Information-theoretic partition-based models for interpretable machine learning

Yang, L.

Chapter 7

Conclusions

The importance of interpretability is widely accepted in machine learning tasks in which humans are responsible for making a decision for taking action. In such scenarios, it is crucial for domain experts to trust the machine learning model. As a result, research on interpretable machine learning has received a lot of attention in recent years. This dissertation contributes to this area by proposing intrinsically interpretable and transparent methods for supervised and unsupervised tasks, both for predictive modeling and for knowledge discovery.

In this Chapter we provide our conclusions.

7.1 Summary

Truly Unordered Rule Sets. In the field of rule set models, we considered the problem of increasing the interpretability of rule set models by removing the ad-hoc schemes for handling conflicts caused by overlaps of rules, in which an overlap refers to a subset of instances covered by multiple rules simultaneously.

In order to achieve this goal, we first considered allowing overlaps for expressing uncertainty and exception, which eliminated the need for imposing implicit orders among rules. Building upon it, we next formally defined truly unordered rule set (TURS) models, which informally only "allow" rules with similar outputs to overlap. Lastly, we showcased through a case study with patient data collected at Leiden University Medical Center that our TURS model paves the way to interactive rule learning. That is, the rule set model can be automatically updated with feedback from domain experts.

Multi-dimensional MDL-based Histograms. We studied multi-dimensional MDL-based histograms, which can be used as a transparent tool for various tasks in machine learning and data mining, including density estimation, explanatory data analysis, discretization, entropy estimation, and conditional mutual information estimation. With a series of papers, we first extended the one-dimensional MDL-based histogram to the two-dimensional case and showcased its use for analyzing spatial datasets. Secondly, we extended MDL-based histograms for analyzing multi-dimensional and mixed-type datasets (with discrete-continuous mixture variables), specifically for analyzing its dependency structures via conditional mutual information estimation.

7.2 Answers to Research Questions

In the following, we revisit the proposed research questions and provide our answers and a discussion for each of them.

Research Question 1: How can we formalize rule sets as probabilistic models such that rules are independent of each other? Further, how to learn such a model from data?

A set of rules, when put together, can form a global model for the whole dataset. While defining a single rule as a local probabilistic model is straightforward (given that the target variable is discrete for classification tasks), defining a global model for a rule set is far more complicated, mostly due to the nuisance caused by overlaps, i.e., one instance covered by multiple rules at the same time.

To remove implicit and explicit orders among rules, we treated rules equally, i.e., one rule does not have a higher "quality" than the other. We started by considering the informal implication of an overlap of two rules; i.e., what is the *implication* of the overlap in the sense that why the instances covered by the intersection of these two rules cannot form a rule on itself (by concatenating the conditions of the two rules)? This leads to our justification of the overlap: if the class probabilities of the instances covered by the overlap are not very different from those of the instances covered by each single rule respectively, it is not desirable to separate the instances in the overlap to nether of the two rules. Informally, this can be caused by close class probability estimates and/or by a small sample size of the overlap (which leads to a large variance). In this case, we interpret overlaps as "uncertainty", in the sense that we do not have enough data to decide that the instances covered by the overlap "belong" to a single rule.

Thus, our first answer to Research Question 1 is that we treat overlap as uncertainty when formalizing rule sets as probabilistic models. This approach is very different from previous methods, which either minimize the size of overlaps or takes post-hoc conflict resolving schemes.

Further, when regarding an overlap as uncertainty, an overlap of two rules, e.g., rule S_i and rule S_j, can be interpreted as "instances that are covered by the overlap "belong" to rule S_i **or** rule S_j", in which the "or" represents uncertainty. This intuition motivated us to consider taking the union of rules for modeling instances covered by the overlap, which leads to our second answer to the proposed

research question.

Our second answer to Research Question 1 is that we formally define a probabilistic model for any given rule set that may have overlaps, i.e., the truly unordered rule set (TURS) model. The key innovation is to define

$$P(Y = y|X = x) := P(Y|\{X \in \cup S_i\}), \forall x \in \cap S_i.$$

In this way, the likelihood of a TURS model incorporates how different the class probability estimates of rules that form an overlap are. Thus, we now only "allow" overlaps that have similar class probability estimates by penalizing the situation when two rules with very different class probability estimates overlap.

Lastly, as learning a TURS model from data requires taking into consideration modeling overlaps, existing formalizations of the problem of learning rules from data cannot be applied to learning a TURS model. Also, existing rule learning algorithms cannot be used directly or with modification easily.

Therefore, our third answer to Research Question 1 is that we formally defined the problem of learning a TURS model as an MDL-based model selection problem, and we developed a novel heuristic algorithm for finding good models.

