

Unlocking Insights and Trust: The Value of Explainable Clustering Algorithms for Cognitive Agents

Federico Sabbatini* Roberta Calegari†

*Università di Urbino †ALMA MATER STUDIORUM – Università di Bologna
*f.sabbatini1@campus.uniurb.it †roberta.calegari@unibo.it

24th Workshop “From Objects to Agents” (WOA)
Rome, Italy – 8 November 2023



This work was partially supported (i) by the European Union's Horizon Europe AEQUITAS research and innovation programme under grant number 101070363 and (ii) by the European Union ICT-48 2020 project TAILOR (No. 952215).



Next in Line...

- 1 Context & Background
- 2 PSyKE design
- 3 Experiments & Results
- 4 Conclusion



Context & Motivation

Explainable clustering

- it is gaining momentum for cognitive agents^[Arrieta et al., 2020]
 - e.g. human users
 - e.g. AI systems
- it offers several advantages
 - e.g. clustering results in a comprehensible and interpretable manner
 - humans and AI systems can discern the rationale behind the results
 - efficacy and transparency of the clustering process^[Rahman et al., 2023]

Basic idea

Exploiting the benefits of explainable clustering algorithms to augment the capabilities of cognitive agents

Background I

- ML-based predictors detect patterns and relationships buried in data
 - the acquired knowledge is reused in similar applications
- However, their knowledge is not explicitly represented
 - lack of interpretability
 - models described as black boxes^[Lipton, 2018]
 - need of more interpretable predictors^[Rudin, 2019]
 - need of inspection techniques^[Guidotti et al., 2018]

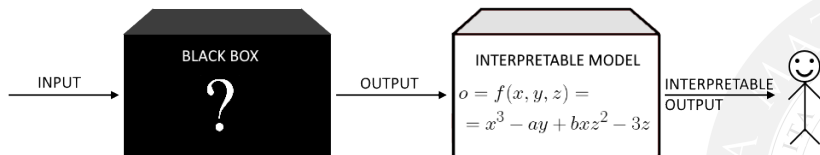


Background II

Some intuitive definitions^[Ciatto et al., 2020]

interpretability property of a system which is easily contemplable for humans

explanation a more interpretable model of a poorly-interpretable object



Background III

ML OOP frameworks

- make ML solutions easily available
- provide abstractions for BB models
- provide various utilities
 - data set pre-processing
 - feature engineering
 - predictive performance measurements
 - data visualisation

e.g. SciKit-Learn for Python, Smile for JVM

Background IV

The PSyKE OOP framework

PSyKE is a general-purpose platform for Python ^[Sabbatini et al., 2022]

- dedicated to symbolic knowledge extraction
- supporting supervised ML tasks
 - e.g. classification and regression
- implementing many pedagogical algorithms
 - Rule-extraction-as-learning (REAL) ^[Craven and Shavlik, 1994]
 - TREPAN ^[Craven and Shavlik, 1996]
 - ITER ^[Huysmans et al., 2006]
 - GridEx ^[Sabbatini et al., 2021] and GridREx ^[Sabbatini and Calegari, 2022]
 - CART ^[Breiman et al., 1984]
 - CREEPY ^[Sabbatini and Calegari, 2023c]
- providing logic rules as output

