Re-interpreting Rules Interpretability

Linara Adilova (linara.adilova@ruhr-uni-bochum.de)
Ruhr University Bochum
Michael Kamp
IKIM, University Medicine Essen
Gennady Andrienko
Fraunhofer Institute IAIS
Natalia Andrienko
Fraunhofer Institute IAIS

Research Article

Keywords: interpretability, descriptive model, global explanation, generalization

Posted Date: April 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1525944/v1

Re-interpreting Rules Interpretability

Linara Adilova[1,2*], Michael Kamp[1,3,4], Gennady Andrienko[2,5] and Natalia Andrienko[2,5]

[1*]Ruhr University Bochum, Bochum, 44801, Germany.
[2]Fraunhofer Institute IAIS, Sankt Augustin, 53757, Germany.
[3]IKIM, University Medicine Essen, Essen, 45131, Germany.
[4]Monash University, Melbourne, VIC 3800, Australia.
[5]City, University of London, London, EC1V 0HB, UK.

*Corresponding author(s). E-mail(s): linara.adilova@ruhr-uni-bochum.de;

Abstract

Trustworthy machine learning requires a high level of interpretability of machine learning models, yet many models are inherently black-boxes. Training interpretable models instead — or using them to mimic the black-box model — seems like a viable solution. In practice, however, these interpretable models are still unintelligible due to their size and complexity. In this paper, we present an approach to explain the logic of large interpretable models that can be represented as sets of logical rules by a simple, and thus intelligible, descriptive model. The coarseness of this descriptive model and its fidelity to the original model can be controlled, so that a user can understand the original model in varying levels of depth. We showcase and discuss this approach on three real-world problems from healthcare, material science, and finance.

Keywords: interpretability, descriptive model, global explanation, generalization

Statements and Declarations

The authors declare that they have no conflicts of interest. This work was supported by the Cancer Research Center Cologne Essen (CCCE). The authors have no relevant financial or non-financial interest to disclose. This research does not require ethics approval and uses only publicly available data sources that are appropriately referenced. The code used in this manuscript is available at https://github.com/gandrienko/TapasExplTreeViewer. The authors contributions are: Adilova L.: text, experiments; Kamp M.: text, experiments; Andrienko G.: idea, tool implementation, analysis; Andrienko N.: idea, tool implementation, analysis, text.

1 Introduction

One of the key challenges for machine learning (ML) models to be adopted in critical applications, such as autonomous driving and healthcare, is that the model must be explainable [1]. The explainability is not only demanded by practitioners, but is in fact required by law in the EU with the European Parliament's General Data Protection Regulation (GDPR) introducing the right to receive explanations of decisions made by AI systems. There are two different types of explanations: (i) local explanations, i.e., a justification for an individual decision, also termed post-hoc explanation [1], and (ii) a global explanation of the overall logic and behavior of a model. The latter one is often a generalization of the former, since from such an understanding of the model individual decisions can be justified as well.

The usual way to have a global explanation is to use a model that inherently allows such an understanding. Typical examples are decision trees, or rule ensembles. Studying the rules, or equivalently the paths in the decision tree, allows a user to understand the logic of the model, as well as to justify individual predictions. If instead an existing model that is not inherently explainable (i.e., a deep neural network) needs a global explanation, then it can be obtained by training an explainable mimic-model [2] that approximates the black-box model's behavior.

For explainable models to achieve high predictive quality often requires them to be very large in terms of their number of rules. This also holds for mimic models that aspire to achieve high fidelity to the original black-box model. For example, a tree may have hundreds of nodes and tens of levels, and a rule ensemble may consist of hundreds of rules with complex conditions. Therefore, models that are interpretable in principle often remain beyond human perceptual and cognitive capabilities due to their size [3].

A lot of prior research focussed on training compact explainable models, e.g., via special tuning or post-processing [4] in order to reduce the model size and improve its understandability. However, the achievable degree of reduction is significantly constrained by the striving to preserve the prediction accuracy. Even more importantly, such a mimic model will have a different logic than the original model, and its relationship to the original model will be unclear. Hence, instead of explaining how the original model comes to its predictions, the mimic model demonstrates alternative ways to come to the same or similar predictions. While this is, perhaps, the only viable possibility when the original model is a true black box, it may be less desirable when the model logic can, in principle, be understood by a human. In the latter case, a preferable approach would be to facilitate the comprehension of the original model logic rather than to substitute it by another logic.

In this paper, we present an approach to facilitating comprehension of an existing model, representable as a set of conjunction rules (e.g., rule ensembles themselves, decision trees, random forests, tree ensembles), that is explainable in principle but not in practice due to its size. The idea is to extract the general logic from the model by uniting its rules based on their similarity. A union rule not only substitutes multiple original rules, but also typically consists of fewer logical conditions than each of the original rules. Thus, the resulting set of union rules, even with additionally possible exceptions from them, becomes more comprehensible.

Unlike the existing methods that aim at reducing the size of a model while preserving its accuracy on the data, our method creates a new model that *describes* the original model at hand. The original model serves as an input for the algorithm and the output is *a descriptive model of the original model*, having also the form of the set of conjunctive rules. The purpose of this descriptive model is not to make predictions for data instances but to tell how the original model works. Therefore, the descriptive model is not evaluated in terms of the accuracy of its predictions but in terms of its correspondence to the original model. For this purpose, we introduce a novel measure called *Coherence Coefficient* showing how consistent the descriptive rules are with the rules they are intended to describe. This measure allows for a user to regulate the degree of inconsistency of the descriptive model with the model at hand.

Hence, the very idea of our approach is principally different from the ideas behind the existing methods for model simplification that strive to preserve and improve the performance. Our contributions thus are:

- Introduction of an approach that produces a descriptive model of a model that is explainable in principle but too large for comprehension for the purpose of facilitating the understanding of the model logic.
- The achievable degree of simplification is not restricted by the requirement to preserve the prediction quality of the original model, different from the multitude of known approaches for training more compact rule-based models.
- Union rules of the descriptive model can be explored in detail by tracing the hierarchy of more specific rules that were involved in the derivation of the union rules.
- The construction of the descriptive model is fully transparent, and its relationship to the original model is absolutely clear.

The remainder of this paper is organized as follows. We first discuss the relation of the proposed approach to training compact (mimic) models in Section 2. We then present our approach in Section 3, followed by exploring the algorithm via visualizations on a typical application in Section 4. We empirically evaluate the approach in Section 5 and conclude by a discussion of the contribution and its limitations, as well as future work in Section 6.

2 Related work discussion

In ML, certain types of models, namely, decision trees, decision tables, and rules are considered to be inherently interpretable [2], as they can be represented in a human-readable form. However, the actual comprehensibility of such a model greatly depends on its complexity [5, 6], which is typically roughly estimated in terms of the model size [2]. Therefore, the existing ML algorithms that generate decision trees or rules usually strive to reduce the model size by pruning the tree or compressing the set of rules (e.g., RuleFit [7]) so that the smaller model is still as accurate as the big one. Making models more compact is a vast area of research mainly due to the expected improvements in generalization and stability properties of the obtained solutions. Al-Akhras et al. [8] discuss a popular approach to avoiding overfitting in decision trees in which a more compact tree is produced via reducing the amount of instances used to build the model. The approaches directed to reducing the amount of instances were surveyed by Wilson and Martinez [9]. Pruning of decision trees is another popular way to achieve higher stability [10]. Helmbold and Schapire [11] propose an alternative algorithm that avoids pruning. Compactness of sets of rules is also a matter of concern. Dash et al. [12] propose an algorithm for creating compact whilst sufficiently accurate sets of rules using integer programming. In general, enforcing sparseness of the learned rules is a popular problem addressed by, e.g., Su et al. [13]. Alternative approaches propose a different interpretable class of models that is trained in a way to be sparse [14, 15].

The research on compressing intelligible models is mostly based on the regularization techniques applied while training. Thus, Joly et al. [16] propose to use L1 compression for random forests in order to decrease the prohibitively long computation time for the big forests. Alternatively, Painsky and Rosset [17] propose to encode a random forest in a lossless or lossy with guarantees way, which allows not to store the full models—motivated by the limitations of the storage space. Sometimes an interpretable by design model is even compressed into a black box model, like a neural network [18], in order to sustain the small storage space and high performance. In general, aforementioned works aim at achieving more stable, smaller and better generalizing intelligible models while training. Big restriction to the degree of the compactness is always final performance of the model [19].

On the other hand, while the creation of rule-based mimic models is a typical approach to explaining the behavior of black-box models, such as neural networks [2], the research on improving the thereby obtained explanations is on-going. It is clear, that applying pruning or other compression training techniques when trying to mimic a complex black-box model will lead to a loss in fidelity. To achieve both goals, Qiao et al. [20] recently proposed a novel approach in which a set of decision rules is generated by a neural network with a special two-layer architecture. The authors also proposed a sparsity-based regularisation approach to balance between classification accuracy and the simplicity of the derived rules. For now this is a limited approach, that does not allow to work with any black-box model at hand.

So, current research, on the one hand, acknowledges the problem of comprehending large rules sets or decision trees, on the other hand, does not consider the possibility of creating approximate simpler descriptive models instead of directly training more compact ones. Note that creating a descriptive model that helps to interpret the original model is fundamentally different from training a more compact model with similar accuracy. The descriptive model seeks to explain a given model at hand, while training a different, more compact model seeks to replace it and makes it much harder to connect functionality of the initial black-box with the interpretable and compact mimic model.

Freitas [21] discusses that decision trees and rules have different properties in terms of interpretability and that decision trees are usually perceived better when transformed to rules sets. This is also confirmed by Quinlan [22], who considers multiple approaches to pruning decision trees and finalises with the transformation to rules as a help for understanding. A random forest model consisting of multiple trees can also be transformed to a set of rules, for example, using a novel approach from Bénard et al. [23], which is close to the RuleFit [7]. Furthermore, it is argued that a representation in the form of rules can be more compact than a decision tree, because rules can include only significant clauses and have no repeated occurrences of the same variable [2]. Another work [24] discusses high redundancy in decision trees and proposes a method for extracting non-redundant rule-like explanations from a decision tree. The arguments about advantages of rules over trees substantiate the focus of our research on sets of rules.

Since our approach involves unions of rules, it is partly related to the works where rules or decision trees are merged for various purposes. Hierarchical merging of several trees was addressed in the context of the problem of learning decision trees from multiple sources of the data—so the challenge is to produce one tree that will cover the decisions of others [25]. Another problem that is addressed is construction of consensus trees from different ones with the goal of producing a more stable model [26]. A framework for combining multiple rule-based models that have been created for different subproblems is proposed by Strecht et al. [27]. Rules from different models are combined by computing their intersections. After resolving conflicts, the resulting rules set is minimised by uniting nearly-identical rules. A similar approach to joining rules is taken by Andrzejak et al. [28]. Our approach also involves an operation of rule union, but, unlike others, it allows controlled decrease of rule accuracy for achieving a higher degree of simplification.

