Robust Rule Learning for Reliable and Interpretable Insight into Expertise Transfer Opportunities

Willa Potosnak

Department of Engineering, Duquesne University Rangos School of Health Sciences, Pittsburgh, PA
Auton Lab, Carnegie Mellon University School of Computer Science, Pittsburgh, PA
potosnakw@duq.edu, wpotosna@andrew.cmu.edu

Abstract

Intensive care in hospitals is distributed to different units that care for patient populations reflecting specific comorbidities, treatments, and outcomes. Unit expertise can be shared to potentially improve the quality of methods and outcomes for patients across units. We propose an algorithmic rule pruning approach for use in building short lists of human-interpretable rules that reliably identify patient beneficiaries of expertise transfers in the form of machine learning risk models. Our experimental results, obtained with two intensive care monitoring datasets, demonstrate the potential utility of the proposed method in practice.

Introduction

Intensive care in hospitals is distributed to different units, or sites, that care for patient populations reflecting specific comorbidities, treatments, and outcomes (Nguyen, Perrodeau, and et al. 2014). Site expertise can be shared to potentially improve the quality of methods and outcomes for patients across sites (Caldas et al. 2021). An explicit and translatable understanding of which patients would benefit from site expertise is important as externally derived knowledge may not be applicable to an entire site population. Machine learning (ML) can be used to identify subpopulation beneficiaries of site expertise, however, ML model reliability is essential. Interpretable ML models, such as decision lists, bear considerable relevance for this purpose as their decisions can be understood and verified by domain experts.

We aim to identify knowledge transfer opportunities among specialized medical and surgical intensive care units (ICU) that could benefit patients across units. We propose an algorithmic rule pruning approach for use in building short lists of human-interpretable rules that reliably identify patient beneficiaries of expertise transfers in the form of ML risk models. Rule pruning is performed using permutation testing with the novel contribution of using the k-nearest neighbors (KNN) ML algorithm as a non-parametric approach to probability estimation.

Related Work

Decision Lists

Decision lists are ML models that comprise an ordered list of rules. The appeal of decision lists is their human-interpretable component rules organized in simple list-like structures. Decision list construction is composed of two main tasks: rule generation and rule selection. Rule pruning with permutation testing can be applied prior to rule selection as a non-parametric approach to estimate the probability that rules are derived from a non-permutated (original) dataset and not influenced by sampling variance (Frank 2000). Permutation testing is commonly applied with non-parametric probability estimation using chi-square or Fisher's exact tests. However, permutation testing with chi-square or Fisher's exact tests typically require sample class predictions, which are not feasible without classifier score decision thresholds. Our non-parametric approach using the KNN algorithm does not require sample class predictions.

Federated Classifier Selection

The federated classifier selection (FRCLS) algorithm generates a decision list with rules that identify regions of the feature space for which a classification model developed using data from an external site provides more accurate outcome predictions than a classification model developed using data from the local site (Caldas et al. 2021). Rule selection to generate the list is performed using a heuristic that maximizes the lower confidence bound on a variable that estimates external site model competence, which is a parametric approach that approximates the model competence variable distribution as normal (Caldas et al. 2021). Rules that describe small numbers of samples violate normal distribution assumptions. Inappropriate selection of these rules can result in lengthy decisions lists that overfit to training sample data. Rule pruning has application to remove rules influenced by sampling variance for more reliable rule selection.

Methods

We demonstrate the utility of our algorithmic rule pruning approach using two datasets, CH and MIMIC-II, with 1,563 and 1,776 samples, respectively. Each dataset was partitioned into two separate sites based on patient ICU stay with either medical (MICU) or surgical (SICU) focus. An

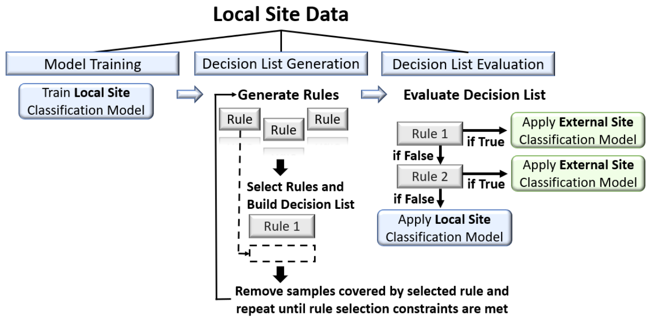

Figure 1: FRCLS pipeline is shown for a cross-validation fold with site models trained using random forest classifiers.