Contribution of the paper

New design for PSyKE

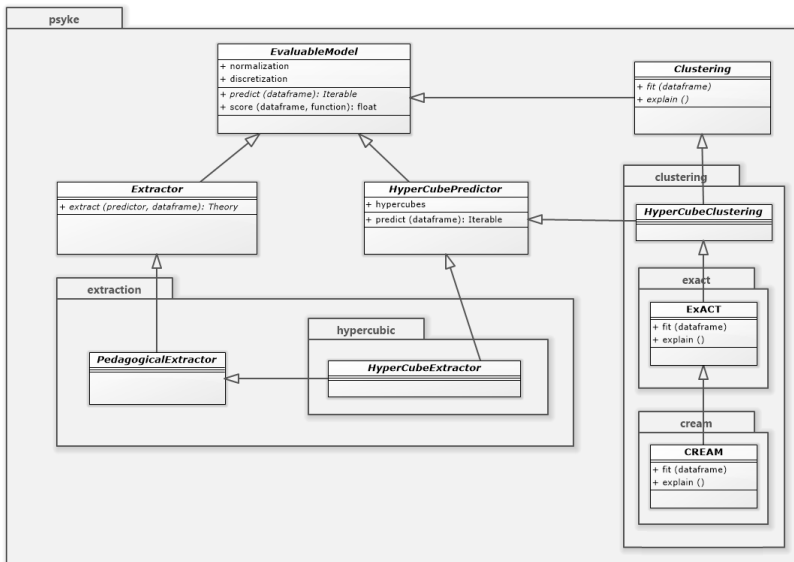
- PSyKE framework generalisation and extension
 - version 0.5
- explainable clustering support
 - via 2 different algorithms
 - ExACT^[Sabbatini and Calegari, 2023a]
 - CREAM^[Sabbatini and Calegari, 2023b]
- goal: enhance the capabilities of cognitive agents

Next in Line...

- 1 Context & Background
- 2 PSyKE design**
- 3 Experiments & Results
- 4 Conclusion



New PSyKE design



Supported Algorithms I

	Knowledge extraction					Clustering	
Algorithm	REAL	TREPAN	CART	ITER GridEx GridREx	CREEPY	EXACT	CREAM
Translucency Pedagogical Decomp.	×	×	×	×	×		
Task Classification Regression Clustering	×	×	×	×	×	×	×
Input feature Binary Discrete Continuous	×	×	×	×	×	×	×
	×	×	×	×	×	×	×
	×	×	×	×	×	×	×
	×	×	×	×	×	×	×

* Only if binary values are encoded as numbers, e.g., 0 and 1

‡ Only if discrete values are binarised, e.g., via one-hot encoding

† Only if discrete values are numeric

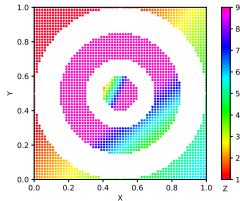
Supported Algorithms II

	Knowledge extraction					Clustering	
Algorithm	REAL	TREPAN	CART	ITER GridEx GridREx	CREEPY	EXACT	CREAM
Output							
String label	×	×	×	×	×	×	×
Constant			×	×	×	×	×
Linear eq.				×	×	×	×
Cluster id						×	×
Knowledge							
Rule list	×			×			
Decision tree		× [§]	× [§]		× [§]	× [§]	× [§]
Interpretation							
Global	×	×	×	×	×	×	×
Local	× [◇]	× [◇]	× [◇]	× [◇]	× [◇]	× [◇]	× [◇]

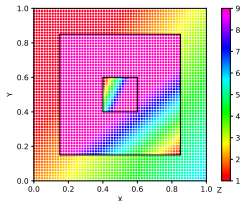
§ Decision trees may be linearised into an ordered list of rules

◇ Local interpretability may be achieved by considering individual items of the explanation

