Late-Breaking Developments in the Field of Artificial Intelligence
Papers Presented at the Twenty-Seventh AAAI Conference on Artificial Intelligence

An Interpretable Stroke Prediction Model Using Rules and Bayesian Analysis

Benjamin Letham
Operations Research Center
MIT
Cambridge, MA 02139
bletham@mit.edu

Cynthia Rudin
MIT Sloan School of Management
MIT
Cambridge, MA 02139
rudin@mit.edu

Tyler H. McCormick
Department of Statistics
Department of Sociology
University of Washington
Seattle, WA 98195
tylermc@u.washington.edu

David Madigan
Department of Statistics
Columbia University
New York, NY 10027
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 generative model called the Bayesian List Machine for fitting decision lists, a type of interpretable classifier, to data. We use the model to predict stroke in atrial fibrillation patients, and produce predictive models that are simple enough to be understood by patients yet significantly outperform the medical scoring systems currently in use.

Introduction

In many domains, interpretability is a fundamental desirable quality in a predictive model (Giraud-Carrier 1998). Domain experts tend to prefer models that explain which factors were used to make a particular prediction. 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 (Leondes 2002), 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 (Gage et al. 2001). 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$_2$DS$_2$-VASc (Lip et al. 2010) 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 and Markov chain Monte Carlo 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.

The Bayesian List Machine

We now present a generative model for decision lists which we call the Bayesian List Machine (BLM). The setting is multi-class classification with labels $1, \dots, L$ and training data $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ are the features of observation i and $y_i \in \{1, \dots, L\}$ are the labels. To generate a class label for the ith observation x_i:

1. Generate a list of rules $r = 1, \dots, R$ using a rule-mining algorithm.
2. Sample a permutation over rules π from Prior(p, C).
3. Using this ordering, select the first rule that applies, in that it matches the observed features x_i. Call the rule $\tilde{r}_i$.
4. Draw a label y_i from a Dirichlet-Multinomial distribution $\boldsymbol{\theta}^{(\tilde{r}_i)}$, with Dirichlet parameters $\alpha_1, \dots, \alpha_L$ and counts $n_{\tilde{r}_i 1}, \dots, n_{\tilde{r}_i L}$ for rule $\tilde{r}_i$ chosen in the previous step.

We now give a full description of each step.

In applications with binary features, such as ours, a list of rules can be generated using an algorithm for frequent itemset mining. We used the FP-Growth algorithm (Borgelt 2005) which finds all itemsets that satisfy constraints on minimum support and maximum cardinality.

We chose a prior that favors shorter decision lists (small number of rules before the *else* statement), and prefers rules with a small number of conditional statements (small length of each rule). The parameter C in the prior trades off between horizontal and vertical sparseness. A separate parameter p controls the overall strength of the prior. The prior is:

$$\text{Prior}(\pi) \propto \frac{1}{\left(R_\pi + C\frac{A_\pi}{M}\right)^p},$$

where R_π is the number of rules in the list above the rule corresponding to the *else* statement (called the default rule), A_π is the average length of the rules, and M is the maximum allowed length of the rules (for example, the maximum cardinality constraint used in itemset mining).

An outcome y_i is then generated as a single draw from a Multinomial distribution with $\boldsymbol{\theta}^{(\tilde{r}_i)} = \theta_1^{(\tilde{r}_i)}, \dots, \theta_L^{(\tilde{r}_i)}$ the vector of class probabilities. $\boldsymbol{\theta}^{(\tilde{r}_i)}$ in turn follows a Dirichlet distribution with parameters $\alpha_1, \dots, \alpha_L$, which are set to be weakly informative. Define $\tilde{\mathbf{r}} \in \mathbb{R}^n$ as a vector of rule labels such that element $\tilde{r}_i = r$ if x_i is classified by rule r. This is used to compute multinomial counts $n_{r\ell}$ for each rule r and class ℓ as the number of observations x for which r was the first rule in the list that applied, and which have label $y = \ell$. The likelihood then follows the Dirichlet-Multinomial distribution:

$$p(y_1, \dots, y_n|\alpha_1, \dots, \alpha_L, \tilde{\mathbf{r}}) \propto \prod_{r=1}^{R} \frac{\prod_{\ell=1}^{L} \Gamma(n_{r\ell} + \alpha_\ell)}{\Gamma(\sum_{\ell=1}^{L} n_{r\ell} + \alpha_\ell)}.$$

In practice, many datasets are extremely imbalanced. For example, many fewer medical patients have a stroke than do not have a stroke. In such circumstances, we might simply weight the counts in the likelihood, by replacing $n_{r\ell}$ with $n_{r\ell}L/\mathbb{P}(y = \ell)$.

