LEARNING INTERPRETABLE CLASSIFICATION RULES USING SEQUENTIAL ROW SAMPLING

Sanjeeb Dash, Dmitry M. Malioutov, and Kush R. Varshney

IBM Research
1101 Kitchawan Road, Yorktown Heights, NY 10598, USA

ABSTRACT

In our previous work we have presented an approach to learn interpretable classification rules using a Boolean compressed sensing formulation. Our approach uses a linear programming (LP) relaxation and allows us to find interpretable (sparse) classification rules that achieve good generalization accuracy. However, the resulting LP representation for problems with either a large number of samples or large number of continuous features tends to become challenging for off-the-shelf LP solvers. We have explored a screening approach which allows us to dramatically reduce the number of active features without sacrificing optimality. In this work we explore reducing the number of samples in a sequential setting where we can certify reaching a near-optimal solution while only solving the LP on a small fraction of the available data points. In a batch setting this approach can dramatically reduce the computational complexity of the rule-learning LP formulation. In an online setting we derive stochastic upper and lower bounds on the the LP objective for unseen samples. This allows early stopping when we detect that the classifier will not change significantly with additional samples. The upper bounds are related to the learning curve literature in machine learning, and our lower bounds appear not to have been explored. Finally, we discuss a quick approach to compute the complete regularization path balancing rule interpretability versus accuracy.

Index Terms— Linear programming duality, rule learning, row sampling, sparse signal approximation, supervised classification

1. INTRODUCTION

One of the guiding principles for successful applications of machine learning is Occam's razor: among the models that are supported by the data, pick the one that is the simplest. In addition to ensuring that the observed empirical loss on the training set will be a good predictor (generalize) to the error on the test set, keeping the model simple helps it to be interpretable, i.e. able to provide intuition to the human analysts examining it. We restrict ourselves to binary classification rules, and study the question of how much data is sufficient to learn a classification rule under a budget of interpretability. In particular, via linear programming (LP) duality theory we can detect when we have obtained a sufficient number of training examples to learn a near-optimal classification rule.

In recent work [1], a formulation for learning interpretable classification rules was proposed based on tools of Boolean compressed sensing. Classification rules on their own are already among the most well accepted and trusted classification techniques by practitioners, precisely due to the insight they provide. The formulation in [1] goes a step further by explicitly balancing the objectives of interpretability (as encoded by the sparsity in the number of terms used by the rule) versus classification accuracy, and models this problem as a binary optimization problem with a Lasso-like LP relaxation.

We assume that we have access to a sequence of i.i.d. training samples (features and labels) for a binary classification problem. This can be viewed either as an online classification setting, or as a way to obtain a near-optimal classification rule while examining only a small subset of the data in a batch setting. By considering the linear program based on the available samples as part of a bigger LP based on all the available samples, we can develop upper and lower bounds on the objective function for the bigger LP by carefully *extending* the solution of the smaller LP. We consider two ways to measure the duality gap, one for the stochastic setting where we compute the expected size of the duality gap, and another where we can compute the duality gap exactly by an inexpensive linear scan of the remaining samples not used in the smaller LP.

In related work, the *learning curve* literature in machine learning [2–5] has considered how the generalization error evolves as a function of the received number of samples. In the context of ordinary (non-Boolean) compressed sensing, [6] has developed sequential tests to establish that a sufficient number of measurements has been obtained to recover the correct sparse signal (or its accurate approximation). A somewhat different flavor of reducing the number of training samples for support vector machine classification uses screening techniques to identify those training samples that are guaranteed to not join the support vector set [7]. Our previous work on screening for Boolean compressed sensing-based rule learning screened the features, not the training samples [8].

The outline of the paper is as follows: in Section 2 we review the Boolean compressed sensing approach to learn interpretable classification rules. Section 3 describes our formulation for obtaining the duality-based bounds on optimality in the batch setting. Stochastic bounds in the sequential setting are presented in Section 3.1. We present numerical experiments on large-scale machine learning datasets in Section 4 showing that one can obtain accurate near-optimal interpretable classification rules while being trained only a small subset of the training samples. We also describe how to efficiently compute the solution path to balance accuracy and interpretability and allow cost-sensitive classification in Section 5.

