Proceedings of Machine Learning Research 176:288–291, 2022 NeurIPS 2021 Competition and Demonstration Track

# AIMEE: Interactive model maintenance with rule-based surrogates

**Owen Cornec** O.CORNEC@IBM.COM
**Rahul Nair** RAHUL.NAIR@IE.IBM.COM
**Öznur Alkan** OALKAN2@IE.IBM.COM
*IBM Research Europe*

**Dennis Wei** DWEI@US.IBM.COM
*IBM Research, Yorktown Heights, NY, USA*

**Elizabeth M. Daly** ELIZABETH.DALY@IE.IBM.COM
*IBM Research Europe*

**Editor:** Douwe Kiela, Marco Ciccone, Barbara Caputo

## Abstract

In real-world applications, such as loan approvals or claims management, machine learning (ML) models need to be updated or retrained to adhere to new rules and regulations. But how can a new model be built and new decision boundaries be formed without having new training data available? We present the AI Model Explorer and Editor tool (AIMEE) for model exploration and model editing using human understandable rules. It addresses the problem of changing decision boundaries by leveraging user-specified feedback rules that are used to pre-process training data such that a retrained model will reflect user changes. The pre-processing step uses synthetic oversampling and relabeling and assumes black-box access to the algorithm that retrains the model. AIMEE provides interactive methods to edit rule sets, visualize changes to decision boundaries, and generate interpretable comparisons of model changes so that users see their feedback reflected in the updated model. The demo shows an end-to-end solution that supports the full update lifecycle of an ML model.

**Keywords:** Interactive AI; Interpretability; Surrogate models

## 1. Introduction

Machine learning (ML) models may involve decision boundaries that change over time due to updates to rules and regulations, such as in loan approvals or claims management. But how do we build a new model that reflects the new rules or decision boundaries without having to wait for sufficient new training data to capture the updated information? We present the AI Model Explorer and Editor (AIMEE) system for exploration and model editing using human-interpretable rules. The rules serve as surrogates of arbitrary ML classification models and aim to convey global behaviour of models. Users can then interact with these rules, propose new ones, or edit existing ones. Any user-specified feedback is used to pre-process the training data, such that a retrained model better reflects decision boundaries implied by user-input rules.

Several challenges need to be effectively addressed to enable such a system:

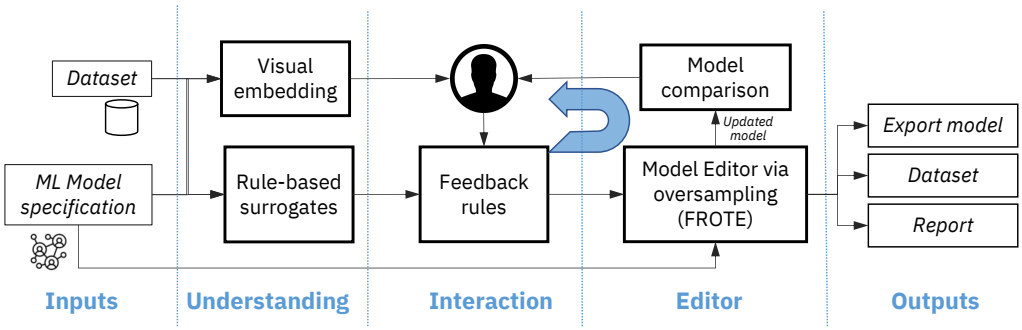


Figure 1: Overall flow of the AIMEE system

- **Interpretable surrogates:** To enable users to interact with the model, we first generate a global surrogate that reflects the model's decision logic. The logic is in the form of Boolean decision rules and generated using an algorithm described in Dash et al. (2018). This produces a compact representation of the model behavior that users can assess and interact with.

- **Real time user interaction and modification of ML decision boundaries:** After inspecting the model logic through decision rules, AIMEE allows users to add new rules to change the decision boundaries through using a pre-processing algorithm. This underlying algorithm described in Alkan et al. (2022) first selects a base population from the original dataset based on the user-provided rules. It then generates instances using this base population through a method similar to the well-known oversampling technique, SMOTE (Synthetic Minority Oversampling Technique) (Chawla et al., 2002). Once the data is pre-processed to reflect new rules, AIMEE retrains the model using the new dataset, and the users can inspect the new decision boundaries.

