A Semiquantitative Group Testing Approach for Learning Interpretable Clinical Prediction Rules

Amin Emad[1], Kush R. Varshney[2], and Dmitry M. Malioutov[2]
[1]University of Illinois at Urbana-Champaign, 1308 West Main Street, Urbana, Illinois 61801
[2]IBM Thomas J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, New York 10598

I. 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, known as the CHADS$_2$ score, is shown in Table I [3]. The health worker determines which of the five diagnostic indicators a patient exhibits and adds the corresponding points together. 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 and notably contains few conditions with small integer point values 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., l_1-regularized logistic regression [6].

In this work, we build upon our recent research on the supervised learning of interpretable classification rules using Boolean compressed sensing ideas [7]. With the same goal as [4], [5], we develop a method for learning interpretable clinical prediction rules using sparse signal representation techniques. In that previous work of ours, the form of the classifier was a sparse AND-rule or OR-rule whereas here, we would like to find a sparse set of medical conditions or features with small integer coefficients that are added together to produce a score. Such a model is between the "1-of-N" and "N-of-N" forms implied by OR-rules and AND-rules. In [7], the Boolean compressed sensing formulation had a close connection with the group testing problem whereas here, the connection is to the *semiquantitative* group testing (SQGT) problem [8].

II. Formulation

We first introduce SQGT. Let $\mathbf{A} \in [q]^{m\times n}$ be a q-ary test matrix, where $[q] = \{0, 1, \dots, q-1\}$. Let $\mathbf{w} \in [2]^n$ be an unknown sparse binary vector and let $\mathbf{y} \in [Q]^m$ be a Q-ary vector representing the results of tests. We assume that $\mathbf{y} = f_\eta(\mathbf{A}\mathbf{w})$, where $f_\eta(\cdot)$ is a quantizing function with thresholds $\boldsymbol{\eta} = \{\eta_0 = 0, \eta_1, \eta_2, \dots, \eta_Q\}$. More precisely, $f_\eta(x) = r$ if $\eta_r \leq x < \eta_{r+1}$. Note that we have the standard binary group testing problem used in [7] when $q = 2$, $Q = 2$, $\eta_1 = 1$, and $\eta_2 = \infty$. Throughout the remainder of this paper, we assume that $Q = 2$; however, the formulation and the algorithm can be generalized to include $Q > 2$.

In clinical prediction rule learning, the goal is to learn an interpretable function $\hat{y}(\cdot) : \mathcal{X} \to \{0, 1\}$, given m labeled training samples $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$, where $\mathbf{x}_i \in \mathcal{X}$ are features and $y_i \in [2]$ are Boolean labels indicating the presence or absence of a medical condition.[1] We form a matrix $\mathbf{A}$ with elements $a_{ij} = a_j(\mathbf{x}_i)$, where $a_j(\cdot) : \mathcal{X} \to [q]$, $j = 1, \dots, n$, are Boolean terms (e.g. age ≥ 75) multiplied by small positive integers. In particular, there may be several columns of $\mathbf{A}$ corresponding to the same Boolean term with different positive integer multiples to allow the determination of point values in the clinical prediction rule.

In the framework of [7], $\hat{y}$ is encoded by $\mathbf{w}$, the sparse solution or approximation to the Boolean matrix-vector product equation $\mathbf{y} = \mathbf{A} \vee \mathbf{w}$, which leads to a "1-of-N" rule table. We generalize this model to "M-of-N" rule tables, which classify a sample as $y = 1$ if at least M out of N terms are satisfied. For a fixed properly chosen value of M, one can easily follow the SQGT notation with two thresholds where $\eta_1 = M$ and $\eta_2 = \infty$.

A vector $\mathbf{w}$ that satisfies the constraint imposed by the SQGT model with thresholds $\eta_1 = M$ and $\eta_2 = \infty$, must also satisfy the pair of ordinary linear inequalities $\mathbf{A}_\mathcal{P}\mathbf{w} \geq M\mathbf{1}$ and $\mathbf{A}_\mathcal{Z}\mathbf{w} < M\mathbf{1}$, where $\mathcal{P} = \{i | y_i = 1\}$ is the set of positive samples, $\mathcal{Z} = \{i | y_i = 0\}$ is the set of negative samples, and $\mathbf{A}_\mathcal{P}$ and $\mathbf{A}_\mathcal{Z}$ are the corresponding subsets of rows of $\mathbf{A}$. To learn the prediction rule solution, we apply two relaxations: first, we minimize the l_1 norm instead of minimizing $||\mathbf{w}||_0$, and second, relax the binary constraint on $\mathbf{w}$. If M or an estimate of M is known in advance, we formulate the problem as

$$\begin{aligned} \min \quad & \sum_{j=1}^{n} w_j \qquad (1) \\ \text{s.t.} \quad & 0 \leq w_j \leq 1,\ j = 1, \dots, n \\ & \mathbf{A}_\mathcal{P}\mathbf{w} \geq M\mathbf{1},\ \mathbf{A}_\mathcal{Z}\mathbf{w} < M\mathbf{1}, \end{aligned}$$

We also introduce slack variables as in [7]. If M is unknown, we treat it as a free variable and jointly solve (1) with the extra constraint $1 \leq M \leq n(q-1)$.

