Pivot Tree - An Explainable Model with Pivotal Instances

Alessio Cascione[1] Mattia Setzu[1] Riccardo Guidotti[1,2]
a.cascione@studenti.unipi.it mattia.setzu@unipi.it riccardo.guidotti@unipi.it

[1]University of Pisa, Italy [2] ISTI-CNR, Italy

Explainability... case by case

Explanations tend to be domain-dependent, and on non-relational data they can be...

- **Flat.** How do different cases, e.g., prototypes, relate to one another?
- **Global.** How does a case, e.g., a prototype applies to a specific, or a set of, explanations?
- **Factual-only.** How do we change the model's behavior?

We wish to have a case-based algorithm that *relates prototypical instances*, and provides *counterfactuals* and *different levels of explanation locality*.

Case-based explanations

Explanations are often provided in form of

- Feature importance
- Decision rules
- Counterfactual instances

which, in a relational domain, are often faithful. Still, many of the assumptions underlying their algorithms break in the non-relational domain, e.g., images, graphs, or text:

- Feature correlation
- Linearity of the model
- Space density

Cases and pivots

Case-based explanations rely on a set *prototypes* or *pivots* P, which a model f leverages to predict, e.g., through linear combination [1, 2], or simply through voting, as in k-nearest neighbor. These models are often

- Explainable by design: since we can inspect which cases they are leveraging, and how
- Provide multiple layers of explanations, e.g., instances as well as feature importance
- Limited to local or global explanations

But they lack the ability to provide **counterfactuals**, and to **reveal any relationships among cases** and instances.

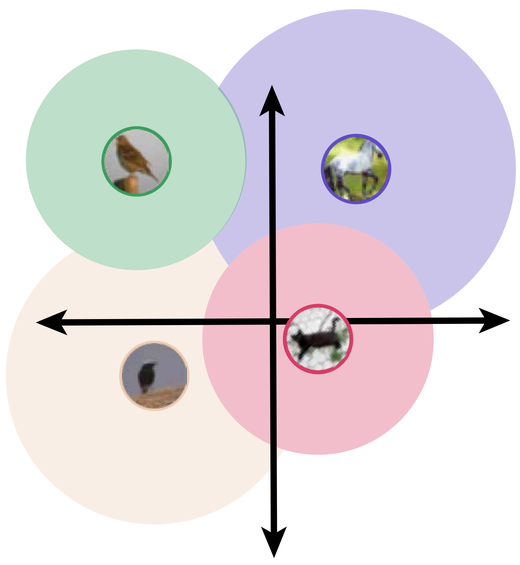


Figure: Input space, partitioned by pivots: each pivot is assigned to a class (color-coded), and its neighborhood adheres to that class.

Pivot Tree at a glance

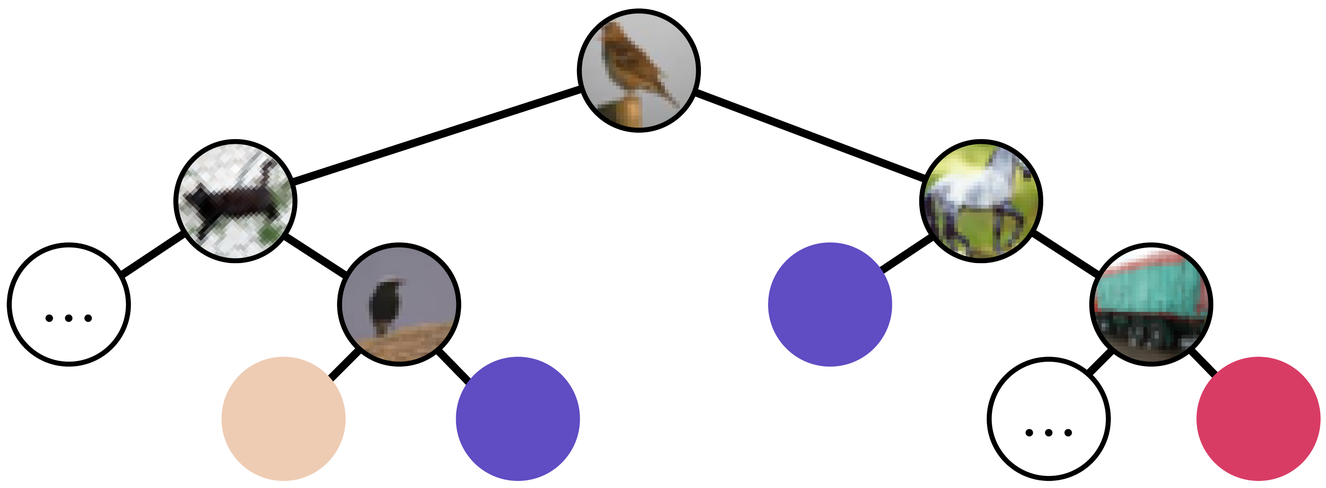


Figure: A Pivot Tree. Similarity to pivots within the tree guides routing, providing a set of instances as explanation. Branching on separate paths provides counterfactuals, while intermediate nodes provide more general instances.

