Information-theoretic partition-based models for interpretable machine learning

Yang, L.

Chapter 4

Case Study: Towards Interactive Rule Learning for ICU Readmission Analysis

Chapter Abstract

Interactive machine learning systems that can incorporate human feedback for automatic model updating have great potential use in critical areas such as health care, as such systems can combine the strength of data-driven modeling and the prior knowledge from domain experts. Designing such a system is a challenging task as it must enable mutual understanding between humans and computers, which hence relies on interpretable and specifically easily comprehensible models. Specifically, we consider the problem of incorporating human feedback for model updating in rule set learning for the task of predicting readmission risks for ICU patients. Building upon the TURS model described in the previous chapters, we further propose a certain format for feedback for rules, together with an automatic model updating scheme. We conduct a pilot study and demonstrate that the rules obtained by updating the TURS model learned from the ICU patients' data can empirically incorporate human feedback without sacrificing predictive performance. Notably, the updated model can exclude conditions of rules that ICU physicians consider clinically irrelevant, and thus enhance the trust of physicians.

4.1 Introduction

In critical areas such as health care, developing machine learning models that domain experts can comprehend and trust potentially has great societal impact. Specifically, in intensive care units (ICU) where patients are monitored intensively, conditions of patients are to a large extent recorded digitally, which provides the foundations for building decision support systems with data-driven models (Hond et al. 2023).

We consider the problem of predicting the probability of readmission to the ICU within a short period (7 days) after a patient is discharged from the ICU and moved to a normal ward. Such readmission risk for patients is clinically relevant, as it is observed that patients who are readmitted often become much worse in comparison to their condition when they were in the ICU previously (Kramer et al. 2013; Woldhek et al. 2017). Thus, the readmission itself is a key factor that is highly correlated with the patient's condition; as a result, predicting the readmission risk can both facilitate efficient ICU resource management and prevent discharging patients improperly. In practice, beds in the ICU are a very scarce and costly resource; thus, discharging patients from the ICU smartly can help distribute the resource to patients who need it more.

As physicians are responsible for estimating the risk of discharging a patient from the ICU, data-driven models only brings benefits if physicians trust the model and are willing to use it in practice. To build trust, the data-driven model needs to have interpretability for domain experts to comprehend what is going on (Li et al. 2023). Further, beyond interpretability, the situation when physicians and machine learning models disagree must be properly handled (Holzinger 2016; Mosqueira-Rey et al. 2023; Teso and Kersting 2019). That is, when the model gives a probabilistic prediction together with explanations, what if the physician disagrees with the prediction and/or the explanation? For instance, the model could identify a factor that is known to be irrelevant clinically as important for predicting readmission risk for a single patient. In this situation, it would be ideal if the physician would give this feedback to the machine learning model; further, if the model can be automatically updated when receiving the feedback from human, the physician could trust the model next time when the model gives the same explanation and prediction for a similar patient in the future.

Thus, interaction between humans (i.e., physicians in the ICU in this case) and the machine learning model is crucial in such a scenario, which requires the human to understand the machine, and at the same time, the machine to understand the human. While rule-based models, and especially truly unordered rule set (TURS) models, are in principle comprehensible to domain experts, the challenges remain unresolved that *1) how and in what formats feedback from domain experts can be incorporated, and 2) how rule-based models can be updated according to human feedback*.

To tackle these challenges, we introduce a human-guided rule updating scheme based on the TURS model. The TURS model paves the way towards an interactive rule learning process with the following two advantages over existing methods for learning rule lists (in which rules are explicit ordered) and rule sets (in which external and ad-hoc methods are mostly used to handle the conflicts caused by overlaps).

The first advantage is that rules in the TURS model can be empirically regarded as truly unordered and hence independent from each other. Thus, deleting and/or editing one rule (that a domain expert dislikes) has little influence on other, potentially overlapping rules. In contrast, for rules with (implicit) orders obtained by other existing methods, editing or deleting one rule may cause "a chain of effects" on how instances covered by other rules are modeled. Secondly, the TURS model reduces the workload for domain experts to find out which rules need to be edited. Specifically, when comprehending a single rule, there is no need to go over all other rules that are ranked (explicitly or implicitly) higher, as unlike other existing methods, our TURS model does not impose any order among rules.

