 2022] II



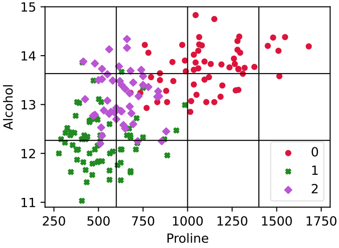


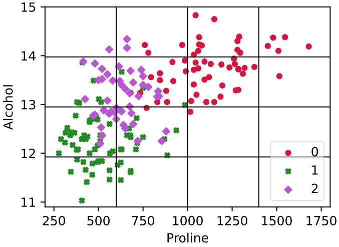
1. find the most distinctive dimensions
2. split each dimension into n user-defined partitions
3. merge pairs of adjacent similar regions
   - similarity w.r.t. expected output
4. split regions having predictive error greater than the user-defined threshold (recursion)
   - same metrics adopted for the BB

! input dimensions are split at each recursion into equal partitions

✓ sensitivity to region relevance
✓ sensitivity to feature relevance
✗ symmetric partitioning

GridEx [Sabbatini et al., 2021] and GridREx [Sabbatini and Calegari, 2022] III

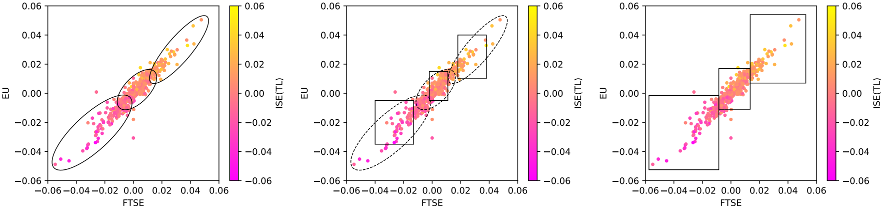




- stopping criteria
  - maximum number of allowed recursions
  - absence of regions associated with high predictive error
- generally, by augmenting the recursion depth and/or the number of splits
  - ✓ higher predictive performance
  - ✗ smaller human readability
- overall performance dependent on the amount of performed splits in symmetric data sets
  - the splits' position is critical

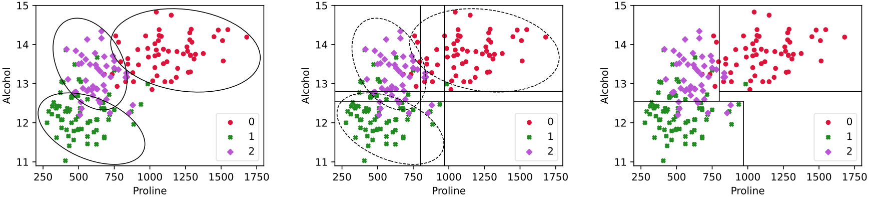
A clustering-based bottom-up workflow

1. apply a clustering technique to the data to identify relevant regions
2. enclose the regions (or part of them) inside hypercubes
3. refine the hypercubes to enhance coverage and predictive performance of the explainable model
4. remove overlaps, or impose a priority order to avoid ambiguity
5. describe each hypercube with human-interpretable rules



A clustering-based top-down workflow

1. apply a clustering technique to the data to identify relevant regions
2. cut the input space to separate different clusters while avoiding spreading instances of a single cluster over multiple regions
3. create hypercubic regions approximating the identified optimum cuts, avoiding overlapping cubes
4. refine the hypercubes by recursively repeating the previous steps for each cube, to enhance the predictive performance of the model
5. describe each hypercube with human-interpretable rules



Next in Line. . .

1 Context & Motivation

2 Hypercube- and Clustering-Based Knowledge Extraction

3 Conclusions

Conclusions

In this paper

- we present 2 workflows for symbolic knowledge extraction
    - top-down vs. bottom-up
- based on hypercubic input feature space partitioning
    - to enhance human interpretability
- exploiting clustering mechanisms
    - to perform a density-driven partitioning
- theoretically achieving better results

  e.g. computational complexity, input space coverage

  e.g. predictive performance, human-readability extent

Future Works

Future works & open issues

- implement knowledge extractors adhering to the presented concepts
  - to be included within the PSyKE framework [Sabbatini et al., 2022]
- select the correct number of clusters to be identified
- handle outliers in the construction of hypercubes
- clusters associated with overlapping hypercubic regions
- discern different clusters approximated by same hypercubes

Bottom-Up and Top-Down Workflows for Hypercube- and Clustering-based Knowledge Extractors

Federico Sabbatini[*] Roberta Calegari[†]

[*]Department of Pure and Applied Sciences (DiSPeA) – University of Urbino
[†]Alma Mater Research Institute for Human-Centered Artificial Intelligence – Alma Mater Studiorum—University of Bologna

f.sabbatini@unibo.it, roberta.calegari@unibo.it

5th International Workshop on EXplainable and TRAnsparent AI and Multi-Agent Systems (EXTRAAMAS 2023)
May 29, 2023, London, UK

This work has been partially supported by the EU ICT-48 2020 project TAILOR (No. 952215).

References I

Akbilgic, O., Bozdogan, H., and Balaban, M. E. (2014).
A novel hybrid rbf neural networks model as a forecaster.
Statistics and Computing, 24:365–375.

Forina, M., Leardi, R., Armanino, C., Lanteri, S., Conti, P., and Princi, P. (1988).
PARVUS: An extendable package of programs for data exploration, classification and correlation.
Journal of Chemometrics, 4(2):191–193.

Huysmans, J., Baesens, B., and Vanthienen, J. (2006).
ITER: An algorithm for predictive regression rule extraction.
In *Data Warehousing and Knowledge Discovery (DaWaK 2006)*, pages 270–279. Springer.

Kenny, E. M., Ford, C., Quinn, M., and Keane, M. T. (2021).
Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies.
Artificial Intelligence, 294:103459.

Lipton, Z. C. (2018).
The mythos of model interpretability.
Queue, 16(3):31–57.

References II

Rocha, A., Papa, J. P., and Meira, L. A. A. (2012).
How far do we get using machine learning black-boxes?
International Journal of Pattern Recognition and Artificial Intelligence, 26(02):1261001–(1–23).

Rudin, C. (2019).
Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.
Nature Machine Intelligence, 1(5):206–215.

Sabbatini, F. and Calegari, R. (2022).
Symbolic knowledge extraction from opaque machine learning predictors: GridREx & PEDRO.
In Kern-Isberner, G., Lakemeyer, G., and Meyer, T., editors, *Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning, KR 2022, Haifa, Israel. July 31 - August 5, 2022*.

Sabbatini, F., Ciatto, G., Calegari, R., and Omicini, A. (2022).
Symbolic knowledge extraction from opaque ML predictors in PSyKE: Platform design & experiments.
Intelligenza Artificiale, 16(1):27–48.

References III

Sabbatini, F., Ciatto, G., and Omicini, A. (2021).
GridEx: An algorithm for knowledge extraction from black-box regressors.
In Calvaresi, D., Najjar, A., Winikoff, M., and Främling, K., editors, *Explainable and Transparent AI and Multi-Agent Systems. Third International Workshop, EXTRAAMAS 2021, Virtual Event, May 3–7, 2021, Revised Selected Papers*, volume 12688 of *LNCS*, pages 18–38. Springer Nature, Basel, Switzerland.