Can clustering improve the performance of classifiers? Introduction of a new ensemble technique utilizing cluster analysis methods in classification tasks.

Jakub Piwko, Jędrzej Ruciński, Dawid Płudowski, Antoni Zajko, Franciszek Filipek, Anna Kozak, Katarzyna Woźnica
Warsaw University of Technology

Motivation

Problem Outline:

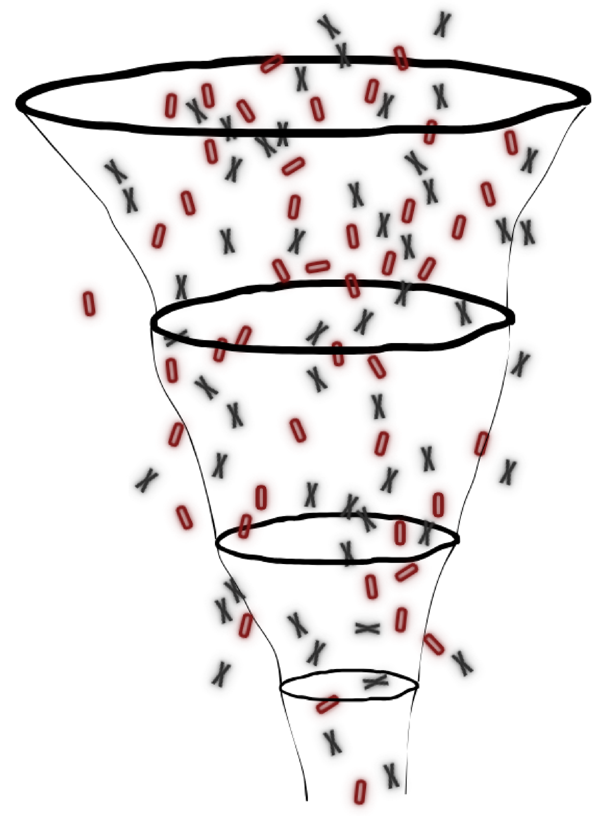
- Traditional ensemble methods treat entire datasets uniformly, often overlooking varied subsets within real-world data. This uniform approach can miss out on the benefits of tailored classification strategies for different data clusters.

Our Solution: Sequential Model Training and "Circles of Hell":

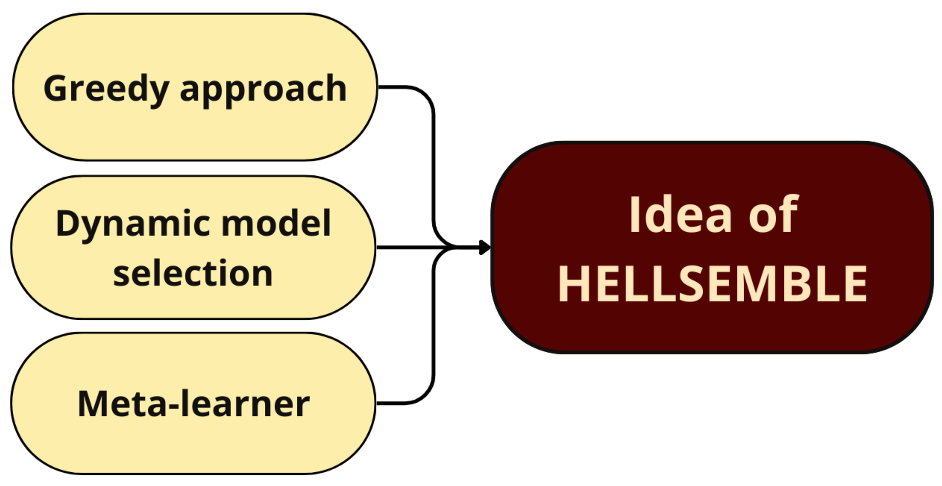
- Each model in the ensemble is trained sequentially, focusing only on observations that previous models failed to predict correctly. A routing model will then learn to map each sample to its appropriate model.
- This process results in distinct groups or "circles of difficulty" representing clusters with escalating complexity. We termed these clusters "circles of hell" to symbolize their rising difficulty levels, giving rise to our framework, **Hellsemble**.

Objective:

- Develop a method for correctly assigning observations to their appropriate circle of difficulty.
- Use this structured approach to boost ensemble performance on increasingly difficult subsets.



Related works



Interesting approaches in enhancing ensembles:

- Greedy model selection, as demonstrated in [1], builds ensembles by iteratively adding models that maximize performance metrics.
- Dynamic ensemble selection (DES) further refines this by tailoring ensemble choices to individual samples, selecting classifiers based on local performance, as seen in [2] and [3].
- Alternatively, [4] introduced a meta-learning framework, where meta-features guide a meta-classifier in selecting effective classifiers for specific instances.