In the following, we conduct an empirical pilot study by applying the TURS model to a dataset collected at the ICU of Leiden University Medical Center (LUMC) in the year 2020. To this end, we ask a domain expert from LUMC to identify rules with clinically irrelevant variables, and we also propose an updating scheme for the TURS model.

4.2 Updating Rule Sets with Human Feedback

We now describe in what format we allow ICU physicians to give feedback, and how the TURS model can be updated automatically with the feedback.

4.2.1 Human feedback format

Although it seems tempting to allow feedback in flexible formats (and the most flexible format would be in natural language), we argue that it is desirable to constrain human feedback to have certain formats, in order to transform the feedback into *transparent human guidance* to the algorithm for updating the model. In other words, we aim to propose certain human feedback formats so that the consequence of such human feedback can be easily explained to domain experts.

However, such feedback format should also allow domain experts to express clearly and sufficiently why they dislike the current model. This requires a deep understanding about what might cause dissatisfaction from domain experts. Hence, how to design such feedback formats may depend on the application task at hand, and may require collaboration between computer scientists and domain experts.

Focusing on the task of ICU readmission risk analysis, we constrain ourselves to a simple yet fundamental feedback format and leave as future work incorporating other feedback formats. Formally, given a truly unordered rule set model with K rules denoted as $M = \{S_1, ..., S_K\}$, we consider feedback from domain experts in the following form: *remove rule* S_j *due to irrelevant variables* $\{X_i\}_{i \in I}$, in which S_j denotes a single rule and I denotes an index set. Notably, feedback in this format contains not only information regarding whether a rule is disliked, but also the reason why a rule is disliked.

4.2.2 Updating a rule set

We now present how we can equip the TURS model with an "self-updating" scheme after receiving feedback from a domain expert.

Removing a rule. Given the rule set $M = \{S_1, ..., S_K\}$, assume that a domain expert gives the feedback that rule S_i does not make sense as it contains irrelevant variable X_j. Then, removing S_i from M is straightforward as there exist no implicit or explicit orders among rules. That is, following the procedure of formalizing a rule set as a TURS model, we simply have a new rule set $M' = M \setminus \{S_i\}$, for which the likelihood can be calculated according to how the TURS model is defined.

We next analyze for which instances the empirical class probabilities are affected. First of all, when S_i is eliminated from model M, it has an effect on the estimated probabilities of both 1) instances covered by S_i, and 2) instances not covered by any rule (i.e., covered by the else rule). Specifically, instances previously covered by S_i only (i.e., before removing S_i) are now combined with instances originally covered by the else-rule, which are now used for obtaining new class probability estimates for the new else rule after eliminating S_i from M. Meanwhile, for instances covered by the overlap of S_i and some other rule(s), the class probability estimates will be updated accordingly.

Learn a new rule with constraint. Building upon the new TURS model M', we next consider learning a new rule that can be added to M' as the replacement for the removed rule, for which we leverage the dual-beam diverse-patience algorithm for learning the next "best" rule given the current status of a rule set, as proposed in Chapter 3.

As the conditional likelihood of class labels can be calculated given the dataset and the rule set M' given the definition of the TURS model, the MDL-based model selection score for the rule set M' can be calculated accordingly. Further, when adding a rule S' to M', the model selection score can be calculated for $M' \cup \{S'\}$ as well.

Thus, the algorithm can search for the next best rule S' such that when adding S' to M' the *learning speed score* $r(S')$ is optimized (as defined in Chapter 3), in which $r(S')$ measures how much the MDL-based model selection score decreases per extra covered instance when adding S' to M'.

4.3 An Empirical Pilot Study

We conduct a pilot study in collaboration with Leiden University Medical Center (LUMC) using the real-world patient dataset to showcase how the TURS model together with the model updating scheme can be used for interactive rule learning with humans in the loop. We next describe the experiment setup and present our results.