Our research involves not only the development of an algorithmic method to obtain a descriptive model, but also the creation of interactive visual techniques for exploring sets of rules and investigating the behavior and results of the algorithm. Combining computational methods with interactive visual interfaces is at the core of Visual Analytics (VA) [29]. In particular, VA techniques allow human experts to be involved in the creation of ML models [30]. This way, humans can contribute not only their background knowledge, but also new knowledge gained in the process of interactive data analysis [31] through discovery and abstraction of patterns existing in data [32]. Currently, the problem of explaining ML/AI models is receiving much attention in VA [33]; however, the techniques proposed so far address mostly the needs of model developers rather than domain experts. As an exception, RuleMatrix [34] visualizes rule sets for users with little machine learning experience, but it does not address the problem of model simplification and is severely limited by the size of the rule set.

In the area of visualization research, a comparative evaluation of four basic techniques for visual representation of rules sets, namely, symbolic and graphical encoding of conditions with and without vertical alignment of conditions referring to the same features, has been conducted recently [35]. The experiments showed the superiority of the representations that use feature alignment, which is valid for our table view. Graphical encoding is advantageous to textual, although the effect is less pronounced compared to that of feature alignment. However, the experiments were conducted using small sets of rules, whereas effective visualization of large models is still a challenging task.

3 Rules set simplification

3.1 Main concepts

In the following we define the terms that we will be using throughout the paper. We assume that one wants to interpret a predictive model $h : \mathcal{X} \to \mathcal{Y}$ at hand with input and output spaces $\mathcal{X}$, resp. $\mathcal{Y}$ that can be rewritten as a collection of rules, i.e., $h = \mathcal{R}$, where each rule $R \in \mathcal{R}$ consists of an antecedent which is a *conjunction of conditions* and consequent which is a *prediction* r *of the rule*. The input space $\mathcal{X}$ consists of instances with d features $f_i, i = [d]$, which can be numerical or categorical. The output space $\mathcal{Y}$ can be either categorical for classification or numerical for regression tasks.

A *condition* c is a logical expression of the form $f_i \in V$, where V can be a set of values (for a categorical f_i) or an interval (for a numeric f_i) that is restricting the values that f_i can get. Such c can be a splitting condition from a decision tree node or a part of a conjunctive logical rule.

Definition 1. *A condition* $c_1 = (f_i \in V_1)$ ***subsumes*** *another condition* $c_2 = (f_j \in V_2)$ *iff* $i = j$ *and* $V_2 \subseteq V_1$. *A rule* R_1 ***subsumes*** *or* ***covers*** *rule* R_2 $R_1 \supseteq R_2$, *if every condition of rule* R_1 *subsumes some condition of rule* R_2.

By definition, any rule covers itself. When $R_1 \supseteq R_2$ and $R_1 \neq R_2$, then R_1 is more general and R_2 is more specific. Note that R_2 may include conditions involving features that do not appear in conditions of R_1, i.e., R_1 may have fewer conditions than R_2. For each rule $R \in \mathcal{R}$, where $\mathcal{R}$ is the set of all rules in the model, we can identify set of rules $\mathcal{R}^{\supseteq}$ that are covered by it. When the set of rules $\mathcal{R}$ is optimal in the sense of our approach such sets are trivial $\mathcal{R}^{\supseteq} = \{R\}$.

Definition 2. *Predictions of two rules* R_1 *and* R_2, *denoted by* r_1 *and* r_2, *are* ***congruent*** $r_1 \cong r_2$ *if one of the following conditions holds:*

- $r_1 = r_2$;
- $|r_1 - r_2| \leq \epsilon$ *when* r_1 *and* r_2 *are numbers;*
- $max(r_1^{up}, r_2^{up}) - min(r_1^{low}, r_2^{low}) \leq \epsilon$ *when* r_1 *and* r_2 *are numeric intervals,* $r_1 = [r_1^{low}, r_1^{up}]$, $r_2 = [r_2^{low}, r_2^{up}]$,

where ϵ *is a* ***tolerance threshold*** *used during the run of the algorithm.*

Note, that the last case of intervals is needed for the further work of the algorithm with union rules (defined in the following).

Definition 3. *We say that* $R_1 \supseteq R_2$ *correctly, if* $r_1 \cong r_2$. *In this case the coverage of* R_2 *by* R_1 *is* ***correct****; otherwise, the coverage is* ***wrong****. If* $R_1 \supseteq R_2$ *wrongly, then* R_2 *is an* ***exception*** *of the covering rule* R_1.

Definition 4. *The* ***coherence coefficient*** *(*CC*) of a rule is the ratio of the number of correctly covered rules to the total number of covered rules:*

$$CC(R) = \frac{|\mathcal{R}^{\supseteq correct}|}{|\mathcal{R}^{\supseteq}|}$$

Definition 5. *A rule whose* $CC < 1$, *i.e., a rule having at least one exception, is called a* ***rough rule****.*

Definition 6. *A* ***roughness threshold*** $\rho \in [0, 1]$ *defines the minimal acceptable value of* CC *of a rule included in a descriptive model during the run of the algorithm.*

So, specifying $\rho = 1$ means that no rough rules are allowed, and the smaller ρ gets, the more exceptions rough rules are allowed to have.

For a better understanding of the concept of rule coverage, imagine the multidimensional space of the features (assuming, for simplicity, that all features are numeric). Conditions of a rule antecedent define a multidimensional shape (namely, a rectangular hyper-parallelepiped) in this space. When some feature f_i is not used in a rule explicitly, it can be treated as being involved in an implicit condition $f_i \in V$ where V is the whole range of possible feature values. A rule R_1 covers rule R_2 (Definition 1) when the shape p_1 defined by R_1 includes the shape p_2 defined by R_2. Please note that *any* rectangular parallelepiped p in this space corresponds to some conjunction of conditions, even if there is no rule with such an antecedent. For two or more shapes, it is possible to create a rectangular parallelepiped that encloses all these shapes. The smallest parallelepiped $p^{\cup}$ enclosing the shapes p_1 and p_2 defined

by the conditions of rules R_1 and R_2 represents the *union* of the antecedents of R_1 and R_2.

When we apply the union operation also to the predictions r_1 and r_2 of the rules R_1 and R_2, we obtain a new rule $R^{\cup}$, which is the *union of the rules* R_1 and R_2. The rule $R^{\cup}$ is meaningful only when the predictions r_1 and r_2 are congruent (Definition 2); so, our algorithm makes unions only from rules with congruent predictions.

Accidentally, $p^{\cup}$, apart from p_1 and p_2, may also include parallelepipeds corresponding to antecedents of some other rules; hence, a union $R^{\cup}$ of two rules R_1 and R_2 may additionally cover other rules. Some of those other rules may have predictions incongruent to the prediction of $R^{\cup}$. In such a case, $R^{\cup}$ is a rough rule (Definition 4), and the rules with incongruent predictions are its exceptions (Definition 3).

Let us now define the union of two rules more formally.

Definition 7. *A **union of two conditions** $c_1 = (f_i \in V_1)$ and $c_2 = (f_i \in V_2)$ involving the same feature f_i is the condition $c^{\cup} = (f_i \in V^{\cup})$, where*

- $V^{\cup} = (V_1 \cup V_2)$ *if V_1 and V_2 are sets of discrete values;*
- $V^{\cup} = [min(v_1^{low}, v_2^{low}), max(v_1^{up}, v_2^{up})]$ *if $V_1 = [v_1^{low}, v_1^{up}]$ and $V_2 = [v_2^{low}, v_2^{up}]$ are intervals.*

Definition 8. *A **union of two predictions** r_1 and r_2, denoted $r^{\cup}$, is defined as*

- $r^{\cup} = r_1 = r_2$ *when $r_1 = r_2$,*
- $r^{\cup} = r_1 \cup r_2$ *when r_1 and r_2 are distinct sets of discrete values,*
- $r^{\cup} = [min(r_1^{low}, r_2^{low}), max(r_1^{up}, r_2^{up})]$ *when r_1 and r_2 are numeric intervals, $r_1 = [r_1^{low}, r_1^{up}]$, $r_2 = [r_2^{low}, r_2^{up}]$.*

Definition 9. *A **union of two rules** R_1 and R_2 with* congruent predictions *r_1 and r_2 is a rule $R^{\cup}$ where each condition is a union of conditions from R_1 and R_2 according to Definition 7 and the prediction $r^{\cup}$ is the union of r_1 and r_2 according to Definition 8. Since union is defined for congruent rules, it follows that $R^{\cup} \supseteq R_1$ and $R^{\cup} \supseteq R_2$ correctly.*

3.2 Distance function

In order to perform the hierarchical merging of rules, we define a distance function on the space of rule antecedents. We set the distance between two rule antecedents to be the sum of the distances between the value intervals V of the same feature f_i in the conditions of the rules. So if $c_1 = f_i \in [v_1^{low}, v_1^{up}]$ and $c_2 = f_i \in [v_2^{low}, v_2^{up}]$, then distance between c_1 and c_2 is

$$d_{f_i} = \frac{|v_1^{low} - v_2^{low}| + |v_1^{up} - v_2^{up}|}{2\,(v_{max} - v_{min})}$$

where v_{max} and v_{min} are the absolute maximal and minimal, respectively, values of the feature f_i that may occur in practice. This distance metric is, in fact, a specific formulation of the Hausdorff distance [36] for numeric intervals. The division by $(v_{max} - v_{min})$ is done for normalization of all distances between conditions to the interval $[0, 1]$.

For categorical features, rule conditions contain discrete sets of categorical values instead of numeric intervals. In this case, the distance between two conditions can be defined as the Jaccard similarity index [37] subtracted from 1, i.e., if $c_1 = f_i \in A$ and $c_1 = f_i \in B$, where A and B are sets, then

$$d_{f_i} = 1 - |A \cap B| / |A \cup B|$$

The distance equals 0 when A and B are identical and 1 when the sets have no common elements.

Based on the distances between corresponding conditions, the distance between the rules R_1 and R_2 is $\sum_{f_i} d_{f_i}$, where $f_i \in \{\text{features used in } R_1 \text{ and } R_2\}$. It corresponds to the definition of the Manhattan distance. The interval endpoints are normalised to values between 0 and 1: When some feature is absent in the conditions of one of the rules, it is assumed to have an interval from 0 to 1. Note that since we are not aiming at creating a new compact model that will be used on novel data, this assumption makes sense.