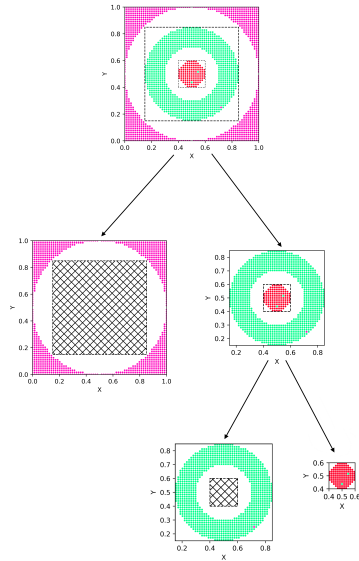
ExACT



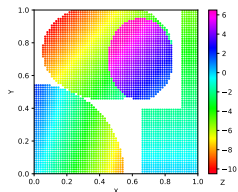
Data set



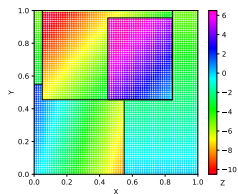
ExACT's partitioning



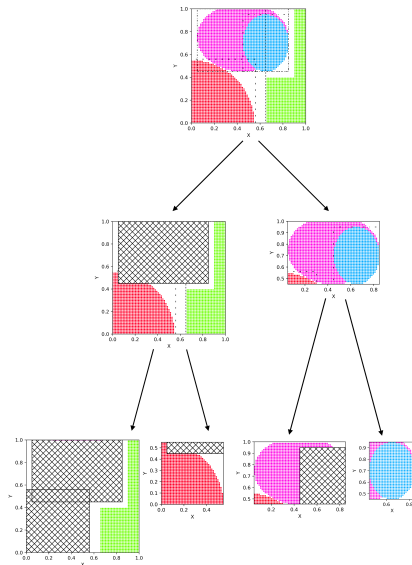
CREAM



Data set



CREAM's partitioning



Example: CREAM

Fundamental steps

- ➊ find the surrounding cube associated with the tree root
- ➋ set the root as the current node
- ➌ apply GMM clustering to the current node and identify n clusters
- ➍ apply DBSCAN to the n clusters for outlier removal
- ➎ approximate each cluster with a minimal hypercube
- ➏ create n pairs of cubes and corresponding difference cubes
 - ! difference cube = current node - approximating cube
- ➐ calculate the predictive error e for each pair, select the best pair
- ➑ expand the current tree node
 - right child $\rightarrow$ best pair's approximating cube
 - left child $\rightarrow$ best pair's difference cube
- ➒ repeat from 3. for every child of the current node having $e > \text{threshold}$

Next in Line...

- 1 Context & Background
- 2 PSyKE design
- 3 Experiments & Results**
- 4 Conclusion

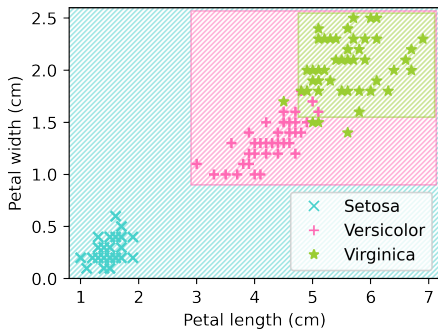


Case study

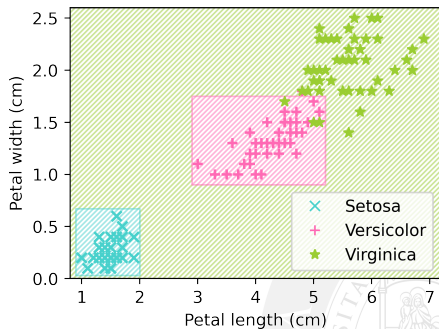
- We tested PSyKE in a simple scenario (Iris data set)
 - classification task
 - 4 continuous input features
 - 1 output label (3 distinct classes)
 - 150 instances split into training and test sets (75% + 25%)
- We trained 2 different explainable clustering algorithms
 - EXACT
 - CREAM
- EXACT and CREAM hyper-parameters:
 - optimum depth set with ORCHID [Sabbatini and Calegari, 2023b]
 - optimum error threshold set with ORCHID
 - number of Gaussian components = 3 (number of distinct classes)



Results I



(a) ExACT.



(b) CREAM.

Results II

Classification rules obtained with ExACT for the Iris data set.

```
1      Class Virginica if
2          PetalWidth in [1.6, 2.5] and
3          PetalLength in [4.8, 6.9] and
4          SepalWidth in [2.5, 3.8] and
5          SepalLength in [5.7, 7.9].
6
7      Class Versicolor if
8          PetalWidth in [0.9, 2.5] and
9          PetalLength in [3.0, 6.9] and
10         SepalWidth in [2.2, 3.8] and
11         SepalLength in [4.9, 7.9].
12
13     Class Setosa otherwise.
```

Results III

Classification rules obtained with CREAM for the Iris data set.

```
1      Class Versicolor if
2          PetalWidth in [0.9, 1.7] and
3          PetalLength in [3.0, 5.2] and
4          SepalWidth in [2.0, 3.4] and
5          SepalLength in [5.0, 7.0].
6
7      Class Setosa if
8          PetalWidth in [0.0, 0.7] and
9          PetalLength in [0.9, 2.0] and
10         SepalWidth in [2.3, 4.4] and
11         SepalLength in [4.3, 5.8].
12
13     Class Virginica otherwise.
```

Next in Line...

