This part looks alike this: identifying important parts of explained instances and prototypes

Jacek Karolczak, Jerzy Stefanowski
Poznan University of Technology, Institute of Computing Science
jacek.karolczak@cs.put.poznan.pl, jerzy.stefanowski@cs.put.poznan.pl

The 3rd World Conference on eXplainable Artificial Intelligence
9-11 July, 2025 – Istanbul, Turkey

Introduction

While prototypes provide intuitive links between predictions and representative training instances, their interpretation can be difficult for tabular data with many features. This work enhances prototype-based explanations by identifying and leveraging the most important shared features, improving interpretability in both local explanations and prototype selection.

Identifying *alike parts*

Having instance $\mathbf{x}_i$ and its nearest prototype $\mathbf{p}_j$, for instance SHAP can be used to compute feature importance scores $\phi(h, \mathbf{x}_i^l)$ and $\phi(h, \mathbf{p}_j^l)$, which are then squared and normalized to ensure comparability and prevent cancellation effects:

$$\hat{\phi}(h, \mathbf{x}_i^l) = \frac{(\phi(h, \mathbf{x}_i^l))^2}{\sum_{k=1}^{d} (\phi(h, \mathbf{x}_i^k))^2}, \quad \hat{\phi}(h, \mathbf{p}_j^l) = \frac{(\phi(h, \mathbf{p}_j^l))^2}{\sum_{k=1}^{d} (\phi(h, \mathbf{p}_j^k))^2}. \tag{1}$$

Feature alignment is quantified as the product of normalized importance scores:

$$w_l = \hat{\phi}(h, \mathbf{x}_i^l) \cdot \hat{\phi}(h, \mathbf{p}_j^l). \tag{2}$$

A binary mask $\mathbf{m} \in \{0, 1\}^d$ selects features with above-mean weights to identify those most influential for both the instance and prototype:

$$m_l = \mathbb{1}\left(w_l > \frac{1}{d} \sum_{k=1}^{d} w_k \right). \tag{3}$$

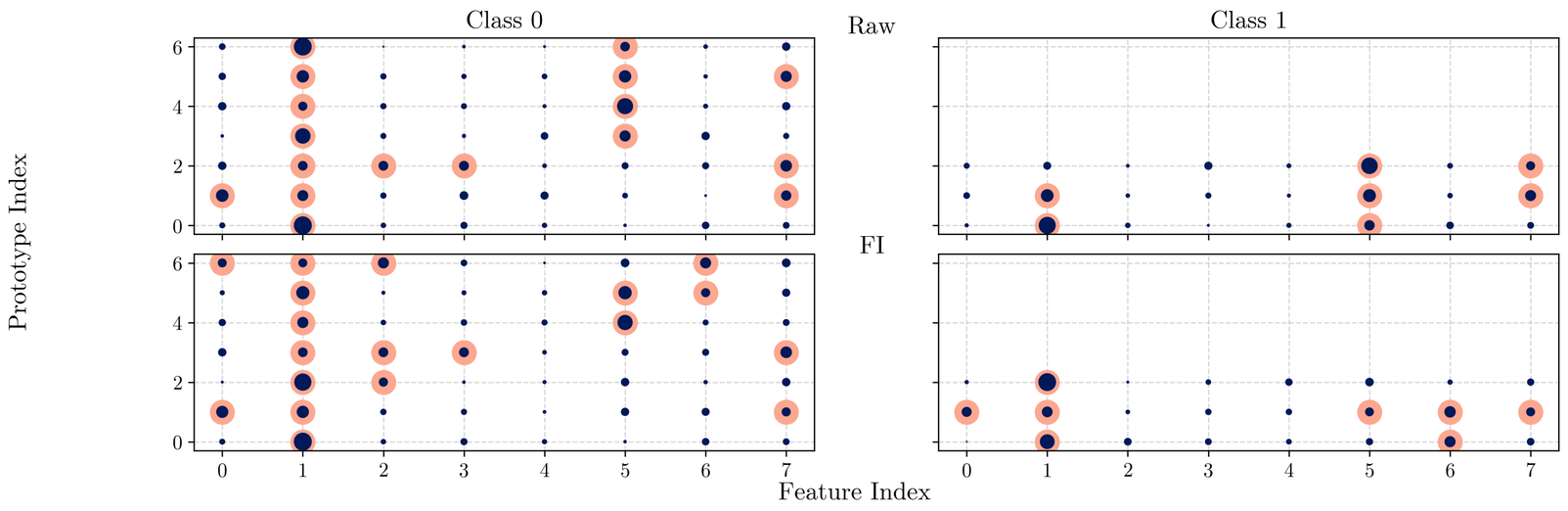
Toy exmaple

Table 1: Feature importance and weights for an instance and its prototype from Apple Quality, with a binary mask highlighting shared key features.