We obtain a posterior over decision lists using Metropolis sampling, with three step types to propose a new list π^* from the current list π_t: 1) Swap two rules on the decision list. 2) Add a rule to the decision list (a rule ordered below the *else* default rule). 3) Remove a rule from the decision list (move it below the default rule). The step types and which rules to move are chosen independently and uniformly at random. Steps are accepted or rejected according to the Metropolis sampling rules, and sampling proceeds until chain convergence. This sampling algorithm is related to those used for Bayesian Decision Tree models (Chipman, George, and McCulloch 1998). We make predictions in our experiments using the decision list with highest posterior probability.

Stroke prediction compared to CHADS$_2$

We applied BLM 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 that the Observational Medical Outcomes Partnership (OMOP, http://omop.fnih.org) has mapped to a common data model (Stang et al. 2010). 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). Of these, 1,786 (14%) had a stroke within a year of the atrial fibrillation diagnosis. This is a much larger dataset than the one originally used to develop the CHADS$_2$ score (n=1,733 with 94 strokes). 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 to obtain a list of similar complexity to the CHADS$_2$ score, and evaluated the fit using 5-fold cross validation.

if hemiplegia **then** *stroke risk 58.0%* (14.5%)
else if cerebrovascular disorder **then** *stroke risk 46.6%* (12.5%)
else if transient ischaemic attack **and** essential hypertension
  then *stroke risk 23.2%* (8.3%)
else if occlusion and stenosis of carotid artery
  then *stroke risk 16.4%* (7.8%)
else if age≤60 **then** *stroke risk 3.7%* (7.4%)
else *stroke risk 8.5%*

Figure 1: Decision list for determining 1-year stroke risk following diagnosis of atrial fibrillation from patient medical history. For each rule we give in parentheses the base risk for all patients that make it to that depth on the list.

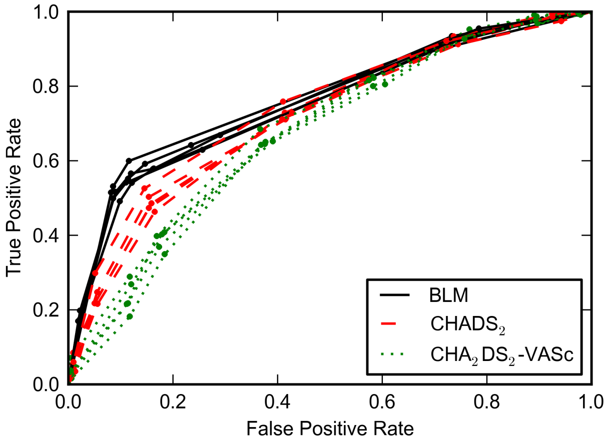

Figure 2: ROC curves for stroke prediction on the MDCD database for each of 5 folds of cross-validation, for BLM (solid), CHADS$_2$ (dashed), and CHA$_2$DS$_2$-VASc (dotted).

In Figure 1 we show the decision list recovered from one of the folds. For each rule we give the stroke risk estimated from the training data as the number of patients satisfying that rule (and no preceding rule) that had a stroke. We give in parentheses the stroke risk across all patients that did not satisfy any of the preceding rules in the list. For example, the second line in the list indicates that among patients without hemiplegia the stroke risk was 12.5%, which increased to 46.6% when patients had a cerebrovascular disorder. 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.

Figure 2 shows ROC curves for all 5 folds for BLM, CHADS$_2$, and CHA$_2$DS$_2$-VASc. In Table 1 we report mean AUC (in parentheses, standard deviation) across the folds. These results show that with complexity and interpretability similar to CHADS$_2$, the BLM decision lists performed significantly better at stroke prediction than both CHADS$_2$ and CHA$_2$DS$_2$-VASc ($p < 0.01$, t-test). Interestingly, we also found that CHADS$_2$ outperformed CHA$_2$DS$_2$-VASc despite CHA$_2$DS$_2$-VASc being an extension to CHADS$_2$. This is likely because the model for the CHA$_2$DS$_2$-VASc score, in which risk factors are added linearly, is a poor model of actual stroke risk, and highlights the difficulty in constructing

	BLM	CHADS$_2$	CHA$_2$DS$_2$-VASc
AUC	0.750 (0.007)	0.721 (0.014)	0.677 (0.007)

Table 1: Mean AUC for stroke prediction with standard deviation in parentheses, across 5 folds of cross-validation.

these interpretable models manually.

References

Borgelt, C. 2005. 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, 1–5.

Chipman, H. A.; George, E. I.; and McCulloch, R. E. 1998. Bayesian CART Model Search. *Journal of the American Statistical Association* 93(443):935–948.

Gage, B.; Waterman, A.; Shannon, W.; Boechler, M.; Rich, M.; and Radford, M. 2001. Comparing Hospitals on Stroke Care: The Need to Account for Stroke Severity. *Journal of the American Medical Association* 285:2864–2870.

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

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

Lip, G.; Nieuwlaat, R.; Pisters, R.; Lane, D.; and Crijns, H. 2010. 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.

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