- 1 Context & Background
- 2 PSyKE design
- 3 Experiments & Results
- 4 Conclusion**



Summing Up

What we did

- extension of the PSyKE framework's design
- added support to explainable clustering
- augmented capabilities for cognitive agents

What we will (probably) do

- development and integration of new explainable clustering algorithms
- testing explainable clustering within cognitive agents
 - simulation of real-world scenarios

Unlocking Insights and Trust: The Value of Explainable Clustering Algorithms for Cognitive Agents

Federico Sabbatini* Roberta Calegari†

*Università di Urbino †ALMA MATER STUDIORUM – Università di Bologna
*f.sabbatini1@campus.uniurb.it †roberta.calegari@unibo.it

24th Workshop “From Objects to Agents” (WOA)
Rome, Italy – 8 November 2023



This work was partially supported (i) by the European Union's Horizon Europe AEQUITAS research and innovation programme under grant number 101070363 and (ii) by the European Union ICT-48 2020 project TAILOR (No. 952215).



References I

- [Arrieta et al., 2020] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020).
 Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.
Information fusion, 58:82–115
- [Breiman et al., 1984] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984).
Classification and Regression Trees.
 CRC Press
- [Ciatto et al., 2020] Ciatto, G., Schumacher, M. I., Omicini, A., and Calvaresi, D. (2020).
 Agent-based explanations in AI: Towards an abstract framework.
 In Calvaresi, D., Najjar, A., Winikoff, M., and Främling, K., editors, *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, volume 12175 of *LNCS*, pages 3–20.
 Springer, Cham
 DOI:10.1007/978-3-030-51924-7_1
- [Craven and Shavlik, 1994] Craven, M. W. and Shavlik, J. W. (1994).
 Using sampling and queries to extract rules from trained neural networks.
 In *Machine Learning Proceedings 1994*, pages 37–45. Elsevier
 DOI:10.1016/B978-1-55860-335-6.50013-1

References II

- [Craven and Shavlik, 1996] Craven, M. W. and Shavlik, J. W. (1996).
Extracting tree-structured representations of trained networks.
In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference*, pages 24–30. The MIT Press
<http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf>
- [Guidotti et al., 2018] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018).
A survey of methods for explaining black box models.
ACM Computing Surveys, 51(5):1–42
DOI:10.1145/3236009
- [Huysmans et al., 2006] 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
DOI:10.1007/11823728_26

References III

[Lipton, 2018] Lipton, Z. C. (2018).

The mythos of model interpretability.

Queue, 16(3):31–57

DOI:10.1145/3236386.3241340

[Rahman et al., 2023] Rahman, A., Hossain, M. S., Muhammad, G., Kundu, D., Debnath, T., Rahman, M., Khan, M. S. I., Tiwari, P., and Band, S. S. (2023).

Federated learning-based ai approaches in smart healthcare: concepts, taxonomies, challenges and open issues.

Cluster computing, 26(4):2271–2311

[Rudin, 2019] 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

DOI:10.1038/s42256-019-0048-x



References IV

- [Sabbatini and Calegari, 2022] 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*
DOI:10.24963/kr.2022/57
- [Sabbatini and Calegari, 2023a] Sabbatini, F. and Calegari, R. (2023a).
ExACT explainable clustering: Unravelling the intricacies of cluster formation.
In *Proceedings of the 2nd International Workshop on Knowledge Diversity, KoDis 2023, Rhodes, Greece, September 2–8, 2023 (to appear)*
- [Sabbatini and Calegari, 2023b] Sabbatini, F. and Calegari, R. (2023b).
Explainable Clustering with CREAM.
In *Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning*, pages 593–603
DOI:10.24963/kr.2023/58



References V

- [Sabbatini and Calegari, 2023c] Sabbatini, F. and Calegari, R. (2023c).
Unveiling opaque predictors via explainable clustering: The CReEPy algorithm.
In *Proceedings of BEWARE-2023, Rome, Italy, November 6–9, 2023, (submitted to)*
- [Sabbatini et al., 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
DOI:10.3233/IA-210120
- [Sabbatini et al., 2021] 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
DOI:10.1007/978-3-030-82017-6_2