Introducing the MDL principle removes the regularization parameter for controlling the model complexity. Setting such regularization parameters in an ad-hoc way reduces the algorithm transparency, while tuning it with cross-validation can be time-consuming. Moreover, our algorithm is equipped with several algorithmic innovations, including 1) taking "patience" into account, 2) using a dual-beam, and 3) using a look-ahead strategy based on a MDL-based local test. Our algorithm is shown to have competitive predictive performance and simple model complexity; further, more importantly, the TURS models learned by our algorithm are shown to be empirically "truly unordered", in the sense that the predictive performance is hardly affected by randomly chosen rules for making predictions for instances covered by overlaps.

Research Question 2: How can we construct parameter-free two-dimensional histograms with transparent and informative patterns (bins)?

Eliminating user-defined parameters for controlling the bin sizes of histograms

increases the transparency of how a histogram model is obtained from data. Hence, the ambiguity to data analysts caused by different histograms showing different data summarization is removed.

In order to remove parameters that control the bin sizes of histograms, we formalize the problem of learning such histogram models as an MDL-based model selection problem. That is, we adopted the MDL principle to define the "optimal" model for such an unsupervised task and formalized the best model as the one that leads to the best compression of data and model together.

In addition, to obtain more interpretable bins (patterns) in the sense that 1) instances within each bin can be considered to have homogeneous density, and 2) neighboring bins have different densities, we proposed a greatly flexible model class that includes any data partition formed by unions of disjoint rectangles. Lastly, we developed an efficient algorithm that combines top-down and bottom-up search, and showcased that the learned two-dimensional histograms carry meaningful patterns that generalize well to unseen data, both on simulated datasets with known ground truth and real-world case study datasets.

Thus, our answer to Research Question 2 is to formalize the problem of learning a two-dimensional histogram based on the MDL principle, and to obtain more informative patterns (bins) by 1) considering dependencies between dimensions and 2) using more flexible data partitions.

Research Question 3: How can we construct a multi-dimensional adaptive histogram-based model for interpretable CMI estimation?

Learning dependency structure via estimating the conditional mutual information (CMI) is a challenging task, especially when the data contains mixed types (discrete, continuous, and possibly also discrete-continuous mixtures).

To construct histograms for mixed type data, we first formalized the problem of estimating CMI for mixed type data. Specifically, we adopted measure-theoretic tools to prove that the CMI for mixed-type data can be written as the sum of four entropy terms, just like the CMI for purely continuous and discrete data.

Further, we proposed an entropy estimator based on multi-dimensional histogram models, and consequently a plug-in estimator for CMI. Next, we formalized the problem of learning a multi-dimensional adaptive histogram as an MDL-based model selection task. Leveraging the MDL principle reduced the

hyper-parameters to be set and hence increased the transparency of how a model is obtained. Lastly, we proposed an alternating algorithm for learning such a multi-dimensional histogram from data and showcased the effectiveness of such an approach by benchmarking against competitor methods in various tasks that involve CMI estimation.

In conclusion, our answer to Research Question 3 is 1) to adopt the MDL principle to formalize the learning problem, 2) to leverage the measure-theoretic definition of entropy for mixed-type of data, and 3) to design an alternating algorithm for learning such a histogram form data.

7.3 Future Work

We conclude this chapter by discussing a few possible future work directions following this dissertation.

First, we consider a crucial problem to formally define *human comprehensibility* as a measure in interpretable machine learning, which characterizes how easy a machine learning model can be comprehended by a human. Notably, the concept of human comprehensibility may be defined both for intrinsically interpretable models and explainable artificial intelligence (XAI) methods that provide post-hoc explanations for black-box models. One key challenge is to properly define the "required level" of human comprehension, which can be different for various machine learning tasks.

Second, it is a fundamental research question to formalize as an optimization problem the task of automatic model updating given human feedback, which is the cornerstone of any interactive machine learning system. One potential approach is to borrow the idea from the *subjective interestingness* in information-theoretic data mining (De Bie 2011a,b). However, subjective interestingness in data mining is, informally, about maximizing the "surprisingness" to the data miner based on their prior "beliefs" about the dataset. Thus, the goal is to search the pattern with the maximum amount of information in the data that the user does not know. In contrast, to formalize automatic human-guided model updating, the goal could be set as searching for a model that maximizes the "trust" from human users. As an example, such a model could be a probabilistic rule set that contains rules that

the user knows or considers trustworthy.

Third, further research about how to develop a flexible interactive data exploration system may be another crucial component for human-in-the-loop machine learning systems. It may be useful for building trust between humans and models if we allow human users to explore subsets of datasets with the help of specific types of machine learning models, including examining the statistical characteristics of (subsets of) datasets.