2. LEARNING INTERPRETABLE CLASSIFICATION RULES VIA BOOLEAN COMPRESSED SENSING

Interpretable classification rules, such as:

- a breast cancer patient will not have long-term survival if she has greater than nine nodes *and* is greater than or equal to 40 years of age *and* is less than 60 years of age [9];
- an iris is of species versicolor if its petal is less than or equal to 5.350 cm in length *and* its petal is less than or equal to

1.700 cm in width *and* its petal is greater than 0.875 cm in width [1]; and

- a salesman will voluntarily resign if his job role is specialty software sales rep *and* his base salary is less than or equal to $75,168 *and* his months since promoted is less than or equal to 30 *and* his months since promoted is greater than 13 *and* his compensation plan is not quota-based [10],

are often more actionable and trusted by human decision makers than more opaque classification algorithm outputs (e.g. neural networks, random forests) because they can be easily understood [1, 9, 11].

In [1], we posed the interpretable classification rule learning problem as one of Boolean compressed sensing (CS), which attempts to recover a sparse binary vector from a collection of binary measurements which computes disjunctions of subsets of its entries [12–15]. Viewing these measurements as matrix multiplication by a binary sensing matrix in the Boolean algebra, where disjunction and conjunction replace linear algebraic addition and multiplication, establishes a close connection to traditional CS [16]. Sparsity comes into the rule learning formulation because the three terms in the cancer survival example, the three terms in the botany example, and the five terms in the worker attrition example are selected from a large dictionary of potential terms. Non-zero entries in a sparse binary vector dictate which terms are included in the decision rule, and all other potential terms correspond to zeros in the vector.

Formally, in the supervised classification problem, we are given m i.i.d. labeled training samples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, where the $\mathbf{x}_i \in \mathcal{X}$ are the features and the $y_i \in \{0, 1\}$ are the Boolean labels. We would like to learn a function $\hat{y}(\cdot) : \mathcal{X} \to \{0, 1\}$ that will accurately generalize to classify unlabeled feature vectors drawn from the same distribution as the training samples. We represent individual Boolean terms derived from the features, such as 'patient is less than 30 years of age,' by functions $a_j(\cdot) : \mathcal{X} \to \{0, 1\}$, $j = 1, \ldots, n$. Then for each of the training samples, we can calculate the truth value for each of the terms, leading to an $m \times n$ truth table $\mathbf{A}$ with entries $a_{ij} = a_j(\mathbf{x}_i)$. Writing the true labels of the training set as a vector $\mathbf{y} \in \{0, 1\}^m$, we have:

$$\mathbf{y} = \mathbf{A} \vee \mathbf{w} \oplus \mathbf{n}, \tag{1}$$

where $\mathbf{w} \in \{0, 1\}^n$ is the sparse vector to be learned that indicates which terms are included in the decision rule, and $\mathbf{n}$ is noise that flips some values through the exclusive disjunction operation. The notation $\mathbf{y} = \mathbf{A} \vee \mathbf{w}$ is shorthand for:

$$y_i = \bigvee_{j=1}^{n} a_{ij} \wedge w_j, \quad i = 1, \ldots, m. \tag{2}$$

As in the standard sparse signal recovery problem, we would like to find $\mathbf{w}$ satisfying (1) while keeping $\|\mathbf{w}\|_0$ and the noise $\mathbf{n}$ small. Expressing the Boolean constraint $\mathbf{y} = \mathbf{A} \vee \mathbf{w}$ through ordinary linear equalities and inequalities, relaxing the ℓ_0 problem to the ℓ_1 problem, relaxing the binary constraint on the vector $\mathbf{w}$ to $0 \leq \mathbf{w} \leq 1$, and introducing slack variables to account for noise, the rule learning problem is captured in the following LP [1]:

$$\begin{aligned} \min \quad & \sum_{j=1}^{n} w_j + \lambda \sum_{i=1}^{m} \xi_i \\ \text{s.t.} \quad & 0 \leq w_j \leq 1,\ j = 1, \ldots, n \\ & 0 \leq \xi_i \leq 1,\ i \in \mathcal{P},\ \ 0 \leq \xi_i,\ i \in \mathcal{Z} \\ & \mathbf{A}_{\mathcal{P}}\mathbf{w} + \boldsymbol{\xi}_{\mathcal{P}} \geq \mathbf{1} \\ & \mathbf{A}_{\mathcal{Z}}\mathbf{w} = \boldsymbol{\xi}_{\mathcal{Z}}, \end{aligned} \tag{3}$$