4.3.1 Experiment setup

Dataset description. We specifically considered the dataset collected at the ICU of LUMC in the year 2020, in which the patients who are readmitted within 7 days are labelled as "positive".

The original dataset is multi-modal and contains information in different forms, including time series measurements (e.g., cardiology monitor records), lab results over time (e.g., blood tests), medication use records, as well as static information for each patient (e.g., age, gender, etc). This dataset was described and pre-processed into a tabular dataset by an external expert in previous research (Van der Meijden 2021). The resulting processed dataset was further split randomly for training and test, which contains 9737 and 2435 patients respectively (approximately 80%/20% splitting), with 550 feature variables. The dataset is very imbalanced, as the overall probability of readmission is roughly 0.07.

Human feedback collection. We ask one domain expert from LUMC to give feedback to the rules, with the procedure as follows. First, a TURS model is learned on the training set, with beam width set as 5 and the number of candidate cut points (for continuous-valued features) set as 20, which is the "default" setting that we also used in Chapter 3.

Second, the rule set is shown to the domain expert; specifically, the condition of each rule together with the class probability estimates (obtained using the training set) are shown to the domain expert. Moreover, the algorithm configuration (e.g., the beam width) is revealed to the domain expert as well.

Next, we ask the domain expert to go through each of all rules, and to give feedback to the ruleset in the format as we described in Section 4.2. Subsequently, the feedback is used to update the TURS model, and we use the test set of the ICU dataset for assessing the predictive performance of the TURS model before and after the human feedback. We refer to the latter as the human-guided model. Lastly, note that the test set of the whole dataset is only used for this final assessment step, and therefore the domain expert has no access to it during the procedure of giving feedback to rules.

4.3.2 Rule set for the ICU dataset

Learning a TURS model using our proposed method in Chapter 3, we obtain a surprisingly simple rule set with 5 rules only, which has average rule length of 2. The obtained rule set is shown in Table 4.1:

Rule Conditions	Prob. of Readmission	# Patients
Ureum-max-all $\geq$ 12.1 Ademfrequentie-median-value-last24h $\geq$ 23.5	0.223	494
APTT-max-all $\geq$ 43.1 Ureum-mean-all $\geq$ 16.338	0.199	548
Leukocyten-mean-last $\geq$ 20.81	0.162	464
Kalium-count-first $\geq$ 6.0 specialty-Organization-value-sub-ICCTC = FALSE	0.131	1979
Trombocyten-count-first $\geq$ 2.0 Ureum-last-last $<$ 9.2 specialty-Organization-value-sub-ICCTC = TRUE	0.019	3922
None of the above	0.059	3220

Table 4.1: Rule sets describing the probability of readmission for LUMC ICU patients.

The literals contain feature names that are mostly consisting of three parts, with the first part indicating the basic meaning of this feature variable (in Dutch). For instance, "Ureum" indicates the "Urea" in blood. The second part of feature names indicates how the results are aggregated, among which "count", "mean", "median", and "max" are commonly used. Last, the third part of feature names indicates the time window for which the aggregated values are obtained, in which "first" represents the first 24 hours, "last" represents the last 24 hours, and "all" represents the whole period in ICU. A detailed explanation of the feature names can be found in previous work (Van der Meijden 2021).

4.3.3 Rule-based competitor methods

To benchmark the performance of the TURS model induced from the training dataset, we apply several commonly used probabilistic rule-based models to the ICU dataset. The motivation for such benchmark is to show that the TURS model has competitive predictive performance and thus implicitly describes the data relatively well, which is the foundation for involving humans in the loop.

The comparative predictive performance is summarized in Table 4.2. Notably, the TURS model shows advantages over competitor methods in several aspects. First, the results with respect to ROC-AUC and PR-AUC show that the ICU

dataset is difficult to model using widely used rule-based models (as listed in the table), since the ROC-AUC of C4.5 and RIPPER are roughly equal to 0.5. Further, the TURS model shows its robustness in achieving the best ROC-AUC and PR-AUC, and notably with significantly simpler rules (except when compared to RIPPER, which seriously "underfits" the data).