Note that the distance metric is defined solely for the rule antecedents and does not take into account the rule predictions. Since merging is applied only to the rules with congruent predictions, there is no need to include the predictions in the calculation of the rule similarity. Besides, the

distance metric defined in this way can be used for detection of similar rules with incongruent predictions, which may be useful in examining the quality of a rule set.

3.3 Basic algorithm for rules set generalization

Input

- A classification or regression model in the form of a set of rules or a decision tree. In the latter case, the tree is transformed to an equivalent set of rules by one of existing methods (e.g., [38]).
- A roughness threshold ρ (Definition 6).
- Optional: For a regression model, a tolerance threshold ϵ (Definition 2).

Output

- A set of rules such that:
  1. each original rule is correctly covered by some resulting rule (Definitions 1, 3);
  2. the resulting set of rules has smaller cardinality than the original one. In case when the resulting set of rules has the same cardinality as the original one, we say that the algorithm failed;
  3. the coherence coefficient (Definition 4) of any union rule in the resulting set is not less than the roughness threshold, i.e., $CC \geq \rho$ (Definition 6).

The pseudocode of the rules set generalization algorithm is given below (Alg. 1).

The algorithm repeatedly finds the closest (according to the defined distance metric) pair of rules whose predictions are congruent by Definition 2 and applies the operation of rule union (Definition 9). If the united rule has $CC \geq \rho$ (Definitions 4, 6), it substitutes the two rules it was produced from; otherwise, it is discarded. After accepting a new rule, the algorithm searches for the other rules that are correctly covered by this rule (Definitions 1–3) and, if found, removes them from the resulting set. The algorithm terminates when no new union rule was accepted during an iteration.

3.4 Checking Fidelity in Terms of Data Predictions

Since a union rule is more general than the rules it has been derived from, it may be applicable to additional data instances not described by the original rules. For some of these additional instances, the prediction of the union rule may be incongruent with the predictions of corresponding rules from the original model. If we consider some reference dataset, that we have at hand (not necessarily the training dataset used for the original black-box model), we can define the following notion of fidelity.

Definition 10. *The* ***fidelity of a union rule*** *with respect to the original rules set is the ratio of the number of data instances in some reference dataset (e.g., the set from which the model was derived) for which the union rule gives predictions congruent to the predictions of the original model to the total number of data instances this rule is applicable to.*

Definition 11. *The overall* ***fidelity of a descriptive model****, i.e., a generalized rules set, with respect to the original model is the ratio of the number of data instances in the reference dataset for which the descriptive model gives predictions congruent to the predictions of the original model to the total number of data instances both models are applicable to.*

When some set of data instances described by the original rules set is available, the fidelity of the derived union rules to the original predictions can be additionally checked. A reasonable requirement is that the fidelity must not be less than ρ. A condition for checking the fidelity should be added in the "if" statement on line 16 of the Alg. 1, i.e., the extended condition is $CC(R^{\cup}) \geq \rho \wedge fidelity(R^{\cup}) \geq \rho$.

3.5 Iterative lowering of the roughness threshold

There is a possibility to apply Algorithm 1 in an iterative manner. For this purpose, the user specifies an interval $[\rho^{low}, \rho^{up}]$ and a step $\Delta(\rho)$, where $\Delta(\rho) < \rho^{up} - \rho^{low}$. Algorithm 1 is executed several times with consecutively setting the roughness

Algorithm 1 generalization of the set of rules

Input $\mathcal{R}$
Output $\mathcal{R}^{\mathcal{G}}$

$\mathcal{R}^{\mathcal{G}} \leftarrow \mathcal{R}$
changed $\leftarrow True$
while changed **do**
  $PD \leftarrow \varnothing$ ▷ find distances between all congruent rules in the set
  for each $(R_i, R_j) : R_i \in \mathcal{R}^{\mathcal{G}}, R_j \in \mathcal{R}^{\mathcal{G}}, i \neq j$ **do** ▷ apply Definition 2
    if $congruent(r_i, r_j)$ **then**
      $d_{ij} \leftarrow distance(R_i, R_j)$
      $PD \leftarrow PD \cup \{(R_i, R_j, d_{ij})\}$
    end if
  end for
  changed $\leftarrow False$
  while $PD \neq \varnothing \wedge \neg$changed **do**
    $(i, j) \leftarrow argmin_{d_{i,j}} PD$ ▷ find the minimal distance pair
    $PD \leftarrow PD \setminus \{(R_i, R_j, d_{ij})\}$
    $R^{\cup} \leftarrow R_i \cup R_j$ ▷ unite the closest rules according to Definition 9
    if $CC(R^{\cup}) \geq \rho$ **then** ▷ check if the union is acceptable according to Definition 6
      $\mathcal{R}^{\mathcal{G}} \leftarrow \mathcal{R}^{\mathcal{G}} \setminus \{R_i, R_j\}$
      for each $R_k \in \mathcal{R}^{\mathcal{G}}$ **do** ▷ remove all correctly covered rules (Definitions 1–3)
        if $congruent(r_k, r^{\cup}) \wedge R^{\cup} \supseteq R_k$ **then**
          $\mathcal{R}^{\mathcal{G}} \leftarrow \mathcal{R}^{\mathcal{G}} \setminus \{R_k\}$
        end if
      end for
      $\mathcal{R}^{\mathcal{G}} \leftarrow \mathcal{R}^{\mathcal{G}} \cup \{R^{\cup}\}$
      changed $\leftarrow True$
    end if
  end while
end while

threshold ρ to ρ^{up}, $\rho^{up} - \Delta(\rho)$, ..., ρ^{low}, i.e., starting from ρ^{up} and decreasing the threshold in each following run by $\Delta(\rho)$. The output of run i is used as the input of run $i + 1$.

This extension of the method prioritises more coherent rules, i.e., it will strive to produce united rules with higher CC before attempting to achieve higher compression at the cost of reducing the coherence.

To demonstrate possible differences between the results of the basic algorithm and its multi-step variant, Fig. 1 shows two projection displays where rules are represented by dots. The dots are arranged on a plane based on the distances between the rules. The projections have been obtained using the method t-SNE [39]. The dot colours encode the predictions and the sizes are proportional to the number of the data instances the rule applies to. The lines connect dots representing rules that were united by the generalization algorithm. The display on the left corresponds to the base algorithm and the one on the right to the multi-step variant. The dots marked in black represent a group of original rules that were united in a single rule by the multi-step variant and included in three different unions by the base variant.

The illustrations refer to an example classification model consisting of 109 rules including in total 818 conditions. With the roughness threshold of 0.6, the base variant reduces the original set to 54 rules with 342 conditions, of which 33 rules are the same as in the original set (i.e., the algorithm cannot generalize them) and 21 rules are unions obtained from 76 original rules. The coherence coefficient of the union rules ranges from 0.6 to 1; however, only one union has $CC = 1$, three rules

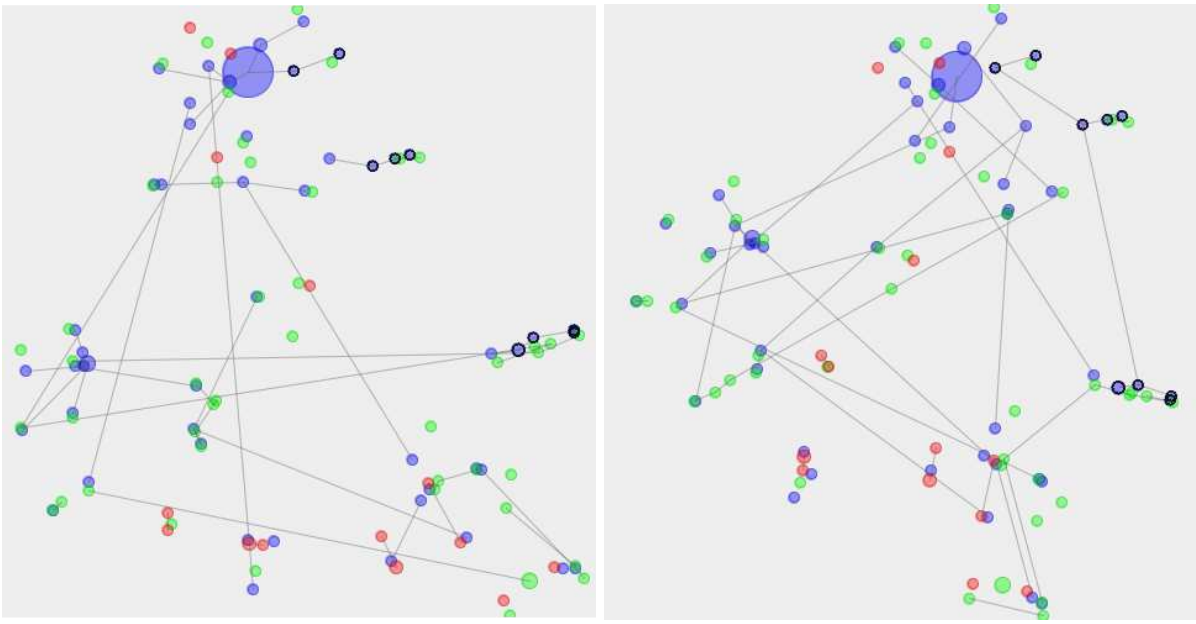


Fig. 1 Two displays demonstrate differences between results of the base algorithm (left) and the multi-step variant (right). The lines connect dots corresponding to rules that were united by the respective variants of the algorithm. The dots marked in black in both displays represent the same selected group of original rules that were united into a single rule by the multi-step variant and included in three different unions by the base algorithm.

have CC from 0.71 to 0.78, and the remaining 17 rules have $CC \leq 0.67$.

The multi-step variant that iteratively lowers the roughness threshold from 1 to 0.6 in steps of 0.05 reduces the original set to 62 rules with 399 conditions, where 43 rules remain the same as in the original set and 19 rules are unions. The coherence coefficient of the union rules ranges from 0.67 to 1, where 5 rules have $CC = 1$, 6 rules have CC from 0.7 to 0.8, and the remaining 8 rules have $CC = 0.67$. Hence, the multi-step variant produces more accurate union rules but achieves a lower degree of generalization and compression than the base algorithm.

4 Visualizations

We have designed and implemented several visualizations that allow researchers to explore rules sets and to investigate the work of our algorithm[1]. This helps to see and interpret the working of the algorithm. Please note that the visualizations are not a part of the rule generalization method and our prototype implementation of the proposed approach is limited to dealing with rules in which the conditions involve only numeric features. However, this limitation does not pertain to the approach itself.