where the regularization parameter λ trades training error and the sparsity of $\mathbf{w}$ (sparsity provides generalizability and interpretability), $\mathbf{1}$ is the vector of all ones of appropriate dimension, $\mathcal{P}$ indexes the set of positive training samples, $\mathcal{Z}$ indexes the set of zero-valued training samples, $\mathbf{A}_{\mathcal{P}}$ and $\mathbf{A}_{\mathcal{Z}}$ are the corresponding rows of $\mathbf{A}$, and $\boldsymbol{\xi}_{\mathcal{P}}$ and $\boldsymbol{\xi}_{\mathcal{Z}}$ are the corresponding slack variables. When we constrain the variables w in the LP in (3) to be binary, we get an integer program (IP), which is the true problem to be solved; solving LP yields a lower bound on the optimal solution value of IP, and one can obtain an approximate solution to IP by rounding an optimal solution of LP, or an exact solution to IP via branch-and-bound. The learned decision rule $\hat{y}(\mathbf{x})$, which is a conjunction of Boolean terms, is obtained from the LP solution $\hat{\mathbf{w}}$ [1].

The LP becomes very large when the number of samples m is very large. We address this issue in this paper through row sampling, as discussed in the next section.

3. ROW SAMPLING

We describe how to find interpretable rules but avoid solving very large LPs. Suppose that we have a large number $\bar{m}$ of samples available, and we believe that we can learn a near-optimal interpretable classifier from a much smaller subset of $m \ll \bar{m}$ samples. We proceed to develop a certificate which shows that when m is large enough, the solution of the LP in (3) on the smaller subset of samples also achieves a near optimal solution on the full data-set.

To compare the solutions of LPs defined with different number of samples, we divide the objective by the number of samples to obtain error rates rather than raw errors. Also, as we have seen in [8] we can drop the upper bounds on ξ and $\mathbf{w}$ without affecting the solution:

$$\begin{aligned} \frac{1}{m} \min_{\mathbf{w}, \boldsymbol{\xi}} \quad & \frac{1}{\lambda} \sum_{j=1}^{n} w_j + \sum_{i=1}^{m} \xi_i \\ \text{s.t.} \quad & 0 \leq w_i,\ 0 \leq \xi_j, j = 1, \ldots, n,\ i = 1, \ldots, m \\ & \mathbf{A}_{\mathcal{P}}\mathbf{w} + \boldsymbol{\xi}_{\mathcal{P}} \geq \mathbf{1} \\ & \mathbf{A}_{\mathcal{Z}}\mathbf{w} = \boldsymbol{\xi}_{\mathcal{Z}}. \end{aligned} \tag{4}$$

Let $(\hat{\mathbf{w}}^m, \hat{\boldsymbol{\xi}}^m)$ and $(\hat{\mathbf{w}}^{\bar{m}}, \hat{\boldsymbol{\xi}}^{\bar{m}})$ be the optimal solutions for the small LP with m samples, and large LP with $\bar{m}$ samples in (4). Let f_m and $f_{\bar{m}}$ be the corresponding (scaled) optimal objective values, and let f_m^* and $f_{\bar{m}}^*$ be the corresponding IP optimal objective values. We denote the data-matrices for small LP as $\mathbf{A}$, $\mathbf{A}_{\mathcal{P}}$, $\mathbf{A}_{\mathcal{Z}}$ and the data matrices for the large LP as $\bar{\mathbf{A}}$, $\bar{\mathbf{A}}_{\mathcal{P}}$, $\bar{\mathbf{A}}_{\mathcal{Z}}$. The first m rows of $\bar{\mathbf{A}}$ constitute $\mathbf{A}$ and the first p entries of $\bar{\mathbf{A}}_{\mathcal{P}}$ form $\mathbf{A}_{\mathcal{P}}$. Since we have error rates in the objective, we can compare objective values for different values of m, and we have that $f_m \to f_{\bar{m}}$ as $m \to \bar{m}$.

