Learning Interpretable Clinical Prediction Rules using Threshold Group Testing

Amin Emad[*]
Electrical and Computer Engineering
Univ. of Illinois at Urbana-Champaign
emad2@illinois.edu

Kush R. Varshney, Dmitry M. Malioutov
Thomas J. Watson Research Center
IBM Research
{krvarshn, dmalioutov}@us.ibm.com

1 Introduction

There is a growing belief that in the face of high complexity, checklists and other simple scorecards or algorithms can significantly improve people's performance on decision-making tasks [1]. An example of such a tool in medicine, the *clinical prediction rule*, is a simple decision-making rubric that helps physicians estimate the likelihood of a patient having or developing a particular condition in the future [2]. An example of a clinical prediction rule for estimating the risk of stroke is known as the CHADS$_2$ score [3]: the health worker determines which of five diagnostic indicators a patient exhibits and adds the corresponding points together. (The five conditions are congestive heart failure, hypertension, age ≥ 75, diabetes mellitus, and prior stroke.) The higher the total point value is, the greater the likelihood the patient will develop a stroke. This rule was manually crafted by health workers, notably contains few conditions, and is extremely interpretable by people.

Recent machine learning research has attempted to learn clinical prediction rules that generalize accurately from large-scale electronic health record data rather than relying on manual development [4, 5]. The key aspect of the problem is maintaining the simplicity and interpretability of the learned rule: similar to the hand-crafted version rather than a complicated, uninterpretable 'black-box' model. Such transparency is critical for trust and adoption by users, and is not exhibited by models from, e.g., deep learning, ensemble methods, or even l_1-regularized logistic regression.

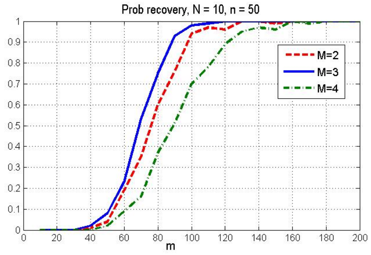
In this work, we build upon our recent research on the supervised learning of interpretable classification rules using Boolean compressed sensing ideas [6, 7]. With the same goal as [4, 5], we develop a method for learning interpretable clinical prediction rules using sparse signal representation techniques. In our previous work, the form of the classifier was a sparse AND-rule or OR-rule ("1-of-N" and "N-of-N" rule tables), whereas here, we focus on "M-of-N" rule classifiers that provides a close match to a clinical decision rule. In [6], the Boolean compressed sensing formulation had a close connection with the group testing problem whereas here, the connection is to the threshold group testing (TGT) problem [8].

2 Formulation

We first introduce TGT (without a gap). Let n, m, and d denote the total number of subjects, the number of tests, and the number of defectives, respectively. Let $\mathbf{A} \in \{0,1\}^{m \times n}$ be a binary matrix, representing the assignment of subjects to each test: for $1 \le i \le m$ and $1 \le j \le n$, $\mathbf{A}(i,j) = 1$ if the jth subject is present in the ith test and $\mathbf{A}(i,j) = 0$, otherwise. Let $\mathcal{D}_t$ be the true set of defectives and let $\mathbf{w}_t \in \{0,1\}^n$ be the binary vector representing which subject is a defective; also, let $\mathbf{y} \in \{0,1\}^m$ be the binary vector representing the error-free results of the tests. In the TGT model, one has

$$\mathbf{y} = f_\eta(\mathbf{A}\mathbf{w}_t), \tag{1}$$

[*] We thank Sanjeeb Dash, Benjamin Letham and Cynthia Rudin for helpful discussions.



(a)

Condition	Write 1 if True
mean texture > 16.57	
worst perimeter > 120.26	
worst area > 724.48	
worst smoothness > 0.125	
worst concave points > 0.179	
Total	

(b)

Figure 1: (a) Phase transition diagram for LP recovery of TGT models (with known ground-truth) with different number of non-zeros M vs. the number of samples: $n = 50$, $N = 10$. A is i.i.d. binary with 0.25 probability of 1. (b) Learned clinical prediction rule from WDBC data set.

where $f_\eta(\cdot)$ is a quantizing function with threshold η, such that $f_\eta(x) = 0$ if $x < \eta$ and $f_\eta(x) = 1$ if $x \geq \eta$. The goal is to recover the unknown vector $\mathbf{w}$ (and also η if needed) given the test matrix $\mathbf{A}$ and the vector of test results $\mathbf{y}$. This model is a special case of semi-quantitative group testing [9, 10].

In clinical prediction rule learning, the goal is to learn an interpretable function $\hat{y}(\cdot) : \mathcal{X} \rightarrow \{0,1\}$, given m labeled training samples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, where $\mathbf{x}_i \in \mathcal{X}$ are features and $y_i \in \{0,1\}$ are Boolean labels indicating the presence or absence of a medical condition.[1] To formulate this problem as TGT, we form the vector of test results according to $\mathbf{y}(i) = y_i$, $i = 1, 2, \ldots, m$; also, we form the test matrix $\mathbf{A}$ according to $\mathbf{A}(i,j) = a_j(\mathbf{x}_i)$, where $a_j(\cdot) : \mathcal{X} \rightarrow \{0,1\}$, $j = 1, \ldots, n$, are simple Boolean terms (e.g. age ≥ 75). Furthermore, we assume that at most d simple Boolean terms, govern the relationship between the labels and the features, i.e. $|\mathcal{D}_t| \leq d$, and this sparse set of terms are encoded in the unknown sparse vector $\mathbf{w}_t$. In addition, we assume that this relationship has the form of a "M-of-N" rule table; in other words, $y_i = 1$ if at least M terms of $\mathcal{C}$ are satisfied and $y_i = 0$, otherwise. Therefore, by setting $\eta = M$ and $d = N$, we can write this relationship as (1). Consequently, in order to find the set of interpretable rules corresponding to a "M-of-N" rule table, we need to recover the sparse vector $\mathbf{w}_t$ given $\mathbf{A}$ and $\mathbf{y}$.