We display a set of rules in the form of a table, as shown in Fig. 2. Each table row corresponds to one rule, each table column corresponds to one feature. Value intervals of the conditions are represented by horizontal bars, which show the relative position of the interval between the minimal and maximal feature values. If a feature is not used in a rule, the corresponding cell is empty. Besides, there in a column entitled "Rule", where each rule is represented as a whole by a glyph with vertical axes corresponding to all available features and vertical bars corresponding to the features used in the rule.

A table showing the results of the rule generalization algorithm (Fig. 3) includes additional columns containing (1) counts of correct and wrong applications of the rule to data instances, (2) counts of correctly and wrongly covered original rules, (3) fidelity, (4) coherence coefficient, (5) number of rules in the derivation hierarchy, and (6) depth of the hierarchy.

[1]Similar visualizations can be used in explaining models to users; however, user-centered design and user evaluation are necessary for ensuring that visualizations are effective, well understood, and easy to use, which is beyond the scope of this paper.

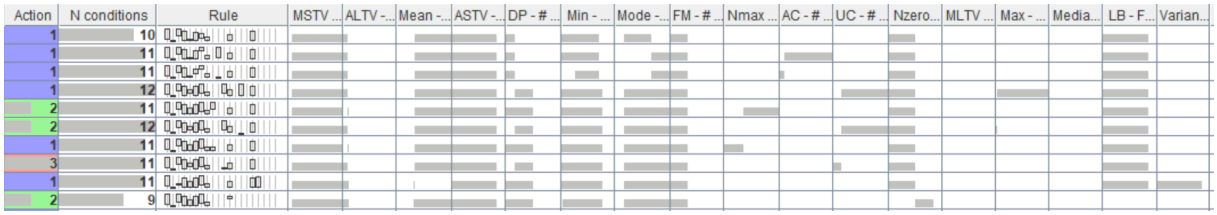


Fig. 2 A fragment of a table representing rules in a graphical form. The rows are ordered by rule similarity using the algorithm OPTICS [40] applying the distance function introduced in Sec. 3.

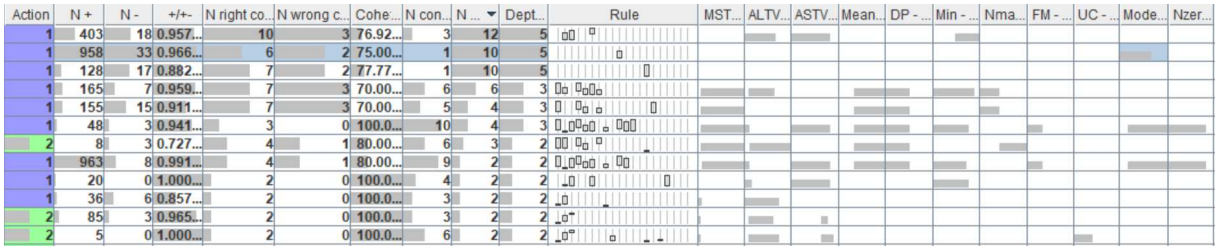


Fig. 3 A fragment of a table representing results of rule generalization.

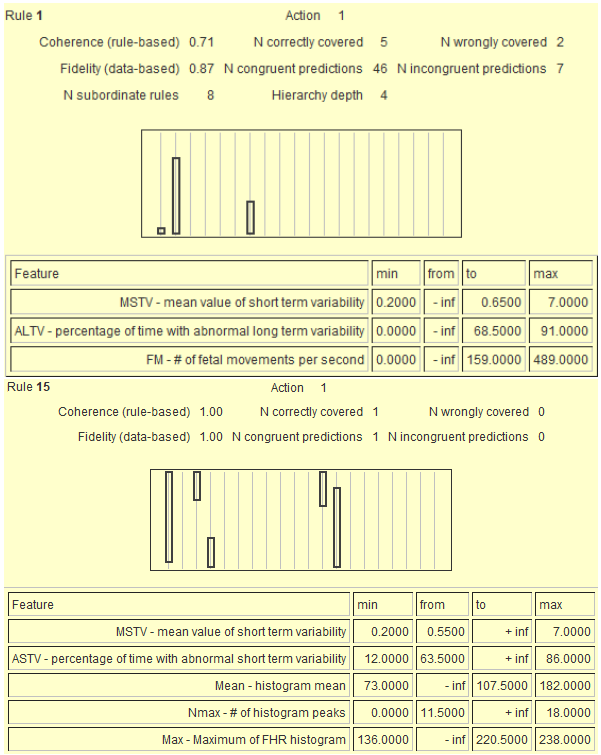


Fig. 4 Examples of popup windows with detailed information about individual rules.

Detailed information about a rule is provided in a popup window (Fig. 4) appearing when the user points on a table row. The window shows all conditions of the rule as well as the minimal and maximal values of all features involved in the rule. Besides, it includes a graphical representation of the rule in a form of a glyph. A glyph representing a rule includes as many vertical axes as there are features used in the whole set of rules (as well as in the "Rule" column of the table). For each feature that is used in the given rule, there is a vertical bar drawn on top of the corresponding axis. The bar represents the interval of the feature values that is specified in the rule condition involving this feature. While the whole length of the axis represents the full range of the feature values (i.e., from the minimal to the maximal values that occur in the dataset), the vertical positions of the lower and upper sides of the bar correspond to the lower and upper ends of the value interval specified in the rule condition.

Another type of display we use is a panel with multiple rules represented by glyphs, as shown in Fig. 5. A popup window with rule details, as in Fig. 4, can be obtained by pointing on a glyph. Comparison of the rules is supported by a mode of glyph drawing in which conditions of one rule are represented by hollow bars with black frames and conditions of one or more previously selected rules are represented by filled semi-transparent bars without frames. The bar shading is darker where feature value intervals from two or more selected rules overlap. For example, in Fig. 5, the conditions of three selected rules (their glyphs have black frames around them) are represented within all glyphs in the display. Rules are selected by clicking on their glyphs or rows in a table view.

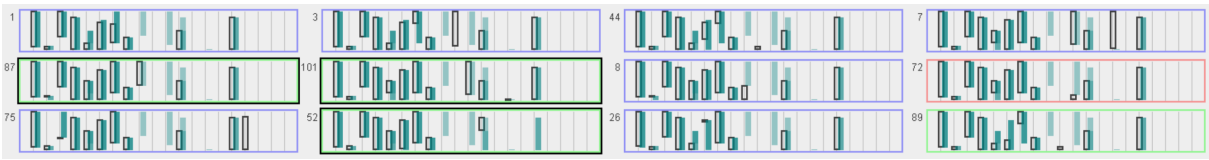


Fig. 5 A display with multiple rules represented by glyphs. The numeric labels are rule identifiers. The colours of the glyph frames encode the predictions made by the rules. Three selected rules are marked by additional black frames. The conditions from the selected rules are represented in all glyphs by semi-transparent vertical bars shaded in cyan.

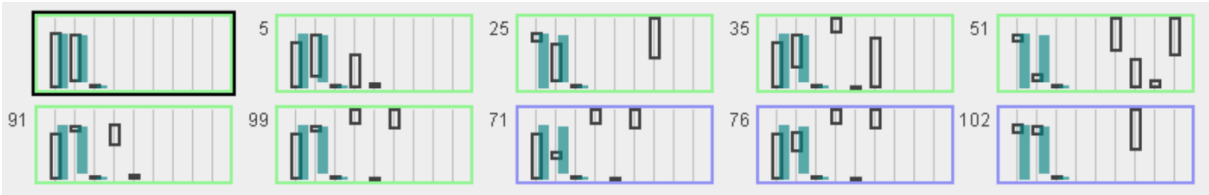


Fig. 6 An illustration of the concept of rule coverage. The colours of the glyph frames represent the predicted classes of the rules. The rule shown on the top left covers the remaining 9 rules, of which 6 are covered correctly and 3 wrongly.

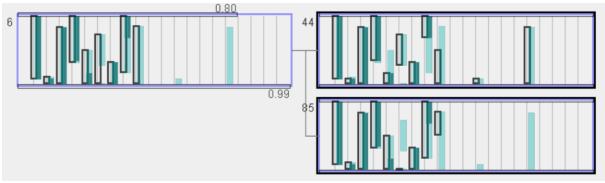


Fig. 7 An illustration of rule union. The glyphs on the right represent two original rules and the glyph on the left their union. The conditions of the original rules are represented by bars shaded in cyan.

Fig. 6 illustrates the concept of rule coverage. Here, the rule shown on the top left (it is selected, so that the glyph is marked with a black frame) covers the remaining rules represented in the image. The conditions of the covering rule are represented in all glyphs by cyan-shaded bars. The colours of the glyph frames encode the predicted classes of the rules. Five of the nine covered rules have the same prediction as the covering rule and three rules have a different one. Hence, five rules are covered correctly and three rules incorrectly.

Fig. 7 illustrates the operation of rule union. Two original rules are shown on the right and their union on the left. The original rules are selected, and their conditions are represented by cyan-shaded bars in all three glyphs. Darker bar shading signifies overlapping conditions. The first four conditions and the seventh condition are identical in the two original rules; so, the same conditions are included in the union rule. In the fifth and eighths conditions, the value interval of one rule includes the value interval of the other rule; so, the union rule includes the larger intervals. In the sixth and ninth conditions, the value intervals do not overlap; so, the union contains the interval from the lower end of the lower interval to the upper end of the higher interval. There are two conditions with features appearing only in one of the original rules. For these features, the union has no conditions.

The numbers 0.80 and 0.99 above and below the glyph of the union rule represent the coherence coefficient and the fidelity of the union rule, respectively. In this example, the union rule covers four original rules correctly and one original rule incorrectly; so, it is a rough rule with $CC = 4/(4+1) = 0.8$. This union rule gives the same predictions as the original model for 963 data instances and different predictions for 8 data instances; hence, its fidelity is $963/(963+8) = 0.99$.