We would like to bound $|f_m - f_{\bar{m}}|$ and $|f_m^* - f_{\bar{m}}^*|$ without solving the large LP and IP respectively. We consider two scenarios: in a deterministic scenario we allow a simple linear scan over the remaining samples that is much cheaper than solving the large LP. For the stochastic scenario in Section 3.1 we receive a small number of additional i.i.d. samples to evaluate the expected duality gap. We first consider the deterministic case and show how to extend the primal and the dual solutions of the small LP and obtain both a lower and an upper bound on the solution of the large LP and IP.

To create a feasible primal solution for the large LP we can extend the vector $\hat{\mathbf{w}}^m$ from the small LP by computing the associated

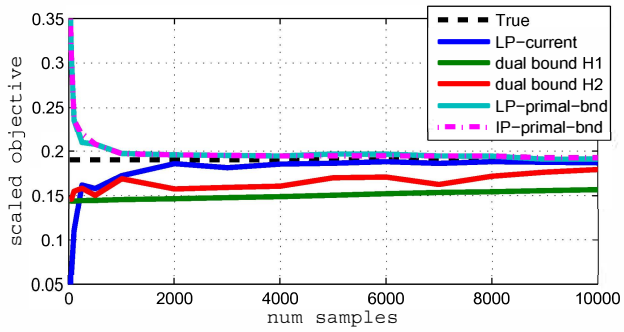

Fig. 1. Illustration of upper and lower bounds on the rule-learning LP and IP objective values for the UCI Adult classification dataset. We obtain tight bounds using only a small fraction of the data.

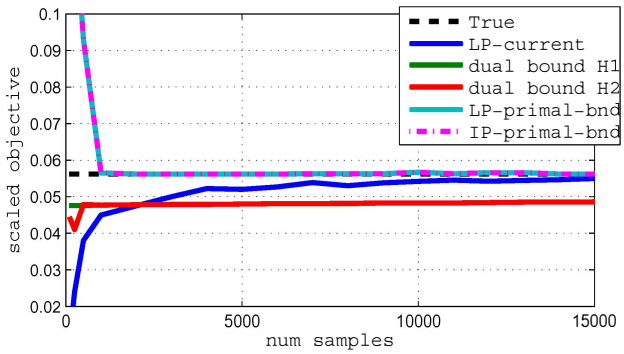

Fig. 2. Illustration of upper and lower bounds on the rule-learning LP and IP objective values for the UCI Census-Income dataset.

errors on the large LP: $\boldsymbol{\xi}^{\bar{m}}_{\mathcal{Z}} = \mathbf{A}_{\mathcal{Z}}\hat{\mathbf{w}}^m$ and

$$\boldsymbol{\xi}^{\bar{m}}_{\mathcal{P}} = \begin{cases} 0 & \text{if } \mathbf{A}_{\mathcal{P}}\hat{\mathbf{w}}^m \geq 1 \\ 1 & \text{otherwise.} \end{cases}$$

This pair $(\hat{\mathbf{w}}^m, \boldsymbol{\xi}^{\bar{m}})$ is feasible for the large LP and the objective value provides an upper bound on $f_{\bar{m}}$. Similarly one can extend an IP solution of the small IP and get an upper bound on $f^*_{\bar{m}}$.

To find a lower bound on $f_{\bar{m}}$ we extend the dual solution of the small LP to give a feasible (but generally sub-optimal) dual solution of the large LP. Recall the dual formulation for the LP in (4), where $\boldsymbol{\mu}$ are the dual variables:

$$\begin{aligned} \frac{1}{m} \max \quad & \sum_{i=1}^{p} \mu_i \\ \text{s.t.} \quad & 0 \leq \mu_i \leq 1,\ i = 1, \ldots, p \\ & \boldsymbol{\mu}^T \mathbf{A}_{\mathcal{P}} \leq \frac{1}{\lambda}\mathbf{1}_n + \mathbf{1}^T \mathbf{A}_{\mathcal{Z}}. \end{aligned} \tag{5}$$

Suppose that $\hat{\boldsymbol{\mu}}^p$ is the optimal dual solution to the small LP. Note that the number of variables in the dual for the large LP increases from p to $\bar{p}$ and the scale factor on the objective changes from $\frac{1}{m}$ to $\frac{1}{\bar{m}}$.

