# Scribe Notes on Comprehensible Classification models: a position paper

m6sharma@ucsd.edu, musharma@ucsd.edu

10 October 2023

## 1 INTRODUCTION

- The majority of classification model evaluations prioritize predictive accuracy.
- In real-world applications, model comprehensibility (interpretability) is crucial.
- Early machine learning research emphasized model interpretability.
- SVMs and ensemble models, designed mainly for accuracy, are often considered black box models, hard to interpret.
- There's been significant progress in enhancing classification model comprehensibility.
- Comprehensible classification models are emphasized in domains like medicine, credit scoring, churn prediction, and bioinformatics.
- Many studies focus on improving the comprehensibility of specific classification model types, like decision trees or classification rules.

### 1.1 Goals and Scope of the Paper

- Discuss pros and cons of interpretability for five classification knowledge representations: decision trees, classification rules, decision tables, nearest neighbors, and Bayesian network classifiers.
- Address how to generally improve classification model comprehensibility.
- Focus is on the broad sense of comprehensibility, not just model size.
- Extraction of comprehensible models from ”black box” models (e.g., SVMs, ANNs) is out of scope.
- This paper emphasizes classification models over classification algorithms.

# 2 THE CASE FOR COMPREHENSIBLE CLASSIFICATION MODELS

- The need for comprehensible classification models is rooted in multiple factors.
- Users often need to understand a model before trusting its predictions and associated recommendations.
- Trust in computational predictions is especially vital in medical fields where lives are involved.
- Model comprehension is crucial for user acceptance in sectors like finance and customer churn prediction.
- In bioinformatics, understanding model predictions boosts trust, influencing the commitment to conduct extensive and costly experiments.
- The true value of bioinformatics model predictions is gauged by the cumulative success of inspired experiments.
- The demand for comprehensible models heightens when a system yields unexpected results, necessitating robust explanations for user acceptance.
- Real-world examples, like the incident at the Three-Mile Island nuclear power plant, showcase the importance of trust in automated system recommendations.
- In certain domains, users must comprehend system recommendations to legally justify their decisions to others.

# 3 INTERPRETING DIFFERENT TYPES OF CLASSIFICATION MODELS

## 3.1 Interpreting Decision Trees

- Decision trees are inherently comprehensible due to their graphical structure.
- They focus on the most relevant attributes.
- Hierarchical structure indicates relative importance of attributes.
- Depth in the tree provides insight into attribute relevance.
- However, comparing relevance purely based on depth has limitations.

## 3.2 Interpreting Classification Rules

- Rules follow the IF (conditions) THEN (class) format.
- Decision trees can be converted into IF-THEN rules.
- Rules are textual, whereas trees are graphical.
- Trees offer hierarchical information; rules do not.
- Ordering rule conditions by relevance can be challenging.
- Converting a decision tree into a set of rules might enhance comprehensibility.
- Rule induction algorithms and tree induction algorithms differ in attribute selection.
- Analyzing exceptions to rules can provide deeper insights.

## 3.3 Interpreting Decision Tables

- Decision tables use tabular knowledge representations.
- Focus is on single-hit tables for enhanced interpretability.
- Inducing decision tables can be done in various ways, including from decision trees or ANNs.
- Larger rule sets might necessitate table contraction methods.
- Decision tables can have many cells, even when using fewer attributes.

## 3.4 Interpreting Nearest Neighbors

- Nearest neighbor algorithms provide instance-specific explanations.
- Challenges include variability of explanations and high-dimensional data.
- Solutions include using prototypes or computing attribute weights.

## 3.5 Interpreting Bayesian Network Classifiers

- Interpreting Naïve Bayes involves analyzing probabilities for all attributes.
- Naïve Bayes assumes attribute independence given the class.
- Advanced Bayesian Network Classifiers, like General Bayesian Networks, consider attribute correlations.
- Size and structure of the network play roles in comprehensibility.
- Directed edges in BNCs can be confusing to users.
- Dependency networks are a proposed alternative to BNCs.

### 3.6 User-based Experiments Comparing the Comprehensibility of Different Types of Classification Model

- Few experiments have directly compared different classification model types.
- One experiment found most users preferred decision tables.
- Another found decision trees more understandable than rule lists.
- A different experiment favored decision trees over decision tables.
- User preferences vary based on their backgrounds.
- However, many users prefer decision trees, rules, or tables over other complex models.

## 4 INTERPRETING CLASSIFICATION MODELS: GENERIC ISSUES

### 4.1 The Drawbacks of Model Size as the Single Measure of Comprehensibility

- Many papers evaluate model comprehensibility based solely on its size, such as the number of nodes in a decision tree or the number of edges in a Bayesian network.
- A model's size only captures its syntactical aspect, ignoring semantics.
- The actual contents of the model, like attributes or conditions, play a significant role in its comprehensibility.
- Larger models can sometimes be more comprehensible because they use more relevant attributes.
- Users may reject overly simplistic models, even if they are technically accurate.
- The definition of ”too large” varies among users.
- Instead of setting a strict maximum model size, it's better to approach the accuracy-comprehensibility trade-off through methods like multi-objective optimization.

### 4.2 Incorporating Semantic Monotonicity Constraints in Classification Models

- Models that respect monotonicity constraints provided by experts are more likely to be trusted by users.
- Monotonic relationships between attributes and class predictions can be either for numerical or nominal attributes.
- Monotonicity constraints can be hard (strict) or soft (can be violated for increased accuracy).
- Monotonicity can be enforced in various ways, depending on the type of classification model.
- While experts can provide monotonicity constraints, they can also be derived from high-quality data.

## 5 SUMMARY AND DISCUSSION

- Knowledge representations have their pros and cons. No single representation is universally the "most comprehensible".
- The choice of representation should consider the user's background, preferences, and the dataset's characteristics.
- While decision trees might be better for datasets with many relevant attribute values, rule-based models can be more suitable for datasets with single relevant values.
- Analyzing decision trees and rules can benefit from considering exceptions or counter-examples.
- Bayesian network classifiers may be harder for users to interpret than decision trees.
- Nearest neighbor classifiers can offer explanations based on the attribute values of the nearest neighbors.
- More experiments involving user evaluations of different model representations are needed.
- The assumption that smaller models are more comprehensible is problematic and doesn't hold universally.
- Monotonicity constraints add a semantic layer to model interpretation and can significantly influence user trust.