Fig. 8 illustrates the work of the algorithm by example of deriving one generalized rule. Original and derived rules are represented by glyphs. The lines represent inclusions of rules into more general rules covering them. In one of the iteration steps, the algorithm unites original rules labelled 25 and 51 (on the upper right of the image) into a union rule shown in the centre of the upper row of glyphs. In another step, the algorithm unites original rules 91 and 99, which are shown in the lower right corner of the display. In one of the following steps, the algorithm unites the earlier produced union of the rules 25 and 51 (top middle) with an original rule labelled 5 (its glyph is in the middle of the figure). The resulting union

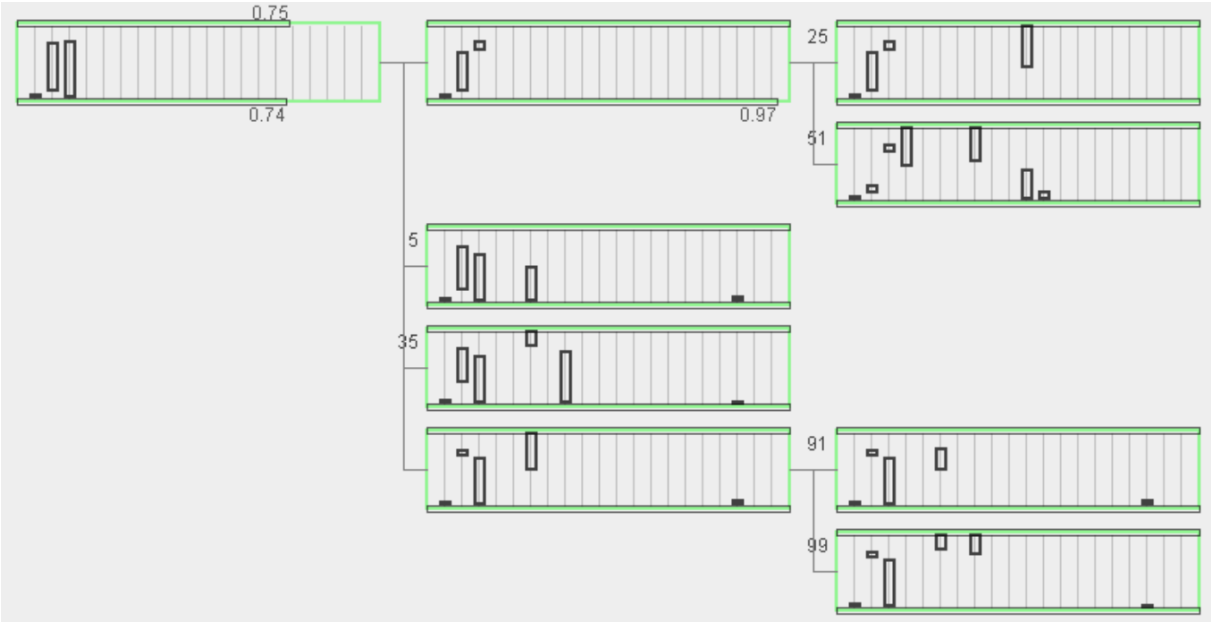


Fig. 8 A representation of the derivation hierarchy of a rule.

rule is shown on the top left. The algorithm finds out that this rule also covers two other rules, an original rule labelled 35 and the earlier obtained union of the rules 91 and 99. The glyphs of these rules are drawn in the central part of the display below the glyphs of the two rules that have been united. Hence, the final union rule generalizes 6 original rules. In deriving it, two intermediate unions have been made. The coherence coefficient of the final rule is 0.75. The rules correctly and wrongly covered by this rule are shown in Fig. 6.

5 Experiments

We describe our investigation of 4 models from three real-world tasks. For each rules set, we ran the basic rule generalization algorithm 9 times setting the parameter ρ to $1.00, 0.95, 0.90, 0.85, ..., 0.60$. Each run was applied to the original rules set. For each run we are analysing the statistics that describe the comprehensibility of the compressed model (the number of the resulting rules, the total number of conditions in all the rules, the mean number of conditions per rule, the number and percentage of the rules including more than 5 conditions, considered as complex), the roughness of the descriptive model (the minimal CC that was actually achieved, the minimal fidelity of a rule, the total fidelity of the whole rules set, and the number of rough rules), and the characteristics of the algorithm work (the number of generated union rules and the maximal depth of a rule derivation hierarchy). While the first group of the results allows to access the interpretability, the second and third ones give a deeper understanding of the mechanics that allows to achieve such compressed descriptive model. It is important to note once again, that our aim is to interpret the global logic of a model at hand, not to understand the data. So we achieve our goal if we can explain the main rules learned by the model, the main features that affect its decisions, the possible outliers that require highly specific rules, etc.

5.1 Cardiocartography dataset

Since medical domain is of high interest for interpretability opportunities, as a primal experiment we looked at a medical dataset. It is an UCI [41] dataset of cardiocartography records [42]. It contains 2126 fetal cardiotocograms for which various diagnostic features were measured. They were classified with respect to a morphologic pattern into 10 classes and to a fetal state into 3 classes. For both cases we directly learned a decision tree and analysed them using our algorithm.

3 classes task

The 3 classes model consists of 109 rules describing 1700 data instances of the training dataset. The statistical characteristics of the generalized rules sets obtained for different settings of the parameter ρ are presented in Fig. 9.

It can be noticed that decreasing the roughness threshold ρ from 1 to 0.85 does not lead to generation of any rough rule, i.e., none of the resulting rules has exceptions. However, the union rules, even when their $CC = 1$, are more general than the original rules and applicable to larger subsets of the data instances, which may include instances with incongruent predictions. Hence, union rules may be fully coherent with regard to the covered original rules, but at the same moment their fidelity may be less than 1.

Another observation is that there are many original rules that cannot be united with others and remain standalone even when the roughness threshold is low. Thus, for $\rho = 0.60$, only 21 out of 54 rules in the resulting model are union rules. Nevertheless, the achievable degree of simplification can be judged as quite high, especially in terms of the number of conditions and the proportion of complex rules with more than 5 conditions. Moreover, such rules help to identify outlier instances that require different logic than most of the other ones. For example, the rule on the right of Fig. 4 describes only one data instance, and it could not be united with any other rule.

An important property of the generalized rules set is that simpler rules (i.e., including fewer conditions) describe a much larger proportion of the data instances than in the original model. So, the minimal number of conditions in one rule is 3 in the original model and 1 in the simplified versions obtained with $\rho = 0.65$ and $\rho = 0.60$ (the maximal number of conditions per rule is 12 in all models). The original model contains 4 rules with 3 conditions describing 47 data instances, 7 rules with 4 conditions describing 62 instances, and 23 rules with 5 conditions describing 163 instances. Taken together, the 34 simpler rules describe 272 data instances out of 1700, i.e., only 16%. In the model obtained with $\rho = 0.65$, the numbers of the rules including from 1 to 5 conditions are, respectively, $2, 1, 6, 9$, and 11, and these 29 rules describe $1009, 36, 201, 200$, and 48 data instances, respectively, i.e., 1494 instances in total. As two or more rules from a generalized model may be applicable to the same data instances, the cumulative number of the data instances correctly (i.e., in congruence with the original model) described by the model with $\rho = 0.65$ is 2709, and thus the simplest rules make 55% $(1494/2709 * 100)$ of the correct descriptions. However, these rules describe 96 data instances incorrectly, i.e., their joint fidelity is $0.94 = 1494/(1494 + 96)$.

Hence, there are multiple aspects of simplification: the number of rules, the number of conditions, the proportion of simple rules, and the proportion of the data described by these simple rules. Moreover, the conditions of the simplest rules applicable to large number of instances indicate which features have higher importance than others. For example, the model with $\rho = 0.65$ contains a rule with a single condition "**If** *histogram mode* < 148.5 **then** class $= 1$" correctly describing 881 data instances and having fidelity 0.97. This rule reveals the importance of the feature *"histogram mode"*. Another example, is that *percentage of time with abnormal short term variability* is rather low for class 1 (healthy), but gets higher for 2 and 3 (suspect and pathology), at the same moment the *histogram mean* is lower for the pathology class, compared to other two.

Additionally we investigated the effects of model pruning on the performance of our algorithm and presented the results in Section 5.4.

10 classes task

This dataset allows us to see the difference between interpretability for simpler and more complex task on the same data features. Decision tree for 10 classes consists of 202 rules describing the same 1700 data instances of the training dataset. The descriptive statistics of the results of the experiments are shown in Fig. 10. As it could be expected, the potential for compression and generalization is lower when the number of classes is higher due to the congruence requirement. Compared to the 3-class model, the 10-class model also consists of more complex rules, i.e., ones that have more conditions. The generalization increases the proportion of simpler rules having up to 5 conditions, which contributes to better comprehensibility, along with the decrease of the number of the rules. The fact that with $\rho = 0.60$ we see

rho	N rules	Total N conditions	Mean N conditions	N rules >5 conditions	% rules >5 conditions	min *CC*	min fidelity	Total fidelity	N rough rules	N union rules	Max depth
	109	818	7.50	75	68.81	1.00	1.00	1.00	0	0	1
1.00	103	762	7.40	68	66.02	1.00	1.00	1.00	0	6	2
0.95	98	708	7.22	63	64.29	1.00	0.95	0.99	0	10	2
0.90	95	678	7.14	61	64.21	1.00	0.91	0.98	0	10	3
0.85	94	678	7.21	62	65.96	1.00	0.85	0.98	0	10	3
0.80	87	609	7.00	54	62.07	0.80	0.81	0.98	3	12	5
0.75	84	594	7.07	52	61.90	0.75	0.75	0.97	5	12	5
0.70	78	542	6.95	47	60.26	0.75	0.71	0.97	7	13	6
0.65	64	415	6.48	35	54.69	0.67	0.67	0.96	17	20	5
0.60	54	342	6.33	28	51.85	0.60	0.62	0.90	20	21	6

Fig. 9 Results of experimenting with the 3-classes classification model trained on the cardiocartography dataset.

many more union rules that are not rough, compared to the 3-class model also confirms that the global logic of the model is more complex.

5.2 Home Equity Line of Credit (HELOC), 2 classes

This example application is based on the Explainable Machine Learning Challenge organised by a group of commercial and academic organisations [2]. Based on an anonymised dataset of applications made by homeowners, the challenge requires creation of a readily explainable model predicting the value of the variable Risk Performance, which may be either "bad" or "good". In order to allow a correct decision tree creation, we excluded records with special values and two categorical features. We first created an obviously incomprehensible random forest model with 50 trees without depth restriction, that achieves perfect accuracy, and then generated a mimic model approximating the behavior of the random forest model. The mimic model consists of 384 rules containing in total 3019 conditions which involve 21 features with numeric value domains. The statistics describing the results of the generalization are presented in Fig. 11.

It can be seen that the mimic model can be slightly simplified even with $\rho = 1$. It means that the model has some redundancies. While increasing the degree of simplification, the total fidelity of the simplified model to the original one decreases gradually but more substantially than it was in other experiments. A probable reason is high similarities between rules giving opposite predictions: when a rule gets more general, it may become applicable to additional data instances that are described by other rules, even if it does not cover those other rules (i.e., the conditions of the rules partly overlap). The projection plot on the left of Fig. 12 supports this guess: blue and red dots representing rules with negative and positive outcomes, respectively, tend to be very close in the plot. An interesting side effect of the simplification is that it increases the separation, i.e., the dissimilarity between the rules with the positive and negative outcomes. This can be seen from comparing the projection of the original rules set on the left of Fig. 12 to the projections of the simplified rules sets obtained with $\rho = 0.85$ (Fig. 12, center) and with $\rho = 0.75$ (Fig. 12, right).