We define a greedy heuristic (H1) to extend $\hat{\boldsymbol{\mu}}^p$ to a feasible dual solution $\bar{\boldsymbol{\mu}}^{\bar{p}}$ of the large LP; note that $\frac{1}{\bar{m}}\sum_{i=1}^{\bar{p}} \bar{\mu}_i$ is a lower bound on the large LP and IP optimal values, i.e., on $f_{\bar{m}}$ and $f^*_{\bar{m}}$. We set $\bar{\mu}_j = \hat{\mu}_j$ for $j = 1, .., p$. We extend the remaining entries $\bar{\mu}_j$ for $j = (p+1), .., \bar{p}$ by setting a subset of its entries to 1 while satisfying $(\bar{\boldsymbol{\mu}}^{\bar{p}})^T \bar{\mathbf{A}}_{\mathcal{P}} \leq \mathbf{1}^T \bar{\mathbf{A}}_{\mathcal{Z}}$ which implies the dual feasibility constraint. In other words the extension of $\bar{\boldsymbol{\mu}}$ corresponds to a subset $\mathcal{R}$ of the row indices $\{p+1, \ldots, \bar{p}\}$ of $\bar{\mathbf{A}}_{\mathcal{P}}$ such that $(\hat{\boldsymbol{\mu}}^p)^T \mathbf{A}_{\mathcal{P}} + \sum_{i \in \mathcal{R}} (\bar{\mathbf{A}}_{\mathcal{P}})_i \leq \mathbf{1}^T \bar{\mathbf{A}}_{\mathcal{Z}}$. We initialize $\mathcal{R}$ to $\emptyset$ and then simply go through the unseen rows of $\bar{\mathbf{A}}_{\mathcal{P}}$ in some fixed order (increasing from $p+1$ to $\bar{p}$), and for a row k, if

$$(\hat{\boldsymbol{\mu}}^p)^T \mathbf{A}_{\mathcal{P}} + \sum_{i \in \mathcal{R}} (\bar{\mathbf{A}}_{\mathcal{P}})_i + (\bar{\mathbf{A}}_{\mathcal{P}})_k \leq \mathbf{1}^T \bar{\mathbf{A}}_{\mathcal{Z}},$$

we set $\mathcal{R}$ to $\mathcal{R} \cup \{k\}$. This first heuristic needs only a single pass through the matrix $\bar{\mathbf{A}}_{\mathcal{P}}$, and is thus very fast.

However, it does not use the optimal solution $\hat{\mathbf{w}}^m$ in any way. Suppose $\hat{\mathbf{w}}^m$ were an optimal solution of the large LP. Then complementary slackness would imply that if $(\bar{\mathbf{A}}_{\mathcal{P}})_i \hat{\mathbf{w}}^m > 1$, then in any optimal dual solution $\boldsymbol{\mu}$, $\mu_i = 0$. Thus, assuming $\hat{\mathbf{w}}^m$ is close to an optimal solution for the large LP, we modify heuristic H1 to obtain heuristic H2, by simply setting $\bar{\mu}_j = 0$ whenever $(\bar{\mathbf{A}}_{\mathcal{P}})_i \hat{\mathbf{w}}^m > 1$, while keeping the remaining steps unchanged.

3.1. Row Sampling: stochastic setting

Now suppose that we operate in an online setting where we can request additional i.i.d. samples, and we would like to declare that we are close to a stationary solution, i.e. that our solution will not change much with additional samples. We describe how to compute expected upper bounds and expected lower bounds on the objective value of the big LP. After receiving m samples we have learned a classifier specified by rule $\hat{\mathbf{w}}^m$. We can compute the expected upper bound on the objective value of a larger LP by drawing a small number $\tilde{m} \ll \bar{m}$ of additional validation samples, extending the primal solution to be feasible as we have done earlier, and evaluating the expected resulting errors. We consider the false positives and mis-detect errors separately: $\sum_{i=1}^{m} \xi_i = \sum_{i \in \mathcal{Z}} \xi_i + \sum_{i \in \mathcal{P}} \xi_i$.