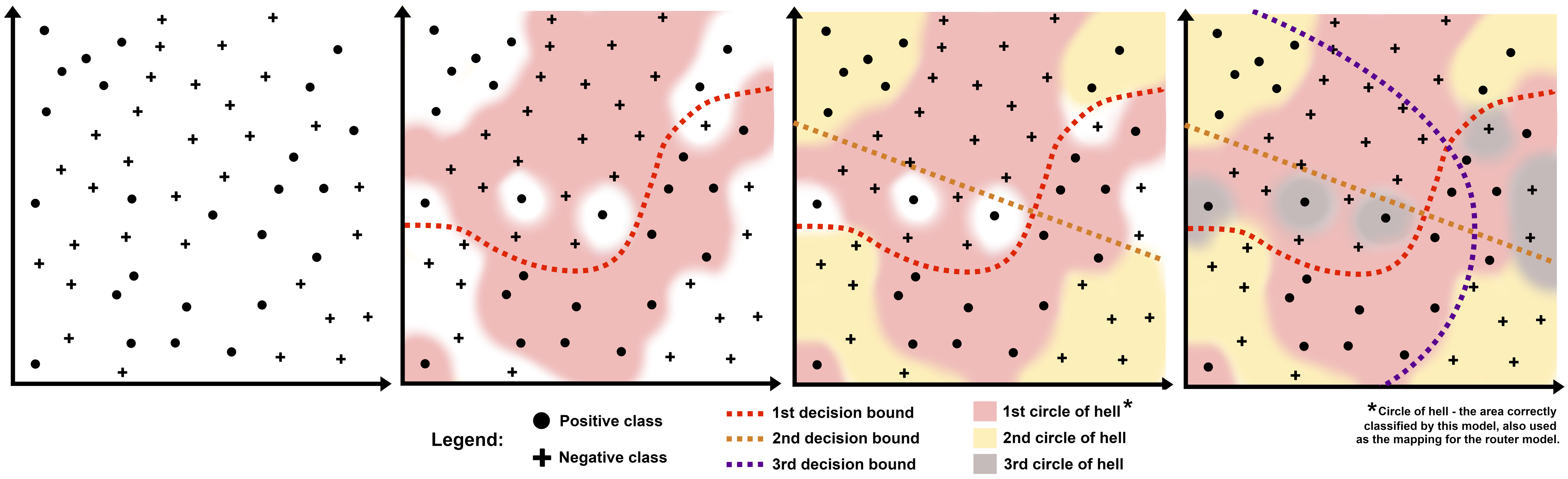
Our work combines the strengths of these approaches: similar to DES, it selects models per sample but uses a greedy method with a routing model similar to meta-learner.

Methodology

Inside the Hellsemble:

- We construct ensemble using classic ML algorithms, iteratively selecting models based on their performance metrics.
- Initially, the model with the best performance is chosen.
- Subsequent model is trained solely on samples misclassified by prior models, thus creating focused training sets.
- For each model, we save indexes of samples that were correctly classified, forming a dedicated training set for a "router model."
- Router model predicts the optimal classifier for each sample at inference, enabling targeted predictions tailored to each instance's characteristics.

This approach builds an adaptable, sequential ensemble that enhances performance with each additional layer.



References

[1] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models.
[2] Kevin Woods, W. Philip Kegelmeyer, and Kevin Bowyer. Combination of multiple classifiers using local accuracy estimates.
[3] Eulanda M Dos Santos, Robert Sabourin, and Patrick Maupin. A dynamic overproduce-and-choose strategy for the selection of classifier ensembles.
[4] Rafael MO Cruz, Robert Sabourin, George DC Cavalcanti, and Tsang Ing Ren. Meta-des: A dynamic ensemble selection framework using meta-learning.
[5] Albert HR Ko, Robert Sabourin, and Alceu Souza Britto Jr. From dynamic classifier selection to dynamic ensemble selection.
[6] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting.

Fire Feedback Zone

Contact info

jakub.piwko2.stud@pw.edu.pl
jedrzej.rucinski.stud@pw.edu.pl
dawid.pludowski.stud@pw.edu.pl
antoni.zajko.stud@pw.edu.pl
franciszek.filipek.stud@pw.edu.pl
KN-Modelowania-Matematycznego/hellsemble

Acknowledgements

Work on this project is financially supported by the Warsaw University of Technology as a part of the *Inicjatywa doskonałości - uczelnia badawcza* program, through a grant program for scientific clubs for the years 2024–2025.