The similarities between rules are demonstrated in Fig. 13, where a table displays a group of rules represented by a cluster of closely positioned dots in the projection plot shown in Fig. 12, left. The cluster has been interactively selected by dragging a frame around it. The table shows that the rules with negative results ($Action = 0$) differ from the closest rules with positive results ($Action = 1$) by just one condition.

Using this example, we can demonstrate how our techniques can be used to answer the question of the challenge organizers: if an applicant who has got a negative result ("bad"), can the model easily explain what should be changed to turn the result to positive ("good")? For this purpose, the rule R^0 that gave the negative result needs to be identified in the projection plot (the localisation of rules is supported by highlighting) and the rules with positive results having close positions in the plot need to be selected, for example, as shown in Fig. 12, left. The rule R^0 can be conveniently compared with the other selected rules using the

[2] See https://community.fico.com/s/explainable-machine-learning-challenge

rho	N rules	Total N conditions	Mean N conditions	N rules >5 conditions	% rules >5 conditions	min *CC*	min fidelity	Total fidelity	N rough rules	N union rules	Max depth
	202	1739	8.61	185	91.58	1.00	1.00	1.00	0	0	1
1.00	197	1682	8.54	177	89.85	1.00	1.00	1.00	0	5	2
0.95	188	1567	8.34	166	88.30	1.00	0.95	0.99	0	14	2
0.90	185	1536	8.30	163	88.11	1.00	0.90	0.98	0	17	2
0.85	183	1518	8.30	161	87.98	1.00	0.88	0.98	0	19	2
0.80	177	1472	8.32	157	88.70	0.83	0.80	0.98	3	19	3
0.75	167	1357	8.13	145	86.83	0.75	0.77	0.95	9	25	4
0.70	163	1326	8.13	139	85.28	0.75	0.70	0.92	10	23	5
0.65	149	1172	7.87	121	81.21	0.67	0.67	0.90	26	36	5
0.60	139	1062	7.64	106	76.26	0.60	0.63	0.85	25	34	5

Fig. 10 Results of experimenting with the 10-classes classification model derived from the cardio dataset.

rho	N rules	Total N conditions	Mean N conditions	N rules >5 conditions	% rules >5 conditions	min *CC*	min fidelity	Total fidelity	N rough rules	N union rules	Max depth
	384	3019	7.86	343	89.32	1.00	1.00	1.00	0	0	1
1.00	351	2603	7.42	287	81.77	1.00	1.00	1.00	0	33	2
0.95	339	2463	7.27	269	79.35	1.00	0.95	0.99	0	42	3
0.90	323	2291	7.09	242	74.92	1.00	0.90	0.97	0	57	3
0.85	313	2195	7.01	231	73.80	1.00	0.86	0.95	0	62	4
0.80	293	2008	6.85	206	70.31	0.80	0.80	0.91	9	65	4
0.75	257	1693	6.59	171	66.54	0.75	0.75	0.85	37	71	5
0.70	234	1497	6.40	146	62.39	0.70	0.70	0.83	38	75	5
0.65	201	1232	6.13	113	56.22	0.67	0.65	0.78	70	86	6
0.60	158	941	5.96	84	53.16	0.60	0.60	0.73	53	67	7

Fig. 11 Results of experimenting with the 2-classes classification model derived from the HELOC dataset.

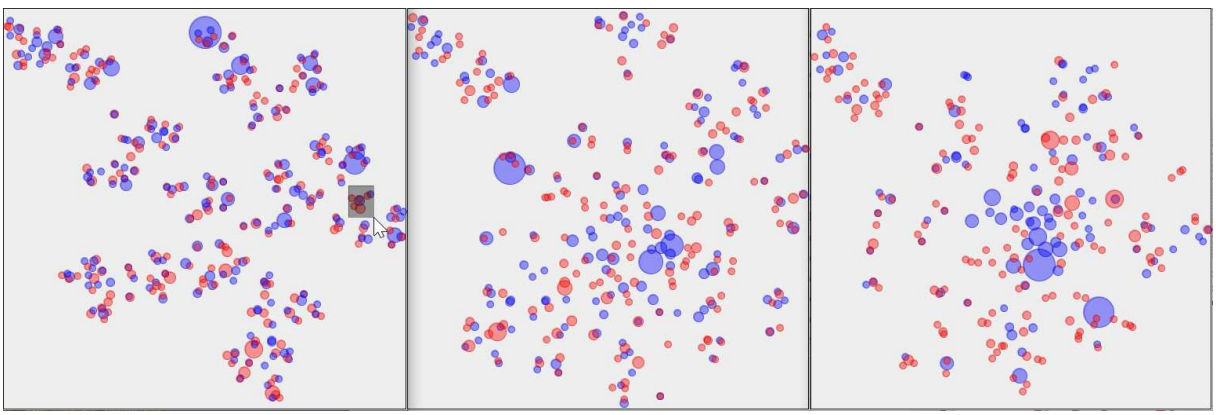


Fig. 12 Similarity-based t-SNE projections of the original HELOC rules set (left) and the simplified versions obtained with $\rho = 0.85$ (center) and with $\rho = 0.75$ (right). Blue dots correspond to rules predicting negative outcome and red to rules with positive predictions.

Action	Rule	Number ...	Consolid...	Percent l...	Number ...	Number ...	Number o...	Net Fracti...	Percent Tr...	Number o...
0										
1										
1										
1										
1										
0										
1										

Fig. 13 A group of rules represented by closely positioned dots in the projection plot in Fig. 12, left (enclosed in a dark grey rectangle) is shown in a table view.

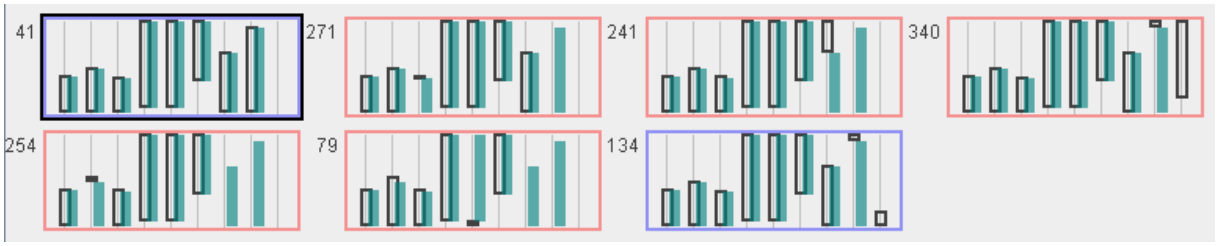


Fig. 14 The same rules as in Fig. 13 are represented by glyphs. The first rule (no. 41) is selected for comparison with the others. One can see how the negative result may be turned to positive by changing the value of just one feature.

table display shown in Fig. 13 or a glyph representation, as in Fig. 14. In this figure, the rule no. 41 is selected as a reference for comparison. Its conditions are represented by cyan-filled bars in all glyphs. It is easy to see that a small increase of the value of the third feature (Percent Installment Trades) will make rule no. 271 with a positive outcome applicable to this case instead of the rule no. 41. Other possibilities are to make rule no. 241 applicable by increasing the value of the 7th feature (Net Fraction Installment Burden), or rule no. 340 by increasing the value of the 8th feature (Percent Trades Never Delinquent), or rule no. 254 by increasing the value of the second feature (Consolidated Version of Risk Markers), or rule no. 79 by decreasing the value of the 5th feature (Number Trades 60+ ever). In a similar way, one can use more general rules of a simplified model version, in which critical features like these can be easier identified.

5.3 Material science, regression

This example is taken from the NOMAD 2018 Kaggle challenge to predict the formation energies and bandgap energies of alloys from transparent conductors [3]. In contrast to the previous examples, this is a regression task.

In material science, a state-of-the-art prediction method is RuleFit [7] which trains a rule ensemble combined with a linear model. Note that the winning methods of the Kaggle challenge (n-gram [43], SOAG [44], MBTR [45]) do not substantially outperform RuleFit on the entirety of the dataset ([46] have shown that they do, however, perform well on well-defined subsets of the data). A prediction is obtained using a weighted sum of rule outputs and feature values. We used RuleFit to train a rule ensemble on the 402 data instances using formation energy as target. The resulting rule ensemble consists of 396 rules. Analyzing these rules, we find that most of them describe single data instances which is not surprising given that the number of rules is nearly equal to the number of data instances. This indicates some potential for compression.

The target values range from 0 to 0.7676. We have set the tolerance threshold ϵ to $0.010, 0.020$ and 0.050, corresponding to the $2\%, 5\%$, and 10% percentiles of target values. The collected statistics are presented in Fig. 15.

The second rows in all three tables demonstrate that the largest part of the simplification is achieved at the cost of decreasing the precision of the predictions from specific numbers to intervals. For example, a union rule predicts that the result will be from 0.2179 to 0.2214 instead of predicting a fixed number like 0.22. Hence, the chosen value of the tolerance threshold ϵ has the highest impact on the resulting degree of simplification and generalization, whereas the impact of the roughness threshold ρ is quite small: the decrease in the number of rules due to decreasing ρ from 1 to 0.6 ranges from only 6% for $\epsilon = 0.010$ to 14% for $\epsilon = 0.050$, and the decrease in the number of conditions ranges from 9% for $\epsilon = 0.010$ to 20% for $\epsilon = 0.050$.