For a fixed $\hat{\mathbf{w}}_m$ the $\boldsymbol{\xi}_{\tilde{\mathcal{P}}}$ errors follow an i.i.d. Bernoulli distribution, so we simply estimate the probability of error $p_e = \frac{1}{\tilde{m}} \sum_{i \in \tilde{\mathcal{P}}} \xi_i$. We can use Agresti-Coull [17] confidence intervals on the sample binomial and its upper bound would correspond to the upper bound on the $\boldsymbol{\xi}_{\mathcal{P}}$ contribution to the objective.

The errors $\boldsymbol{\xi}_{\mathcal{Z}}$ in our model are in general not binary, and can take positive integer values. Since we know $\hat{\mathbf{w}}^m$, the values of $\boldsymbol{\xi}_{\tilde{\mathcal{Z}}}$ are bounded between 0 and $\|\hat{\mathbf{w}}_m\|_0$. Hence we can use the Hoeffding inequality [18] to obtain a confidence interval on the contribution of $\boldsymbol{\xi}_{\tilde{\mathcal{Z}}}$ to the objective. The expectation is $\frac{1}{\tilde{m}} \sum_{i \in \tilde{\mathcal{Z}}} \xi_i$. To obtain the complete expectation of the objective of this feasible solution to the big LP (and its upper confidence bound) we simply add $\frac{1}{m\lambda}\|\hat{\mathbf{w}}^m\|_1$ and these two terms.

To compute the expected lower bound we need to extend the dual solution when we receive additional samples to satisfy:

$$\sum_{i=p+1}^{\bar{p}} \mu_i (\bar{\mathbf{A}}_{\mathcal{P}})_i \leq \mathbf{1}^T \bar{\mathbf{A}}_{\mathcal{Z}} - (\hat{\boldsymbol{\mu}}^p)^T \mathbf{A}_{\mathcal{P}}.$$

In the stochastic setting we do not know $\bar{\mathbf{A}}_{\mathcal{Z}}$ or $\bar{\mathbf{A}}_{\mathcal{P}}$ and can only have access to the expectations. Instead of using the dual extension described in Section 3, we use a simpler extension that is easier to analyze: we set $\mu_i = \min(c, 1)$ for $i = p+1, ..., \bar{p}$, where

$$c = \min_i \frac{(\mathbf{1}^T \bar{\mathbf{A}}_{\mathcal{Z}} - (\hat{\boldsymbol{\mu}}^p)^T \mathbf{A}_{\mathcal{P}})_i}{(\bar{\mathbf{A}}_{\mathcal{P}})_i}.$$

By construction this dual extension is feasible and provides a lower bound on the optimal objective of the big LP: $\frac{1}{\bar{m}} \sum_i \bar{\mu}_i =$

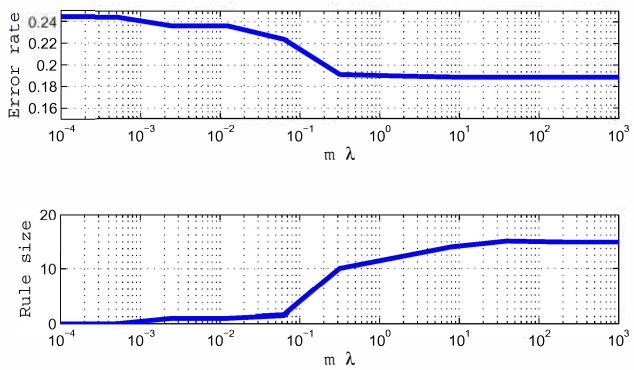

Fig. 3. Solution path as a function of λ for the Adult data-set: trading off classification error vs. size of the rule.

$\frac{1}{\bar{m}}\left(\sum_i \hat{\mu}_i^{\mathcal{P}} + (\bar{m}-m)c\right)$. To estimate the expected value of c and a confidence bound we can use the original sample set $1,...,m$ by splitting it into non-overlapping blocks, and evaluating sample values of c. Confidence bounds on c can be obtained from the sample-mean Chebyshev bound [19]. Note that since this dual extension is not using $\hat{\mathbf{w}}^m$ we can avoid drawing additional samples.

4. EMPIRICAL FINDINGS

In this section, we examine the applicability of the bounds we have developed on two large-scale binary classification datasets from the UCI Machine Learning Repository [20]. We consider the "Adult" dataset with 101 features and 32560 training samples and the "Census Income" dataset with 354 features and 199522 training samples. After converting categorical features into binary indicators, and thresholding continuous features with 10 thresholds we obtain 310 and 812 columns respectively in the $\mathbf{A}$-matrices for the Adult and Census-Income datasets respectively.

In Figure 1 we consider the smaller Adult dataset. We use regularization parameter $\lambda = 1000$. The full training set (our large LP) has $\bar{m} = 32560$ samples, and we plot the various bounds as a function of m: we show the objective value of the full LP (constant dashed line), and of the small LP, the upper bounds on both the LP and IP solutions for the full dataset, and the two dual bounds. We can see that the objective value of the small LP and both the LP and IP upper bounds quickly approach the objective value of the full LP (after about 2000 samples). The dual bounds improve with time, albeit slower than the upper bounds. The second dual extension approach provides a much tighter lower bound.

In Figure 2 we consider the larger Census-Income dataset, again using $\lambda = 1000$. Again we see that the small LP objective values and the upper bounds quickly converge to the LP objective on the full dataset. The dual bounds improve with additional samples, albeit at a slower rate. The second dual extension heuristic does not provide a significant gain over the first heuristic for this example, but they do provide very useful lower bounds. Remarkably, for both UCI examples the LP and IP solutions for the small LP are either the same or very close, allowing quick integral solution via branch and bound. The same holds for the LP and IP upper bounds.

5. COMPUTING THE SOLUTION PATH

The BCS rule-learning framework trades off the sparsity (interpretability) of the rule with respect to its classification accuracy by a regularization parameter λ in (4). A good choice of λ is typically not known a-priori, and it may be of interest to scan through a range of values of λ. Furthermore, for cost-sensitive classification we would like to quantify how the solution changes with varying costs on the false positive and false negative errors. This is accommodated by including a parameter β with $0 \leq \beta \leq 1$ and modifying the objective function in (4):

$$\frac{1}{m}\left(\frac{1}{\lambda}\sum_{j=1}^{n} w_j + \beta\sum_{i\in\mathcal{P}}^{m}\xi_i + (1-\beta)\sum_{i\in\mathcal{Z}}^{m}\xi_i\right) \qquad (6)$$

For both of these problems it is important to have a practical way to quickly scan through a potential range of λ and β. For varying λ, the optimal solution is a piecewise linear function of λ which can in principle be obtained by parametric linear programming. For a simpler practical approach, sensitivity analysis techniques in linear programming can be used. Suppose we obtain an optimal solution to LP for one choice of λ using the simplex algorithm implementation of a modern LP solver such as CPLEX. Changing λ by a small amount corresponds to changing the objective function by a small amount. If the optimal solution to LP does not change, then the LP solver can simply verify optimality by recomputing the dual vector via one linear solve. If the optimal solution changes but the new solution is close to the previous solution, then a small number of pivots in the simplex solver typically obtains the new optimal solution.

In Figure 3 and 4 we evaluate the regularization path as a function of λ and β which allows to decide on the appropriate trade-off between interpretability and classification errors[1], and on the balance of the positive and negative errors. The solution time for the entire parameter grid of 100 values of λ (using CPLEX in Matlab) with $m = 5000$ took 6.19 seconds, while the single slowest λ on the grid took 0.53 seconds, showing the value of warm-starting.

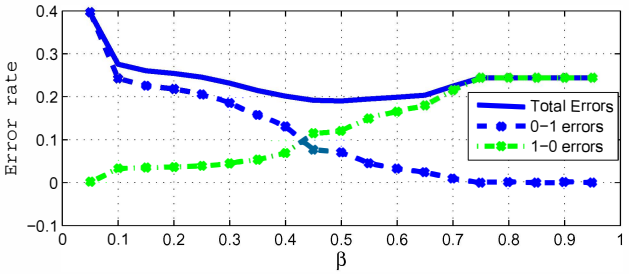

Fig. 4. Cost sensitive classification: number of false positives and false negatives as a function of β.