A vector $\mathbf{w}$ that satisfies the constraints imposed by the TGT model in (1), must also satisfy the pair of ordinary linear inequalities $\mathbf{A}_\mathcal{P}\mathbf{w} \geq \eta\mathbf{1}$ and $\mathbf{A}_\mathcal{Z}\mathbf{w} < \eta\mathbf{1}$, where $\mathcal{P} = \{i | \mathbf{y}(i) = 1\}$ is the set of positive tests, $\mathcal{Z} = \{i | \mathbf{y}(i) = 0\}$ is the set of negative tests, and $\mathbf{A}_\mathcal{P}$ and $\mathbf{A}_\mathcal{Z}$ are the corresponding subsets of rows of $\mathbf{A}$. The vector $\mathbf{1}$ is an all-one vector. For sparsity, we minimize the ℓ_1 norm of $\mathbf{w}$:

$$\begin{aligned} \min \quad & \|\mathbf{w}\|_1 \\ \text{s.t.} \quad & w_j \in \{0,1\},\ j = 1, \ldots, n \\ & \mathbf{A}_\mathcal{P}\mathbf{w} \geq \eta\mathbf{1}, \quad \mathbf{A}_\mathcal{Z}\mathbf{w} < \eta\mathbf{1}. \end{aligned} \tag{2}$$

In practical scenarios, the sets $\mathcal{P}$ and $\mathcal{Z}$ result from noisy labels and thus, we may introduce slack variables as in [6]. If η is unknown, we treat it as a free variable and jointly solve (2) with the extra constraint $1 \leq \eta \leq n$. In general, the problem may be optimized using integer-programming solvers. For medium-sized data sets with a few thousand examples, the solution can often be obtained efficiently using branch and bound. We can also relax the binary constraints on the w_j to interval constraints $0 \leq w_j \leq 1$ to obtain a linear program (LP). Under favorable conditions when the $\mathbf{A}$ matrix satisfies (d, η)-disjunctness [11], we can show that this LP relaxation produces integral solutions. Phase transition diagrams for sparse-rule recovery are shown in Fig. 1(a).

We illustrate the promise of this approach by applying it to the breast cancer diagnosis problem using the Wisconsin Diagnostic Breast Cancer data set from the UCI Machine Learning Repository. The clinical prediction rule that is learned from the data set is given in Fig. 1(b). The learned rule requires that at least 3 of the 5 conditions are satisfied to diagnose a positive case. The training error for this M-of-N rule is 2.8%. The proposed approach may also be applied to other clinical data sets for learning prediction rules for other health conditions.

[1] Note that in typical group testing formulations, the presence or absence of disease in patients is encoded in $\mathbf{w}$, but here it is given in $\mathbf{y}$.

References

[1] A. Gawande, *The Checklist Manifesto: How To Get Things Right*. New York, NY: Metropolitan Books, 2009.

[2] S. T. Adams and S. H. Leveson, “Clinical prediction rules,” *Brit. Med. J.*, vol. 344, p. d8312, Jan. 2012.

[3] 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,” *J. Am. Med. Assoc.*, vol. 258, no. 22, pp. 2864–2870, 13 Jun. 2001.

[4] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, “An interpretable stroke prediction model using rules and Bayesian analysis,” *Proc. Workshops AAAI Conf. Artif. Intell.*, Jul. 2013.

[5] B. Ustun, M. B. Westover, C. Rudin, and M. T. Bianchi, “Clinical prediction models for sleep apnea: The importance of medical history over symptoms,” *J. Clin. Sleep Med.*, Aug. 2015.

[6] D. M. Malioutov and K. R. Varshney, “Exact rule learning via Boolean compressed sensing,” in *Proc. Int. Conf. Mach. Learn.*, Atlanta, GA, Jun. 2013, pp. 765–773.

[7] A. Emad, K. R. Varshney, and D. M. Malioutov, “A semiquantitative group testing approach for learning interpretable clinical prediction rules,” *Proc. Signal Processing with Adaptive Sparse Structured Representations (SPARS) Workshop*, Cambridge, UK, Jul. 2015.

[8] P. Damaschke, “Threshold group testing,” *General Theory of Information Transfer and Combinatorics, LNCS*, vol. 4123, pp. 707–718, 2006.

[9] A. Emad and O. Milenkovic, “Semiquantitative group testing,” *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4614–4636, Aug. 2014.

[10] A. Emad, “New Group Testing Paradigms: From Practice to Theory,” University of Illinois, Ph.D. Dissertation, 2015.

[11] H-. B. Chen and H-. L. Fu, “Nonadaptive algorithms for threshold group testing”, *Discrete Appl. Math.*, vol. 157, pp. 1581–1585, 2009.