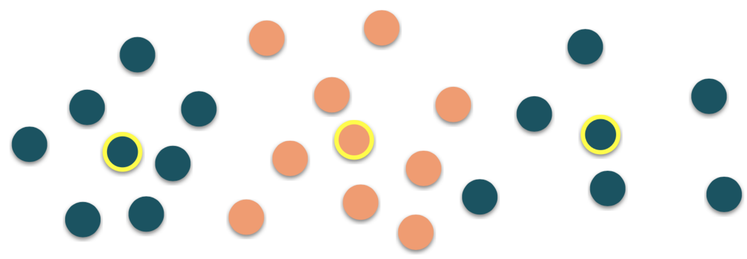
	Size	Weight	**Sweetness**	Crunchiness	Juiciness	**Ripeness**	Acidity
Instance	-2.77	-1.08	-1.72	1.38	0.19	3.65	0.31
Prototype	-0.97	-0.20	-3.07	0.00	-0.52	3.16	-0.52
Weights	**0.18**	0.02	**0.27**	0.00	0.00	**0.51**	0.00
Mask	1	0	1	0	0	1	0

Alike parts found using the new objective function

Figure 1: Comparison of prototypes and important features for the Diabetes dataset. The size of the inner circle represents feature importance, and pink highlights features identified as important for a given prototype.



Prototype explanations



A typical prototype selection algorithms is defined as k- medoids problem with the following objective function:

$$f(\mathcal{P}) = \sum_{i=1}^{|\mathcal{S}|} \min_{\mathbf{p}_j \in \mathcal{P}} d(\mathbf{x}_i, \mathbf{p}_j), \tag{4}$$

solved using greedy approximation [1, 2]. Distance can be a dot product between trainable embeddings, or in tree ensembles, a specialized tree distance metric [1, 2].

New objective function

To strengthen diversification in feature importance, we propose extending the objective function:

$$fi(\mathbf{x}_i, \mathbf{p}_j) = \sum_{l=1}^{d} \frac{(\phi(h, \mathbf{x}_i^l))^2}{\sum_{k=1}^{d} (\phi(h, \mathbf{x}_i^k))^2} \cdot \frac{(\phi(h, \mathbf{p}_j^l))^2}{\sum_{k=1}^{d} (\phi(h, \mathbf{p}_j^k))^2}. \tag{5}$$

The revised function is formally defined as:

$$f(\mathcal{P}) = \sum_{i=1}^{|\mathcal{S}|} \min_{\mathbf{p} \in \mathcal{P}_j} \left(d(\mathbf{x}_i, \mathbf{p}_j) + \beta \cdot fi(\mathbf{x}_i, \mathbf{p}_j) \right). \tag{6}$$

1-NN accuracy

Table 2: Accuracy comparison of raw prototype selection algorithms and their feature importance (FI) enhanced versions.

		Apple Quality	Australia Rain	Breast Cancer	Diabetes	Passenger Satisfaction
A-Pete	FI	.520	.767	.798	.623	.837
	Raw	.487	.424	.488	.427	.783
G-KM	FI	.861	.843	.965	.766	.865
	Raw	.785	.822	.939	.739	.781
SM-A	FI	.571	.809	.623	.734	.779
	Raw	.461	.625	.344	.492	.712

Code

github.com/jkarolczak/important-parts-of-prototypes

References

[1] Jacek Karolczak and Jerzy Stefanowski. A-PETE: Adaptive prototype explanations of tree ensembles. In *Progress in Polish Artificial Intelligence Research*, volume 5, pages 2–8. Warsaw University of Technology, 2024.

[2] Sarah Tan, Matvey Soloviev, Giles Hooker, and Martin T. Wells. Tree space prototypes: Another look at making tree ensembles interpretable. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, FODS '20, page 23–34, 2020.

Acknowledgements

This research was funded in part by National Science Centre, Poland OPUS grant no. 2023/51/B/ST6/00545 and in part by PUT SBAD 0311/SBAD/0752 grant.