Nevertheless, the potential of our method for simplifying the explanation of regression models can be considered as high. It is quite reasonable to posit that a user rarely needs an exact explanation for each individual numeric value that can be predicted by a model. Rather, the user can be satisfied with a model description telling what combinations of conditions lead to model results fitting in different ranges of values (e.g., high and low). The user-controlled value of ϵ determines how narrow or wide these intervals will be. Thus, by choosing a larger value, a user can

[3] https://www.kaggle.com/c/nomad2018-predict-transparent-conductors

rho	N rules	Total N conditions	Mean N conditions	N rules >5 conditions	% rules >5 conditions	min *CC*	min fidelity	Total fidelity	N rough rules	N union rules	Max depth
	396	3099	7.83	359	90.66	1.00	1.00	1.00	0	0	1
1.00	240	1760	7.33	204	85.00	1.00	1.00	1.00	0	96	6
0.95	240	1760	7.33	204	85.00	1.00	1.00	1.00	0	96	6
0.90	240	1760	7.33	204	85.00	1.00	1.00	1.00	0	96	6
0.85	238	1738	7.30	202	84.87	0.86	0.86	0.99	2	96	6
0.80	237	1726	7.28	201	84.81	0.80	0.80	0.99	3	96	6
0.75	234	1701	7.27	198	84.62	0.75	0.75	0.99	5	96	6
0.70	231	1669	7.23	195	84.42	0.71	0.71	0.97	8	95	6
0.65	230	1659	7.21	194	84.35	0.67	0.67	0.97	9	95	6
0.60	225	1603	7.12	188	83.56	0.60	0.60	0.94	12	93	6

rho	N rules	Total N conditions	Mean N conditions	N rules >5 conditions	% rules >5 conditions	min *CC*	min fidelity	Total fidelity	N rough rules	N union rules	Max depth
	396	3099	7.83	359	90.66	1.00	1.00	1.00	0	0	1
1.00	185	1309	7.08	149	80.54	1.00	1.00	1.00	0	97	8
0.95	185	1309	7.08	149	80.54	1.00	1.00	1.00	0	97	8
0.90	185	1309	7.08	149	80.54	1.00	1.00	1.00	0	97	8
0.85	184	1301	7.07	148	80.43	0.86	0.86	1.00	1	96	8
0.80	182	1282	7.04	146	80.22	0.80	0.80	0.99	3	96	9
0.75	180	1266	7.03	144	80.00	0.80	0.75	0.98	3	96	9
0.70	177	1247	7.05	143	80.79	0.71	0.71	0.96	6	92	9
0.65	175	1224	6.99	140	80.00	0.71	0.65	0.94	7	91	9
0.60	166	1138	6.86	129	77.71	0.60	0.60	0.88	14	87	9

rho	N rules	Total N conditions	Mean N conditions	N rules >5 conditions	% rules >5 conditions	min *CC*	min fidelity	Total fidelity	N rough rules	N union rules	Max depth
	396	3099	7.83	359	90.66	1.00	1.00	1.00	0	0	1
1.00	104	677	6.51	78	75.00	1.00	1.00	1.00	0	70	10
0.95	104	677	6.51	78	75.00	1.00	1.00	1.00	0	70	10
0.90	104	677	6.51	78	75.00	1.00	1.00	1.00	0	70	10
0.85	102	654	6.41	75	73.53	0.86	0.88	1.00	0	70	10
0.80	99	631	6.37	73	73.74	0.80	0.80	0.97	3	67	10
0.75	98	622	6.35	71	72.45	0.80	0.75	0.97	3	67	10
0.70	97	612	6.31	70	72.16	0.74	0.71	0.95	4	66	10
0.65	96	603	6.28	68	70.83	0.68	0.68	0.92	6	65	10
0.60	89	542	6.09	60	67.42	0.60	0.60	0.86	9	61	10

Fig. 15 Results of experimenting with the regression model derived from the material science dataset. The tables, from top to bottom, correspond to $\epsilon = 0.010$, $\epsilon = 0.020$, and $\epsilon = 0.050$, respectively.

obtain a compact description of model behavior even without decreasing the coherence coefficient of the rules. For example, the same material science model generalized with $\epsilon = 0.075$ is described by 80 rules with 494 conditions, and $\epsilon = 0.25$ gives only 22 rules with 94 conditions in total and from 3 to maximum 6 conditions in each individual rule. Like in the other cases, such combinations of conditions, as well as the features they involve, can be considered the most influential for the model result.

This example suggests an interesting possible way of using the generalization method for regression tasks. First, a high-level overview of the behavior of a model is gained by obtaining a very rough (large ϵ) generalized representation of it. Then, subsets of the original rules that have been unified in the result of the generalization are investigated in more detail by applying the generalization method separately to these subsets. For example, the t-SNE projection plot in Fig. 16 shows how the original rules of the material science model were joined in a highly generalized version

of the model. One group of 27 linked rules has been selected and generalized with $\epsilon = 0.05$ to 5 simple rules. The latter are shown in a table view, where the first two columns represent the intervals of the predicted values: $r \in [minQ, maxQ]$.

5.4 Experiments with pruned models

In order to investigate the effect of pruning of the original model on the compression that can be achieved with our algorithm, we created decision trees with different level of cost complexity pruning. In Fig. 17 pruning degree 1 denotes the least compressed model and pruning degree 3 - most compressed model. An immediate observation that can be made is that the result of our algorithm is highly dependent on the pruning performed. Thus, the strongly pruned model is highly resistant to generalization. This means that our method can be useful also for practitioners in order to understand if the model is compact enough and does not contain redundancies. Another interesting observation is that the model 2 can be compressed only with significant roughness: the models obtained with $\rho = 0.90$ and $\rho = 0.80$ are identical, while the simplification attempt with $\rho = 1$ fails (no union rules could be produced). This is even more pronounced for the strongest pruning.

It should be noted that each of these pruned models has progressively declining accuracy when trained, which showcases the difference of our approach compared to pruning: while keeping a required degree of coherence and fidelity to the original model, our method gives a simplified description without any effect on the accuracy of the original model.

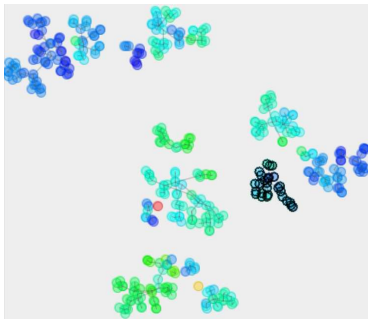
6 Discussion

We proposed an approach to facilitating comprehension of models that are interpretable by design, but too large to be actually intelligible by a human due to cognitive limitations. For this, we explain the logic of a large (in principle) interpretable model by a simplified descriptive model that suits human cognitive properties: while averse to large volumes of information, humans are good in dealing with vague concepts, approximate statements, and fuzzy reasoning. One can think of a data mart[4] as an example of widely used descriptive models in real world: instead of giving a human full data from business, a special high level view is formed in order to understand the processes happening in it.

Our approach differs from the approaches of regularization or compression techniques for obtaining simpler yet accurate enough predictive models, since our goal is not to retrain or improve a model, but to explain it. That is, our aim is to represent the logic of a complicated pseudo-interpretable model at hand. This is independent of whether the model was derived from data or is an interpretable model mimicking the behavior of some black-box model. A descriptive model is not meant to be used as a substitute for the model at hand (i.e., it is not used for making predictions), but its purpose is to provide an explanation for the global logic of that model. The cost of high simplification is loss of predictive accuracy. The more complex the global logic of a model, the harder it is to generalize and represent it by a simple descriptive model of sufficient fidelity. Our algorithm for rule generalization allows users to control how similar to the original model the descriptive model must be in terms of predictions. Besides, the possibility to see the exceptions and the hierarchy of rule generalization allows a human to increase the fidelity (and, hence, the complexity) of the description as desired. In addition to model description, our approach also supports model exploration in terms of important features, their impact on predictions, and which feature combinations would create outliers.

Similar to a mimic model [2] there is a trade-off between interpretability, i.e., size of the descriptive model, and accuracy of the description, i.e., similarity of the descriptive model to the original one. The goal typically is to have the most concise descriptive model that is still sufficiently similar. The similarity, however, is in general hard to assess: A meaningful similarity measure not only depends on the functional similarity, e.g., as measured by a suitable norm on the function space, but also on the expected difference given the data distribution. We use two measures as a surrogate: fidelity to measure difference in predictions and the coherence coefficient to measure structural similarity.

[4] https://en.wikipedia.org/wiki/Data_mart



min Q	max Q	Rule	percen...	perce...	b_c	alpha	ecn_b...	ecn...	ecn...
0.05630	0.08660								
0.12170	0.14760								
0.13960	0.18850								
0.17280	0.21460								
0.22460	0.22640								

Fig. 16 Left: Selection of original rules that contributed to a single generalized rule using a t-SNE projection of the original rules set with lines showing links between rules. Right: A result of generalizing the selected subset consisting of 27 rules with 223 conditions (from 6 to 11 conditions in one rule). With $\rho = 1$ and $\epsilon = 0.05$, the subset can be generalized to 5 rules with 33 conditions; each rule has 6 or 7 conditions.

Pruning degree	rho	N rules	Total N conditions	Mean N conditions	N rules >5 conditions	% rules >5 conditions	min *CC*	min fidelity	Total fidelity	N rough rules	N union rules	Max depth
1		107	807	7.54	78	72.90	1.00	1.00	1.00	0	0	1
1	1.00	104	779	7.49	74	71.15	1.00	1.00	1.00	0	3	2
1	0.90	93	666	7.16	63	67.74	1.00	0.92	0.98	0	10	3
1	0.80	79	533	6.75	48	60.76	0.86	0.80	0.96	4	12	4
1	0.70	77	528	6.86	48	62.34	0.75	0.71	0.96	7	11	5
1	0.60	58	380	6.55	31	53.45	0.60	0.60	0.94	15	17	5
2		48	285	5.94	22	45.83	1.00	1.00	1.00	0	0	1
2	0.90	47	274	5.83	20	42.55	1.00	0.92	0.99	0	1	2
2	0.80	47	274	5.83	20	42.55	1.00	0.92	0.99	0	1	2
2	0.70	43	240	5.58	17	39.53	0.75	0.82	0.99	3	3	2
2	0.60	32	161	5.03	11	34.38	0.60	0.60	0.95	11	12	3
3		26	131	5.04	9	34.62	1.00	1.00	1.00	0	0	1
3	1.00	24	113	4.71	7	29.17	1.00	1.00	1.00	0	2	2
3	0.90	23	103	4.48	5	21.74	1.00	0.99	0.99	0	3	2
3	0.80	23	103	4.48	5	21.74	1.00	0.99	0.99	0	3	2
3	0.70	23	103	4.48	5	21.74	1.00	0.99	0.99	0	3	2
3	0.60	21	96	4.57	5	23.81	0.60	0.80	0.97	2	3	3

Fig. 17 Results of experimenting with the pruned versions of the 3-classes classification model trained on the cardiocartography dataset.

Fidelity is widely used in explainable AI [2] and measures the number of data instances in a reference dataset for which the predictions of the original and descriptive differ by more than ϵ. Fidelity can be inaccurate for two reasons: (i) using a reference dataset as an empirical sample of the data distribution is only an approximation, and (ii) the difference of the two models on those data samples where they do not agree is unbounded.

We introduced the coherence coefficient to measure the number of rules subsumed coherently by a generalization in the description. Rules not covered coherently are kept in the description as exception rules so that no structural parts of the original model are discarded. This guarantees that for all data instances where at least one rule in the original model is satisfied, at least one rule in the descriptive model will be satisfied and their predictions do not differ by more then ϵ. Therefore, the set of points for which at least one rule in the descriptive model is satisfied is a superset of the points for which at least one rule in the original model is satisfied. This measure can still be inaccurate on the points for which a rule in the descriptive model is satisfied but no rule in the original model is. For those points, the difference in prediction is in general unbounded.