Moreover, rules in the TURS model generalize best to the unseen instances in the test set (excluding RIPPER for its low ROC-AUC scores). Specifically, we calculate the difference between the class probability estimates obtained using the training and test dataset, as also reported in the table. We hence conclude that the probability estimate for each single rule of the TURS model shown to physicians are most reliable and trustworthy.

Algorithm	CN2	CART	RIPPER	C4.5	TURS
ROC-AUC	0.641	0.690	0.514	0.539	0.705
PR-AUC	0.114	0.137	0.084	0.076	0.164
Train/test prob. diff.	0.041	0.031	0.001	0.054	0.006
# rules	851	25	1	249	5
Avg. rule length	2.5	4.2	5.0	16.8	2.0

Table 4.2: Rule-based model results on ICU dataset.

4.3.4 Human-AI collaboration

We now showcase that our TURS model can be equipped with the model updating scheme to generate human-guided rule sets. Notably, our approach is very different than existing model editing approaches (Wang et al. 2022), as the end user is not allowed to directly edit the model in our model updating scheme; instead, we only allow user to provide feedback, and the updated model is still learned in a data-driven manner. That is, we let the data always take the leading role, in order to avoid arbitrary (or adversarial) model editing.

Specifically, we consider the rule set obtained in Section 4.3.2, and we collected two pieces of feedback from the domain expert: 1) the domain expert dislikes the 5th rule due to the first variable, and 2) the domain expert dislikes the 3rd rule which contains only one literal.

We thus discard the 5th rule from the rule set, and we next search for a new rule to be added to the rule set, with the constraint that the first variable in the 5th

rule must not be included. We present the new human-guided rule together with the original rule in Table 4.3. We show that our TURS model indeed makes such an interactive process possible, and specifically that it can handle feedback that can be transformed into constraints with respect to excluding certain variables. Further, we demonstrate that for the rule set induced from the ICU patients' dataset, editing a rule based on the human feedback (without the necessity to modify other 'overlapping' rules), can indeed discard certain variables but at the same time keep the predictive performance at the same level.

Note that the updated rule and the original rule are coincidentally very similar; that is, the feedback to the TURS model is only about discarding the first literal of the 5th rule, without asking it to keep the other literals and/or variables in the original rule.

Human-guided	No	Yes
Rule	If Trombocyten-count-first $\geq$ 2.0; Ureum-last-last $<$ 9.2; specialty-Organization-value-sub-ICCTC = TRUE $\rightarrow$ Probability of Readmission: 0.019, Number of patients 3922	If Leukocyten-count-first $\geq$ 2.0; Ureum-last-last $<$ 9.2; specialty-Organization-value-sub-ICCTC = TRUE $\rightarrow$ Probability of Readmission: 0.019, Number of patients 3958
ROC-AUC (rule set)	0.705	0.706
PR-AUC (rule set)	0.164	0.164

Table 4.3: Comparison between the rule before and after a domain expert feedback, together with the ROC-AUC and PR-AUC of the resulting new rule set. Changes in rules conditions before and after human feedback are shown in red and blue respectively.

Next, for examining the effect of the second feedback, we remove the 3rd rule from the original purely data-driven rule set, and search for another rule by excluding the variable "Leukocyten-mean-last" from the search space. We present the results in Table 4.4, which shows that the new rule covers 375 more patients than the original rule. Again, without the need for further modifying other rules, editing the 3rd rule in the original rule set with the updated rule keeps the ROC-AUC and PR-AUC at the same level.

Human-guided	No	Yes
Rule	Leukocyten-mean-last $\geq$ 20.8 $\rightarrow$ Probability of Readmission: 0.162, Number of patients 464	CRP-mean-last-missing $=$ 1 $\rightarrow$ Probability of Readmission: 0.030, Number of patients 839
ROC-AUC (rule set)	0.705	0.704
PR-AUC (rule set)	0.164	0.172

Table 4.4: Another comparison between the rule before and after a domain expert feedback.