III. Discussion

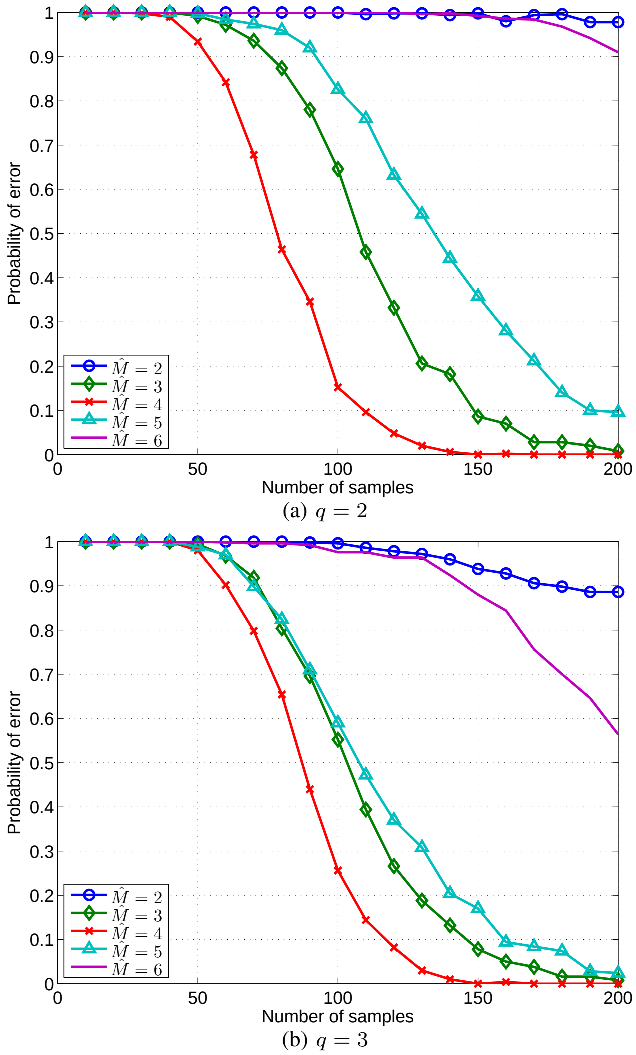
We have generalized [7] for the important task of learning interpretable clinical prediction rules that are the sum of point values. The formulation is more compact than [4], [5] and can be theoretically analyzed from the perspective of SQGT [8]. As a toy example, Fig. 1 demonstrates the effect of knowing a good estimate of M for recovery on the probability of error for two alphabet sizes, $q = 2, 3$.

An interesting extension is to let $\eta_1 = M$ be learned from the training samples rather than being chosen by the user. In this case, the problem can be formulated as a joint learning problem in which $\eta_1 = M$ and $\mathbf{w}$ (and the N of the "M-of-N") are learned. Given that $M = 1$ is one of the possible choices for this joint learning problem, it is clear that the rules learned using this approach will outperform the rules learned from the approach in [7].

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

TABLE I: CHADS$_2$ Clinical Prediction Rule for Stroke Risk

Condition	Points
congestive heart failure	1
hypertension	1
age ≥ 75	1
diabetes mellitus	1
prior stroke, transient ischemic attack, or thromboembolism	2



(a) $q = 2$

(b) $q = 3$

Fig. 1: Average probability of error vs. m for different alphabet sizes using 500 random trials. In each trial, $n = 50$, $N = 10$, and the true value of the threshold $M_{\text{true}} = 4$ were fixed; each entry of the matrix $\mathbf{A}$ was generated randomly according to an independent identically distributed (i.i.d.) distribution. For $q = 2$ an entry was equal to 1 with probability 0.25 and equal to 0 otherwise. For $q = 3$, an entry was equal to 2 with probability 0.1, equal to 1 with probability 0.15, and equal to 0 otherwise. Different values of M (denoted by $\hat{M}$ in the figures) were used to recover the sparse vector $\mathbf{w}$ using the algorithm formulated in (1).

Acknowledgment

The authors thank C. Rudin, B. Letham, and O. Milenkovic for discussions.

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, "Building interpretable classifiers with rules using Bayesian analysis," Dept. Stat., Univ. Washington, Tech. Rep. 609, Dec. 2012.

[5] B. Ustun and C. Rudin, "Methods and models for interpretable linear classification," available at http://arxiv.org/pdf/1405.4047, Oct. 2014.

[6] A. Y. Ng, "Feature selection, l_1 vs. l_2 regularization, and rotational invariance," in *Proc. Int. Conf. Mach. Learn.*, Banff, Canada, Jul. 2004, p. 78.

[7] 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.

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