Pivot Tree

We induce a Pivot Tree by:

1. Constructing a pivot space
2. Inducing a decision tree

Constructing the pivot space

As a case-based model, Pivot Tree explicitly models instances in a similarity space. The feature matrix X is mapped to a similarity matrix S s.t. $s_{i,j}$ measures the similarity between x_i and x_j. This step $\pi : \mathbb{R}^m \to \mathbb{R}^n$ yields a dense feature matrix wherein features become similarities.

For non-relational data, embedding models, such as language models or vision models can be used, making Pivot Tree *data-agnostic*.

Inducing the tree

As a feature matrix, S can be used to induce a tree through any tree induction algorithms, e.g., CART. In such a tree, paths turn into sets of *pivotal* instances, routing instances through similarity, and explanations are sets of similar or nonsimilar pivots. Finally, the tree structure provides native counterfactuals! Simply branch in any other subtree, and find a leaf with different prediction.

Pivot... anything!

Once induced, a Pivot Tree has selected a set P of pivots that we use as a dataset to learn another case-based model! E.g., a k-nearest neighbor, a decision tree, etc.

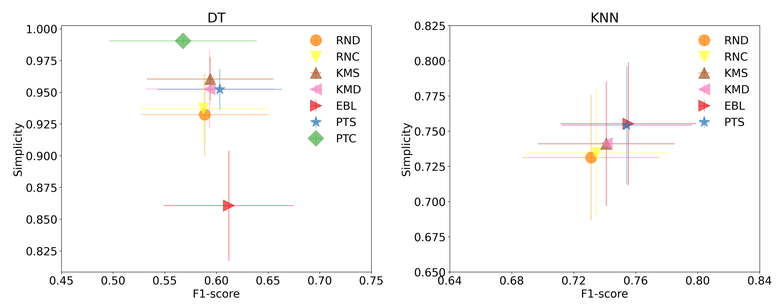
Experiments

Due to its data agnosticism, we test Pivot Tree on tabular (11), time series (5), image (3), and text (5) datasets, comparing against other case-based models such as ε-ball, and k-NN.

References
[1] Prototype Selection for Interpretable Classification. Jacob Bien, and Robert Tibshirani.
[2] Deep Learning for Interpretable Image Recognition. C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin
[3] Interpreting CNNs via Decision Trees. Q. Zhang, Y. Yang, H. Ma, Y. N. Wu.

Results

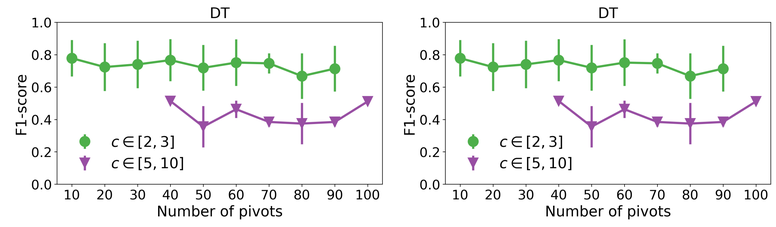
How faithful and complex is Pivot Tree?



(a) Decision Tree (b) k-nearest neighbor

Figure: Case-based methods, including Pivot Tree (PTC), and other downstream case-based methods trained on the pivots extracted by Pivot Tree.

How stable is Pivot Tree?



(a) Decision Tree (b) k-nearest neighbor

Figure: Performance of downstream case-based models on different number of maximum pivots.

An example.

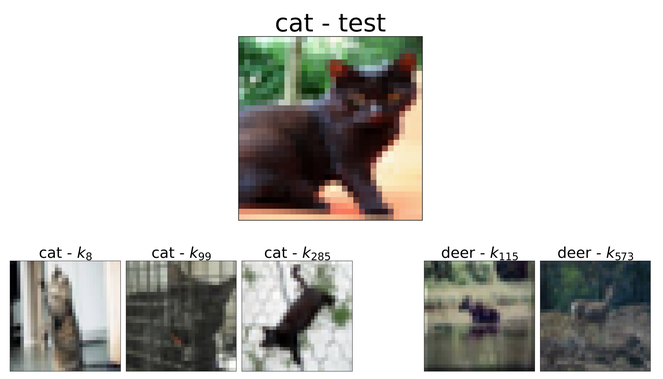


Figure: An instance (top), and an explanation: pivots with high (left) and low (right) similarity.

Highlights

- Pivot Tree is an explainable by-design, prediction and pivot selection model
- Selected pivots can later be used to learn another explainable model
- Selecting pivots by through Pivot Tree yields better performing case-based models

EXPLANATION OF AI DECISION MAKING. MIMOSA PIANO FAIR Future Artificial Intelligence Research Funded by the European Union NextGenerationEU Ministero dell'Università e della Ricerca Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA