An interpretable model for stroke prediction using rules and Bayesian analysis

Benjamin Letham
Operations Research Center
Massachusetts Institute of Technology
bletham@mit.edu

Cynthia Rudin
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
rudin@mit.edu

Tyler H. McCormick
Department of Statistics
University of Washington
tylermc@u.washington.edu

David Madigan
Department of Statistics
Columbia University
madigan@stat.columbia.edu

ABSTRACT

We aim to produce predictive models that are not only accurate, but are also interpretable to human experts. We introduce a Bayesian method for learning decision lists, a type of interpretable classifier, from data. We use the model to predict stroke in atrial fibrillation patients, and produce predictive models that are as interpretable as the current medical scoring systems that are in widespread use, yet are substantially more accurate.

1. INTRODUCTION

In many domains, interpretability is a fundamental desirable quality in a predictive model [5]. A *decision list* is an interpretable classifier consisting of a series of *if... then...* statements, ending with *else*. The *if* statements define a partition of a set of features and the *then* statements correspond to the outcome of interest. Decision lists are a type of associative classifier, and are similar to models used in the expert systems literature [7], which were among the first successful types of artificial intelligence.

The motivation for our work lies in developing interpretable predictive models using massive observational medical data. Most widely used medical scoring systems are designed to be interpretable, but are not necessarily optimized for accuracy, and are derived from few factors. For instance, the CHADS_2 score is a widely used system for predicting stroke in patients with atrial fibrillation [4]. A patient's score is computed by assigning one "point" each for the presence of congestive heart failure (C), hypertension (H), age 75 years or older (A), and diabetes mellitus (D) and by assigning 2 points for history of stroke (S_2). An updated version called CHA_2DS_2-VASc [8] includes three additional risk factors: vascular disease (V), age 65 to 74 years old (A), and female gender (Sc).

Here we use a Bayesian model, which we call Bayesian Rule Lists (BRL), and MCMC sampling to construct a decision list alternative to the CHADS_2 score from a large database of medical histories. The decision list is learned from a large dataset with many features, which provides better accuracy than the few hand-selected features used in the CHADS_2 score, yet the same level of interpretability. Bayesian Rule Lists are an alternative to decision trees, but are not constructed in a greedy way like decision trees. BRL is instead a new framework that designs a probabilistic model over permutations of association rules to form decision lists.

2. BAYESIAN RULE LISTS

The setting for BRL is multi-class classification, where the set of possible labels is $1, \ldots, L$. In the case of predicting stroke risk, there are only two possible labels: stroke or no stroke. The training data are pairs $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ are the features of observation i, and y_i are the labels, $y_i \in \{1, \ldots, L\}$. We let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$.

2.1 Bayesian association rules and Bayesian decision lists

An association rule $a \rightarrow b$ is an implication with an antecedent a and a consequent b. For the purposes of classification, the antecedent is a boolean function of the feature vector x_i and the consequent b would typically be a label y. A Bayesian association rule has a multinomial distribution over labels as its consequent rather than a single label:

$$a \rightarrow y \sim \text{Multinomial}(\theta).$$

The multinomial probability is then given a prior, leading to a *prior consequent distribution*:

$$\theta|\alpha \sim \text{Dirichlet}(\alpha)$$

Given observations $(\mathbf{x}, \mathbf{y})$ classified by this rule, we let $N_{\cdot,l}$ be the number of observations with label $y_i = l$, and $N =$

$(N_{\cdot,1}, \ldots, N_{\cdot,L})$. We then obtain a *posterior consequent distribution*:

$$\theta|\mathbf{x}, \mathbf{y}, \alpha \sim \text{Dirichlet}(\alpha + N).$$

The core of a Bayesian decision list is an ordered antecedent list $d = (a_1, \ldots, a_m)$. Let $N_{j,l}$ be the number of observations x_i that satisfy a_j but not any $a_1, \ldots, a_{j-1}$, and that have label $y_i = l$. This is the number of observations to be classified by antecedent a_j that have label l. Let $N_{0,l}$ be the number of observations that do not satisfy any of $a_1, \ldots, a_m$, and that have label l. Let $\mathbf{N}_j = (N_{j,1}, \ldots, N_{j,L})$ and $\mathbf{N} = (\mathbf{N}_0, \ldots, \mathbf{N}_m)$.

A Bayesian decision list $D = (d, \alpha, \mathbf{N})$ is an ordered list of antecedents together with their posterior consequents. The posterior consequents are obtained by excluding data that have satisfied an earlier antecedent in the list. Graphically, D is:

if a_1 **then** $y \sim \text{Multinomial}(\theta_1)$, $\theta_1 \sim \text{Dir}(\alpha + \mathbf{N}_1)$
else if a_2 **then** $y \sim \text{Multinomial}(\theta_2)$, $\theta_2 \sim \text{Dir}(\alpha + \mathbf{N_2})$
$\vdots$
else if a_m **then** $y \sim \text{Multinomial}(\theta_m)$, $\theta_m \sim \text{Dir}(\alpha + \mathbf{N}_m)$
else $y \sim \text{Multinomial}(\theta_0)$, $\theta_0 \sim \text{Dir}(\alpha + \mathbf{N}_0)$.

Any observations that do not satisfy any of the antecedents in d are classified using the parameter θ_0, which we call the default rule parameter.

2.2 Antecedent mining

We are interested in forming Bayesian decision lists whose antecedents are a subset of a pre-selected collection of antecedents. We use frequent itemset mining to find a collection of itemsets, which are used as the collection of antecedents. In our experiments, the features were categorical so we used the FP-Growth algorithm [1] which finds all itemsets that satisfy constraints on minimum support and maximum cardinality. This means each itemset applies to a sufficiently large amount of data, and that the itemset is a conjunction of at most C conditions on individual features. Because the goal is to obtain decision lists with few rules and few conditions per rule, we need not include any itemsets that apply only to a small number of observations or have larger numbers of conditions per rule (and are not interpretable). Thus frequent itemset mining allows us to significantly reduce the size of the feature space, compared to considering all possible combinations of features. We let $\mathcal{A}$ represent the complete, pre-mined collection of antecedents, and suppose that $\mathcal{A}$ contains R antecedents with the number of conditions up to C.

2.3 A Bayesian model

Our goal is to sample from the posterior distribution over antecedent lists, which is proportional to its prior times the likelihood function:

$$p(d|\mathbf{x}, \mathbf{y}, \mathcal{A}, \alpha, \lambda, \eta) \propto p(\mathbf{y}|\mathbf{x}, d, \alpha) p(d|\mathcal{A}, \lambda, \eta).$$

Given d, we can compute the posterior consequents required to construct a Bayesian decision list as in Section 2.1. There are three prior hyperparameters that must be specified by the user: α, λ, and η. We will see in Sections 2.4 and 2.5 that these hyperparameters have natural interpretations that suggest the values to which they should be set. We now describe the construction for the prior and likelihood.

2.4 The hierarchical prior for antecedent lists

Suppose the list of antecedents d has length m and antecedent cardinalities $c_1, \ldots, c_m$. The prior probability of d is defined hierarchically as

$$p(d|\mathcal{A}, \lambda, \eta) = p(m|\mathcal{A}, \lambda) \prod_{j=1}^{m} p(c_j|c_1, \ldots, c_{j-1}, \mathcal{A}, \eta) \, p(a_j|a_1, \ldots, a_{j-1}, c_j, \mathcal{A}).$$

We take the distributions for list length m and antecedent cardinality c_j to be Poisson with parameters λ and η respectively, with proper truncation to account for the finite number of antecedents in $\mathcal{A}$. We take $p(a_j|a_1, \ldots, a_{j-1}, c_j, \mathcal{A})$ to be uniform over all antecedents in $\mathcal{A}$ of size c_j, excluding those in $\{a_1, \ldots, a_{j-1}\}$.

This prior has the desirable property that when R is large compared to the desired size of the decision list, as will generally be the case when seeking an interpretable decision list, the prior expected decision list length $\mathbb{E}[m|\mathbf{a}, \lambda]$ is approximately equal to λ. The prior hyperparameter λ can then naturally be set to the prior belief of the list length required to model the data. Similarly, if the number of rules of different sizes is large compared to λ, and η is small compared to C, the prior expected average antecedent cardinality is close to η.

2.5 The likelihood function

The likelihood function follows directly from the definition of a Bayesian decision list. Let $\theta = (\theta_0, \theta_1, \ldots, \theta_m)$ be the consequent multinomial parameters for each antecedent in d, together with the default rule parameter θ_0. Then,

$$p(\mathbf{y}|\mathbf{x}, d, \theta) = \prod_{\substack{j=0,\ldots,m, \\ \sum_l N_{j,l} > 0}} \text{Multinomial}(\mathbf{N}_j|\theta_j),$$

with

$$\theta_j \sim \text{Dirichlet}(\alpha).$$

We can marginalize over θ_j in each Multinomial distribution in the above product, obtaining, through the standard derivation of the Dirichlet-Multinomial distribution,

$$p(\mathbf{y}|\mathbf{x}, d, \alpha) \propto \prod_{\substack{j=0,\ldots,m, \\ \sum_l N_{j,l} > 0}} \frac{\prod_{l=1}^{L} \Gamma(N_{j,l} + \alpha_l)}{\Gamma(\sum_{l=1}^{L} N_{j,l} + \alpha_l)}.$$

The prior hyperparameter α has the usual interpretation of pseudocounts. In our experiments, we set $\alpha_l = 1$ for all l, producing a uniform prior.

2.6 Markov chain Monte Carlo sampling

We do Metropolis-Hastings sampling of d, generating the proposed d^* from the current d_t using one of three options:

1) Move an antecedent in d to a different position in the list. 2) Add an antecedent from $\mathcal{A}$ that is not currently in d. 3) Remove an antecedent from d. Which antecedents to adjust and their new positions are chosen uniformly at random at each step. The option to move, add, or remove is also chosen randomly. This sampling algorithm is related to those used for Bayesian Decision Tree models [3, 2, 10].

For every MCMC run, we ran 3 chains, each initialized independently from a random sample from the prior. We discarded the first half of simulations as burn-in, and then assessed chain convergence using the Gelman-Rubin convergence diagnostic applied to the log-posterior values. We considered chains to have converged when the diagnostic $\hat{R} < 1.05$.

3. STROKE PREDICTION COMPARED TO CHADS$_2$

We applied BRL to the MarketScan Medicaid Multi-State Database (MDCD), which contains administrative claims data for 11.1 million Medicaid enrollees from multiple states. This database forms part of the suite of databases from the Innovation in Medical Evidence Development and Surveillance (IMEDS, http://imeds.reaganudall.org/) program that have been mapped to a common data model [9]. We extracted every patient in the MDCD database with a diagnosis of atrial fibrillation, one year of atrial fibrillation-free observation time prior to the diagnosis, and one year of observation time following the diagnosis (n=12,586). We used as features all medications and conditions in the pre-diagnosis medical history (a total of 4,146), together with age and gender. We chose prior hyperparameters $\lambda = 3$ and $\eta = 1$ to obtain a list of similar complexity to the CHADS$_2$ score, and evaluated the fit using 5-fold cross validation. In this work, we designed a medical scoring system that is strictly better than what is currently in widespread use. It is just as interpretable as the predictive models doctors use for stroke prediction, but more accurate.

In Figure 1 we show a point estimate decision list recovered from one of the folds, chosen as the list with highest likelihood and length and cardinality statistics close to the posterior mean. For each rule we give the the posterior consequent mean and, in parentheses, 95% credible interval. The first half of the decision list focuses on a history of stroke and stroke symptoms, in order of severity. The second half of the decision list includes age factors and vascular disease, which are known risk factors and are included in the CHA$_2$DS$_2$-VASc score.

In Table 1 we report mean AUC across the folds for BRL, CHADS$_2$, CHA$_2$DS$_2$-VASc, and a collection of benchmark machine learning algorithms. These results show that with complexity and interpretability similar to CHADS$_2$, the BRL decision lists performed significantly better at stroke prediction than both CHADS$_2$ and CHA$_2$DS$_2$-VASc, and also significantly outperforms standard decision tree algorithms. We also give the training time in Table 1 for each of the methods on the same, single CPU. The time required for MCMC sampling was less than that required for SVM or Random forest with standard parameter settings, and the resulting simple model in Figure 1 has close performance.

if hemiplegia **and** age>60 **then** *stroke risk* 58.9% (53.8% - 63.8%)
else if cerebrovascular disorder **then** *stroke risk* 47.8% (44.8% - 50.7%)
else if transient ischaemic attack **then** *stroke risk* 23.8% (19.5% - 28.4%)
else if occlusion and stenosis of carotid artery without infarction **then** *stroke risk* 15.8% (12.2% - 19.6%)
else if altered state of consciousness **and** age>60 **then** *stroke risk* 16.0% (12.2% - 20.2%)
else if age≤70 **then** *stroke risk* 4.6% (3.9% - 5.4%)
else *stroke risk* 8.7% (7.9% - 9.6%)

Figure 1: A decision list for determining 1-year stroke risk following diagnosis of atrial fibrillation from patient medical history. The risk given is the mean of the posterior consequent distribution, and in parentheses is the 95% credible interval.

4. CONCLUSIONS

We are working under the hypothesis that many real datasets permit predictive models that can be surprisingly small. This was hypothesized over a decade ago [6], however, we now are starting to have the computational tools to truly test this hypothesis. The BRL method introduced in this work aims to hit the "sweet spot" between predictive accuracy, interpretability, and tractability.

Interpretable models have the benefits of being both concise and convincing. A small set of trustworthy rules can be the key to communicating with domain experts and to allow machine learning algorithms to be more widely implemented and trusted. In practice, a preliminary interpretable model can help domain experts to troubleshoot the inner workings of a complex model, in order to make it more accurate and tailored to the domain. We demonstrated that interpretable models lend themselves to the domain of predictive medicine, but there are a wide variety of domains in science, engineering, and industry, where these models would be a natural choice.

BRL provides an efficient and powerful method for obtaining interpretable models without significantly sacrificing accuracy.

Table 1: Mean, and in parentheses standard deviation, of area under the ROC curve and CPU time for training across 5 folds of cross-validation for stroke prediction.

	AUC	Training time (mins)
BRL point estimate	0.756 (0.007)	21.48 (6.78)
CHADS$_2$	0.721 (0.014)	
CHA$_2$DS$_2$-VASc	0.677 (0.007)	
CART	0.704 (0.010)	12.62 (0.09)
C5.0	0.704 (0.011)	2.56 (0.27)
ℓ_1 logistic regression	0.767 (0.010)	0.05 (0.01)
SVM	0.753 (0.014)	302.89 (8.28)
Random forests	0.774 (0.013)	698.56 (59.66)

5. REFERENCES

[1] C. Borgelt. An implementation of the FP-growth algorithm. In *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*, OSDM '05, pages 1–5, 2005.

[2] H. A. Chipman, E. I. George, and R. E. McCulloch. Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948, 1998.

[3] H. A. Chipman, E. I. George, and R. E. McCulloch. Bayesian treed models. *Machine Learning*, 48(1/3):299–320, 2002.

[4] B. F. Gage, A. D. Waterman, W. Shannon, M. Boechler, M. W. Rich, and M. J. Radford. Validation of clinical classification schemes for predicting stroke. *Journal of the American Medical Association*, 285(22):2864–2870, 2001.

[5] C. Giraud-Carrier. Beyond predictive accuracy: what? In *Proceedings of the ECML-98 Workshop on Upgrading Learning to Meta-Level: Model Selection and Data Transformation*, pages 78–85, 1998.

[6] R. C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11(1):63–91, 1993.

[7] C. T. Leondes. *Expert systems: the technology of knowledge management and decision making for the 21st century*. Academic Press, 2002.

[8] G. Lip, R. Nieuwlaat, R. Pisters, D. Lane, and H. Crijns. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the Euro heart survey on atrial fibrillation. *Chest*, 137:263–272, 2010.

[9] P. Stang, P. Ryan, J. Racoosin, J. Overhage, A. Hartzema, C. Reich, E. Welebob, T. Scarnecchia, and J. Woodcock. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. *Annals of Internal Medicine*, 153:600–606, 2010.

[10] Y. Wu, H. Tjelmeland, and M. West. Bayesian CART: prior specification and posterior simulation. *Journal of Computational and Graphical Statistics*, 16(1):44–66, 2007.