- **High dimensional visualization of ML decision boundaries:** To allow users to view the impact of their changes, we employ Gower's distance, a type-agnostic similarity metric (Gower, 1971), which is then passed through UMAP (Uniform Manifold Approximation), a dimensionality reduction technique (McInnes et al., 2018). Users can then visualize a 2D clustered map of all rows in the data set. Each point is then assigned an area using a Voronoi diagram and each area is colored by the point's class in order to display a decision boundary. This boundary can be generated irrespective of the number of columns and supports categorical and numeric data types.

- **Model comparison:** In the final step, a new model that reflects user-provided rules can be compared with the previous versions of the model using methods described in Nair et al. (2021) - which allow for pairwise comparison of models based on their surrogate rules. This provides an interpretable way to inspect if changes accurately reflect the user-provided feedback. Since changes are proposed using rules, the comparison also provides a rule set.

These functions are brought together in a user-centric AI application as shown in Figure 1.

## 2. Demonstration

(a) Representative rules with support percentage (left) and support weight tree map (right)

(b) Two-dimensional boundary viewer (left) and feedback rule builder (right)

(c) Model comparison selection matrix (left) and support overlap comparison (right)

Figure 2: Different views from AIMEE.

The methods for model exploration and editing are combined in an interactive demo for a prototypical classification model[1]. The system works as follows:

1. **Set up:** AIMEE ingests structured datasets, i.e. column based data, either categorical or continuous, with target variables with two or more classes. Columns names and feature attributes are treated as human-interpretable and rule sets are derived directly from this data.

2. **Rule inspection:** Figure 2 shows the rule inspection interface. Our rule induction method generates clauses in disjunctive normal form (DNF), i.e. conditions expressed as ORs of ANDs. The view shows support for each conjunction independently. Rules have support within the data set, i.e. instances which agree (green) or disagree (red) with a given rule. The user can inspect the generated rule set for their given data set and assess its quality. For each rule, the user can observe rule support in percentages (left) or in support weight (right). Rules with larger overall support are more representative of a given data set.

1. Demo available here: http://aimee.eu-gb.mybluemix.net/static/index.html

3. **Boundary inspection and editing:** AIMEE employs a pre-processing algorithm to modify a given data set based on a user-generated feedback rule. Upon creating a new rule based on their domain knowledge, users can submit that rule and the underlying algorithm will augment the dataset to reflect the new rule, which will then be used to retrain the ML model. To facilitate the inspection of those changes, we provide a multi-dimensional boundary viewer. This employs Gower's distance to generate a dissimilarity matrix for a data set, which is then passed on to UMAP to generate a 2D embedding of all instances in the data. Given the label of each of these instances, a Voronoi map is applied to the embedding and colored by an instance's class label. Upon sending a feedback rule, a new boundary embedding for the updated dataset is generated. User can toggle between these two boundaries and compare the evolution of the original decision boundary. For each submitted feedback rule, a new model, edited data set and a representative rule set is saved for comparison.

4. **Model comparison:** The final step of the demonstration looks at the rule surrogates for all edited models and visualizes semantic similarity across model surrogates to validate how revised models reflect feedback (Nair et al., 2021). Users can interact with any combination of edits and view rule sets for each label along with support for each decision rule.

## References

Öznur Alkan, Dennis Wei, Massimiliano Matteti, Rahul Nair, Elizabeth M Daly, and Diptikalyan Saha. FROTE: Feedback rule-driven oversampling for editing models. In *Conference on Machine Learning and Systems (MLSys)*, 2022. arXiv preprint arXiv:2201.01070.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, June 2002. ISSN 1076-9757.

Sanjeeb Dash, Oktay Gunluk, and Dennis Wei. Boolean decision rules via column generation. *Advances in Neural Information Processing Systems*, 31:4655–4665, 2018.

J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27: 857–871, 1971.

L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February 2018.

Rahul Nair, Massimiliano Mattetti, Elizabeth Daly, Dennis Wei, Öznur Alkan, and Yunfeng Zhang. What changed? interpretable model comparison. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, {IJCAI-21}*, pages 2855–2861, 2021.