implementation of the RADSEARCH algorithm was used to generate rules (Moore and Schneider 2002). Two experiments were conducted for each dataset: **experiment 1** with SICU considered the local site and MICU considered the external site and **experiment 2** with the local and external site assignments switched. Two decision lists were generated with and without rule pruning prior to rule selection for each experiment using a 5-fold cross-validation scheme with the FRCLS implementation shown in Figure 1. Rule pruning was performed with the KNN algorithm trained on rule features that measure the number of described samples and external site model competence for these samples to estimate the probability that rules are derived from the non-permutated (original) dataset. Rules with probability estimates below 0.70 were pruned.

The decision lists generated with and without rule pruning prior to rule selection were assessed on whether the external site model applied to rule-identified local patients improves the area under the receiver operating characteristic curve (ROC-AUC) over that obtained with application of solely the local model to all patients. The number of decision list rules was also assessed as it influences the feasibility of the list's use in practice.

Results

Experiment:		ROC-AUC	
Dataset		**Local**	**Local and External**
1:	CH	0.875 (0.021)	**0.879 (0.028)**
	MIMIC-II	0.851 (0.034)	0.830 (0.051)
2:	CH	0.888 (0.012)	**0.893 (0.002)**
	MIMIC-II	0.708 (0.022)	**0.761 (0.050)**

Table 1: ROC-AUC average (standard deviation) for the local site model and for the local site model with the external site model applied to patients identified by rules selected after pruning. Decision lists generated with rule pruning for 3 of 4 experiments successfully identify subpopulations of local ICU patients for whom application of the external site model improves ROC-AUC (in bold), on average, over that obtained with application of solely the local site model.

Experiment:		p-value	
Dataset		Without Pruning	With Pruning
1:	CH	0.531 (0.311)	**0.203 (0.141)**
	MIMIC-II	0.891 (—)	**0.812 (—)**
2:	CH	0.811 (0.324)	**0.555 (0.445)**
	MIMIC-II	0.703 (0.120)	**0.602 (0.107)**

Table 2: Average (standard deviation) p-value of a one-sided binomial test with the null hypothesis that the fraction of rule-identified samples with corrected outcome predictions out of all samples with changed predictions with the use of the external site model is 0.50 (alternative: > 0.50). Rule pruning results in smaller p-values (in bold), on average.

Experiment:		Number of Rules	
Dataset		Without Pruning	With Pruning
1:	CH	1.8 (1.2)	**1.7 (0.5)**
	MIMIC-II	7.8 (9.2)	**1.8 (1.3)**
2:	CH	6.8 (4.3)	**1.7 (0.5)**
	MIMIC-II	14.8 (7.2)	**4.2 (2.8)**

Table 3: Average (standard deviation) number of decision list rules. Decision lists generated with rule pruning are shorter (in bold), on average, and are thus, more interpretable and reliable for use in practice.

Conclusion

Ensuring decision list reliability would enhance insight into beneficial expertise transfer opportunities to improve outcome risk assessments across sites as well as prevent inappropriate transfers that could lead to harmful assessments. This research takes a step to improve decision list rule reliability by incorporating permutation testing with use of a non-parametric ML approach to probability estimation.

Acknowledgments

This work was partially supported by NIH under grants R01HL141916 and R01HL144692, and DARPA under award FA8750-17-2-0130. Thank you to Sebastian Caldas, Dr. Keith A. Dufendach, Dr. Gilles Clermont, Dr. Kyle Miller, and Dr. Artur Dubrawski for research guidance. Thank you to Rachel Burcin and Dr. John Dolan for supporting this research through the CMU RISS program.

References

Caldas, S.; Yoon, J.; Pinsky, M.; Clermont, G.; and Dubrawski, A. 2021. Understanding Clinical Collaborations Through Federated Classifier Selection. In *Proceedings of Machine Learning Research*, volume 149, 126–145.

Frank, E. 2000. Pruning Decision Trees and Lists. Technical report, Dept. of Computer Science, Univ. of Waikato.

Moore, A.; and Schneider, J. 2002. Real-valued All-Dimensions search: Low-overhead rapid searching over subsets of attributes. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, 360–369.

Nguyen, Y. L.; Perrodeau, E.; and et al. 2014. Mechanical ventilation and clinical practice heterogeneity in intensive care units: a multicenter case-vignette study. *Annals of Intensive Care*, (2).