4.4 Conclusion and Discussion

We studied the problem of estimating readmission risk for patients in ICU as an applied machine learning task. In order to resolve the difficult situation when domain experts (physicians) dislike certain rules, which can result in the lack of trust for such data-driven models, we aimed for developing a human-guided rule learning scheme based on our method for learning truly unordered rule set (TURS) models.

We presented a pilot empirical study using the patients data collected at Leiden University Medical Center (LUMC) in the year 2020. Specifically, we firstly presented the learned rule set from the ICU dataset, and compared the predictive performance against other widely used rule-based competitor models, which demonstrated the superiority of the TURS model in terms of both predictive performance and model complexity. This result set the foundation for using the TURS model as a basis for interactive rule learning.

Next, we asked a domain expert from LUMC to give feedback to the purely data-driven rules, and we proposed a simple model updating scheme to incorporate the feedback to induce human-guided rules. We showcased that such a process can lead to new rules as replacements for rules that the domain expert disliked, without sacrificing the predictive performance of the whole model. Notably, the properties of the TURS model enables straightforward, transparent, and efficient model editing, without the need for re-training other rules in the model. We next discuss potential future research directions.

4.4.1 Discussion for future work

We have shown that the truly unordered rule set (TURS) model is "ready" for interactive rule learning, i.e., in a straightforward way it can be equipped with a model updating scheme that incorporates human feedback in certain formats. Following this research line, it may be with great potential to explore the following research questions.

User feedback formats. One natural but crucial question is in what formats we allow domain experts to give feedback to the data-driven model, and further how to inspire and elicit feedback with tools that allow an end user to investigate the data and the rule-based models.

For instance, it may be beneficial to allow domain experts to "zoom in" for each single rule, and examine values of other features for each corresponding subset of patients. While all instances in each rule share the same class probability estimate, domain experts may find one single "typical" patient who should have a different probability estimate than the rest. This may induce feedback in the form of "modifying a given rule by excluding a certain instance from its cover".

Further, we could allow domain experts to name risky factors within each rule; i.e., to allow the domain experts to suggest informative feature to be included in a single rule. Thus, we may allow feedback in the form of "for all patients covered by this rule, those patients whose feature value for variable X_i is larger than a certain threshold may have a higher risk of readmission". Such feedback is useful for 1) obtaining single rules with variables that are congruent with the domain knowledge, and 2) more interestingly, understanding the limits of the data (since the "best" rule with the suggested variables may result in a "worse" score according to the model selection criterion).

Transparent model updating. Introducing the human in the loop extends the meaning of transparency of a machine learning method. Previously, transparency roughly referred to whether the process of obtaining a model based on a given dataset is comprehensible to humans; in contrast, we argue that transparency is also applicable to describing whether the process of model updating based on human feedback is comprehensible to humans. Thus, it is a natural question to ask whether the trust between domain experts and data-driven models depends not only on the interpretability and transparency of the model but also on that

of the model updating scheme.

Further, while it is very transparent to incorporate human feedback as constraints like those we proposed, other ways of processing human feedback are to be explored. For instance, except for considering human feedback in certain formats as constraints, we may also translate human feedback to "prior" preferences.

User study for trust. Trust between domain experts like ICU physicians and data-driven models is a fundamental requirement for deploying a decision-support system in critical areas like health care, because, for instance, if physicians do not trust the data-driven model, they tend to simply ignore the data-driven predictions.

While the goal of involving humans in the loop to obtain human-guided rules to increase the trust by obtaining rules that are (more) congruent with the domain knowledge, whether trust is indeed increased can only be evaluated empirically via user studies. Thus, an interesting research question is how to formally define trust in the task of predicting readmission risk, inevitably with subjectiveness. As a result, it remains a challenge to design questionnaires for evaluating the trust between domain experts and data-driven models.