For exploring the properties of the algorithm, we created a visualization interface and performed a series of experiments applying the rules generalization to four different models. Our case studies showed that the human interaction for setting the acceptable level of description roughness is very helpful—while significant roughness makes the result easier to comprehend, obtaining several descriptive models with different degrees of roughness can help to refine the understanding of the predictions logic. Interesting enough, the experiment with a regression model showcased that in this case imprecision of predictions allows to achieve higher simplification than rules roughness control. Based on this observation, we propose a method for focused exploration of selected parts of a regression model at hand: starting from a very simple but very imprecise descriptive model, a user selects one of the generalized rules, extracts the subset of the original rules it covers, and obtains a more precise descriptive model for this subset. In this way, the understanding of the model logic can be gradually refined and deepened.

We also found out that the distance metric we introduced can be used for answering the prediction justification questions, i.e., determining what features should be changed and how to make a model change its prediction. Knowing the rule by which the current prediction was made, one uses the distance metric to select the closest rules giving the desired outcome and inspects how their conditions differ from the conditions of the rule that was applied.

Our algorithm allows two variants of use (see Sec. 3) and is open to further extensions. For example, it can take into account possible overlaps (partial coverage) between a generalized rule and the original rules. Currently, the coherence coefficient of a general rule is calculated only from the rules fully covered by it. This definition can be extended in an obvious way to including also partial coverage by an appropriate change in the computation of CC.

An interesting direction for future work is to combine generalization of the rules with merging and generalization of the features involved in the rule conditions, which is expected to enable much higher degrees of model logic simplification. Examples of feature merging can be seen in the award-winning solution of the HELOC Challenge[5] [47], where 6 groups of semantically related original features were integrated into composite features thereby reducing the original 23 features to 10 features. Such feature merging is usually hard to perform in the interpretable manner without domain knowledge and human reasoning. However, it may be possible to detect automatically (by analyzing a rule set) which features are likely to be related and propose groups of such features to a human expert for considering and controlling integration. This can significantly strengthen the comprehensibility of a descriptive model.

References

[1] Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. San Diego, California: Association for Computational Linguistics; 2016. p. 97–101.

[2] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys. 2018 Aug;51(5). https://doi.org/10.1145/3236009.

[3] Kovalerchuk B, Ahmad MA, Teredesai A. Survey of explainable machine learning with visual and granular methods beyond quasi-explanations. Interpretable Artificial Intelligence: A Perspective of Granular Computing (Eds W Pedrycz, SM Chen), Springer. 2021;937:217–267.

[4] Letham B, Rudin C, McCormick TH, Madigan D. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. The Annals of Applied Statistics. 2015;9(3):1350–1371.

[5] Arya V, Bellamy RK, Chen PY, Dhurandhar A, Hind M, Hoffman SC, et al. One explanation does not fit all: A toolkit and taxonomy

[5] https://community.fico.com/s/blog-post/a5Q2E0000001czyUAA/fico1670

of ai explainability techniques. arXiv preprint arXiv:190903012. 2019;.

[6] Huysmans J, Dejaeger K, Mues C, Vanthienen J, Baesens B. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decision Support Systems. 2011;51(1):141–154.

[7] Friedman JH, Popescu BE. Predictive learning via rule ensembles. The Annals of Applied Statistics. 2008;2(3):916–954.

[8] Al-Akhras M, El Hindi K, Habib M, Shawar BA, et al. Instance reduction for avoiding overfitting in decision trees. Journal of Intelligent Systems. 2021;30(1):438–459.

[9] Wilson DR, Martinez TR. Reduction techniques for instance-based learning algorithms. Machine learning. 2000;38(3):257–286.

[10] Esposito F, Malerba D, Semeraro G, Kay J. A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1997;19(5):476–491. https://doi.org/10.1109/34.589207.

[11] Helmbold DP, Schapire RE. Predicting nearly as well as the best pruning of a decision tree. Machine Learning. 1997;27(1):51–68.

[12] Dash S, Gunluk O, Wei D. Boolean Decision Rules via Column Generation. Advances in Neural Information Processing Systems. 2018;31:4655–4665.

[13] Su G, Wei D, Varshney KR, Malioutov DM. Learning sparse two-level boolean rules. In: 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE; 2016. p. 1–6.

[14] Lakkaraju H, Bach SH, Leskovec J. Interpretable decision sets: A joint framework for description and prediction. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. p. 1675–1684.

[15] Wang T, Rudin C, Doshi-Velez F, Liu Y, Klampfl E, MacNeille P. A bayesian framework for learning rule sets for interpretable classification. The Journal of Machine Learning Research. 2017;18(1):2357–2393.

[16] Joly A, Schnitzler F, Geurts P, Wehenkel L. L1-based compression of random forest models. In: 20th European symposium on artificial neural networks; 2012. .

[17] Painsky A, Rosset S. Lossless compression of random forests. Journal of Computer Science and Technology. 2019;34(2):494–506.

[18] Buciluǎ C, Caruana R, Niculescu-Mizil A. Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining; 2006. p. 535–541.

[19] Bohanec M, Bratko I. Trading accuracy for simplicity in decision trees. Machine Learning. 1994;15(3):223–250.

[20] Qiao L, Wang W, Lin B. Learning Accurate and Interpretable Decision Rule Sets from Neural Networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35; 2021. p. 4303–4311.

[21] Freitas AA. Comprehensible classification models: a position paper. ACM SIGKDD explorations newsletter. 2014;15(1):1–10.

[22] Quinlan JR. Simplifying decision trees. International journal of man-machine studies. 1987;27(3):221–234.

[23] Bénard C, Biau G, Veiga S, Scornet E. Interpretable random forests via rule extraction. In: International Conference on Artificial Intelligence and Statistics. PMLR; 2021. p. 937–945.

[24] Izza Y, Ignatiev A, Marques-Silva J. On explaining decision trees. arXiv preprint arXiv:201011034. 2020;.

[25] Hulot A, Chiquet J, Jaffrezic F, Rigaill G. Fast tree aggregation for consensus hierarchical clustering: application to multi-omics

data analysis. In: Statistical Methods for Post-Genomic Data (SMPGD); 2019. .

[26] Kavšek B, Lavrač N, Ferligoj A. Consensus decision trees: Using consensus hierarchical clustering for data relabelling and reduction. In: European Conference on Machine Learning. Springer; 2001. p. 251–262.

[27] Strecht P, Mendes-Moreira J, Soares C. Inmplode: A framework to interpret multiple related rule-based models. Expert Systems. 2021;p. e12702.

[28] Andrzejak A, Langner F, Zabala S. Interpretable models from distributed data via merging of decision trees. In: 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). IEEE; 2013. p. 1–9.

[29] Andrienko N, Andrienko G, Fuchs G, Slingsby A, Turkay C, Wrobel S. Visual Analytics for Data Scientists. Springer; 2020.

[30] Sacha D, Kraus M, Keim DA, Chen M. VIS4ML: An Ontology for Visual Analytics Assisted Machine Learning. IEEE Transactions on Visualization and Computer Graphics. 2019 Jan;25(1):385–395. https://doi.org/10.1109/TVCG.2018.2864838.

[31] Andrienko N, Lammarsch T, Andrienko G, Fuchs G, Keim D, Miksch S, et al. Viewing Visual Analytics as Model Building. Computer Graphics Forum. 2018;37(6):275–299. https://doi.org/10.1111/cgf.13324.

[32] Andrienko N, Andrienko G, Miksch S, Schumann H, Wrobel S. A theoretical model for pattern discovery in visual analytics. Visual Informatics. 2021;5(1):23 – 42. https://doi.org/10.1016/j.visinf.2020.12.002.

[33] Spinner T, Schlegel U, Schäfer H, El-Assady M. explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. IEEE Transactions on Visualization and Computer Graphics. 2020;26(1):1064–1074. https://doi.org/10.1109/TVCG.2019.2934629.

[34] Ming Y, Qu H, Bertini E. RuleMatrix: Visualizing and Understanding Classifiers with Rules. IEEE Transactions on Visualization and Computer Graphics. 2019 Jan;25(1):342–352. https://doi.org/10.1109/TVCG.2018.2864812.

[35] Yuan J, Nov O, Bertini E. Visualizing Rule Sets: Exploration and Validation of a Design Space. arXiv preprint arXiv:210301022. 2021;.

[36] Rote G. Computing the minimum Hausdorff distance between two point sets on a line under translation. Information Processing Letters. 1991;38(3):123–127. https://doi.org/https://doi.org/10.1016/0020-0190(91)90233-8.

[37] Jaccard P. THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. New Phytologist. 1912;11(2):37–50. https://doi.org/https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

[38] Quinlan JR. Generating Production Rules from Decision Trees. In: Proceedings of the 10th International Joint Conference on Artificial Intelligence - Volume 1. IJCAI'87. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1987. p. 304–307.

[39] van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research. 2008 11;9:2579–2605.

[40] Ankerst M, Breunig MM, Kriegel HP, Sander J. OPTICS: Ordering Points to Identify the Clustering Structure. SIGMOD Rec. 1999 Jun;28(2):49–60. https://doi.org/10.1145/304181.304187.

[41] Dua D, Graff C.: UCI Machine Learning Repository. Available from: http://archive.ics.uci.edu/ml.

[42] Ayres-de Campos D, Bernardes J, Garrido A, Marques-de Sa J, Pereira-Leite L. SisPorto 2.0: a program for automated analysis of cardiotocograms. Journal of Maternal-Fetal Medicine. 2000;9(5):311–318.

[43] Sutton C, Ghiringhelli LM, Yamamoto T, Lysogorskiy Y, Blumenthal L, Hammerschmidt T, et al. Crowd-sourcing materials-science challenges with the NOMAD 2018 Kaggle competition. npj Computational Materials. 2019;5(1):1–11.

[44] Bartók AP, Payne MC, Kondor R, Csányi G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. Physical review letters. 2010;104(13):136403.

[45] Huo H, Rupp M. Unified representation of molecules and crystals for machine learning. arXiv preprint arXiv:170406439. 2017;.

[46] Sutton C, Boley M, Ghiringhelli LM, Rupp M, Vreeken J, Scheffler M. Identifying domains of applicability of machine learning models for materials science. Nature communications. 2020;11(1):1–9.

[47] Chen C, Lin K, Rudin C, Shaposhnik Y, Wang S, Wang T. A holistic approach to interpretability in financial lending: Models, visualizations, and summary-explanations. Decision Support Systems. 2022;152:113647. https://doi.org/10.1016/j.dss.2021.113647.