Abstract

Author: **Bogdan Gulowaty**
Title: Building transparent classification models

As Artificial Intelligence (AI) and Machine Learning (ML) technologies advance and become deeply integrated into daily life, the need for transparent and interpretable models grows increasingly urgent. AI systems must be understandable, trustworthy, and ethical, especially in critical healthcare, finance, and legal sectors. The rise of Explainable Artificial Intelligence (XAI) seeks to address these challenges by providing explanations of model decisions, making AI systems more transparent. However, much of the progress in XAI has focused on deep neural networks, leaving other complex models like ensemble methods needing to be explored more regarding their interpretability.

The thesis aims to fill that gap by developing novel methods to improve the transparency and interpretability of ensemble classifiers while ensuring these models maintain competitive predictive performance and build inherently transparent models. The central hypothesis of the thesis is that it is possible to construct such transparent or explainable models that perform as well as black-box models in a wide range of classification tasks. The work focuses on three primary methods designed to either explain or replace complex models with transparent alternatives: Non-overlapping Tree Ensemble (NOTE), Optimal Centroids and Quad Split algorithms. The thesis sets out with the following objectives:

- Develop novel algorithms that produce transparent classification models.
- Extract interpretable models from complex ensemble classifiers like Random Forest (RF).
- Use data complexity metrics to assess and improve the performance of transparent models.
- Evaluate the effectiveness of these methods across datasets of varying complexities, ensuring the models are both interpretable and accurate.

The main contribution of the thesis is the introduction and evaluation of three novel algorithms in the domain of XAI:

NOTE The NOTE method is specifically designed to explain tree-based ensemble models like Random Forests. While traditional ensemble models can combine transparent base models like Decision Trees, they often produce overlapping decision boundaries, making the ensemble challenging to interpret. NOTE addresses this by selecting a subset of trees that create non-overlapping decision areas, simplifying the overall model. The proposed method applies graph analysis techniques to identify which rules, coming from Decision Trees in the ensemble, may be used in a way that does not overlap and provide the

best generalizing abilities. The areas covered by selected rules are assigned Decision Trees that are trained and combined to form a transparent version of the original model, allowing users to understand individual predictions more easily. Experimental results showed that NOTE can significantly improve interpretability without significantly losing predictive performance. In datasets of moderate complexity, NOTE performed comparably to the original RF while providing much more precise insights into the decision-making process.

Optimal Centroids The Optimal Centroids algorithm focuses on creating interpretable classification models by splitting the feature space into regions based on centroids and 1-Nearest-Neighbours (NN) classifier. In the method, the feature space is divided according to centroids found by evolutionary algorithms, such as genetic algorithms. Centroid positions are being evaluated to find their best distribution, which would maximize the model accuracy. As part of the evaluation step, transparent models are trained for decision space designated by every centroid. The model created in such a way has high interpretability and solid predictive abilities.

Quad Split The Quad Split is a novel algorithm for creating transparent classification models that has been introduced to compete with more complex models by creating interpretable decision boundaries. The idea behind the Quad Split algorithm is based on the assumption that multiple specialized models applied to parts of the whole dataset will perform better than one. To achieve that, Quad Split recursively splits the decision space based on feature values extracted from the training set. Those values serve as splitting points. The algorithm evaluates those points to find one minimizing data complexity at both sides of the split. Then, when stop criteria are reached, the interpretable model is created in every split, finally resulting in an ensemble of transparent models, which could be reduced to a list of simple rules.

Experimental Evaluation The thesis conducted a comprehensive evaluation of the three proposed methods using 16 binary datasets with varying complexities. These datasets included real-world and synthetic examples, covering a range of domains and challenges such as class imbalance, high-dimensional spaces, and noisy data. Complexity metrics from various categories, such as feature, dimensionality, and linearity-based, were evaluated. The experimental setup focused on several key evaluation metrics:

- Predictive abilities: The performances of each model, as defined by balanced accuracy, F1 score, and geometric mean score, were compared against standard classifiers like Random Forests, Decision Trees and rule models, focusing on overall prediction accuracy.
- Model complexity: Base and proposed model complexities were quantified and evaluated against each other.
- Effectiveness as explainers: The ability of the models to explain traditional black-box models like Random Forests was also assessed. The experiments evaluated how well the proposed methods could serve as surrogate explainers for these complex models.

All the above qualities were evaluated in the domain of changing internal algorithm parameters, as well as different dataset complexities - defined by the aforementioned complexity metrics.

Conclusion and Future Work The thesis contributes to the science of AI by demonstrating that transparent and explainable models can be built without sacrificing predictive abilities, particularly in moderately complex datasets. The proposed methods — NOTE, Optimal Centroids, and Quad Split — offer flexible and interpretable alternatives to traditional black-box ensemble models. The main achievements of the thesis are:

- The development of two novel algorithms that create inherently transparent models while maintaining competitive performance.
- The development of the forest ensemble explaining algorithm that is competitive to the explained RF. Extensive experimental evaluation shows that the proposed methods can explain and, in some cases, replace black-box models with transparent alternatives.
- A demonstration that data complexity metrics play an important role in determining when transparent models are most effective.
- Development of a programming library containing the algorithms mentioned above.

The three proposed methods offer significant improvements in interpretability, making them suitable for applications where trust and transparency are essential, such as in healthcare, finance, or legal systems. Along with that, several observations are made, such as that:

- Evaluated classification models behave immensely differently when applied to datasets with different complexity properties.
- Internal model complexities do not always correlate to dataset complexities.
- Evaluated explaining methods that used knowledge extracted from complex models behaved better than inherently transparent ones when used as explainers.

Future research could focus on further refining the proposed methods in various ways, such as fine-tuning the Genetic Algorithm (GA) parameters of Optimal Centroids or finding better evaluation metrics for NOTE. Additionally, the applicability of the algorithms could be checked in a wider range of complex, black-box methods, such as neural networks. Observations made in the thesis might also help develop meta-algorithms based on datasets' complexity metrics. The thesis makes significant contributions to the field of XAI by demonstrating that transparent models can be built without sacrificing predictive abilities.