6. CONCLUSION

We considered learning interpretable classification rules using Boolean compressed sensing. For large-scale classification problems we showed how it is possible to guarantee a near-optimal solution after training the classifier only on a small subset of the available samples. We confirmed the validity of our approach on large scale binary classification problems from the UCI collection.

[1] Adult dataset asks if a person is a high-earner. A single rule suggests the value of education: *education:Some-college True*. Larger rules use many features suggesting which jobs to avoid: *occupation:Handlers-cleaners False and occupation:Other-service False and workclass:Federal-gov False*, ...

7. REFERENCES

[1] 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.

[2] A. Blum, A. Kalai, and J. Langford, "Beating the hold-out: Bounds for k-fold and progressive cross-validation," in *Proc. Conf. Comput. Learn. Theory*, Santa Cruz, CA, jul 1999, pp. 203–208.

[3] F. Provost, D. Jensen, and T. Oates, "Efficient progressive sampling," in *Proceedings of the fifth ACM SIGKDD*, 1999, pp. 23–32.

[4] G. H. John and P. Langley, "Static versus dynamic sampling for data mining." in *KDD*, vol. 96, 1996, pp. 367–370.

[5] O. Maron and A. W. Moore, "Hoeffding races: Accelerating model selection search for classification and function approximation," *Robotics Institute*, p. 263, 1993.

[6] D. M. Malioutov, S. R. Sanghavi, and A. S. Willsky, "Sequential compressed sensing," *IEEE Trans. Special Topics Sig. Proc.*, vol. 4, no. 2, pp. 435–444, apr 2010.

[7] K. Ogawa, Y. Suzuki, and I. Takeuchi, "Safe screening of non-support vectors in pathwise SVM computation," in *Proc. Int. Conf. Mach. Learn.*, Atlanta, GA, jun 2013, pp. 1382–1390.

[8] S. Dash, D. M. Malioutov, and K. R. Varshney, "Screening for learning classification rules via Boolean compressed sensing," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Florence, Italy, may 2014, pp. 3360–3364.

[9] D. Bertsimas, A. Chang, and C. Rudin, "ORC: Ordered rules for classification," Oper. Res. Center, Mass. Inst. Tech., Working Paper OR 386-11, oct 2011.

[10] K. R. Varshney, J. C. Rasmussen, A. Mojsilović, M. Singh, and J. M. DiMicco, "Interactive visual salesforce analytics," in *Proc. Int. Conf. Inf. Syst.*, Orlando, FL, Dec. 2012.

[11] B. Letham, C. Rudin, T. McCormick, and D. Madigan, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," University of Washington, Department of Statistics Technical Report tr608, 2014.

[12] A. C. Gilbert, M. A. Iwen, and M. J. Strauss, "Group testing and sparse signal recovery," in *Asilomar Conf. Signals Syst. Comp. Conf. Record*, Pacific Grove, CA, Oct. 2008, pp. 1059–1063.

[13] G. K. Atia and V. Saligrama, "Boolean compressed sensing and noisy group testing," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1880–1901, mar 2012.

[14] D. Malioutov and M. Malyutov, "Boolean compressed sensing: LP relaxation for group testing," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Kyoto, Japan, mar 2012, pp. 3305–3308.

[15] C. L. Chan, S. Jaggi, V. Saligrama, and S. Agnihotri, "Non-adaptive group testing: Explicit bounds and novel algorithms," in *Proc. IEEE Int. Symp. Inf. Theory*, Cambridge, MA, jul 2012, pp. 1837–1841.

[16] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, mar 2008.

[17] A. Agresti, *Categorical data analysis*. John Wiley and Sons, 2014.

[18] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American statistical association*, vol. 58, no. 301, pp. 13–30, 1963.

[19] J. G. Saw, M. C. Yang, and T. Mo, "Chebyshev inequality with estimated mean and variance," *The American Statistician*, vol. 38, no. 2, pp. 130–132, 1984.

[20] A. Frank and A. Asuncion, "UCI machine learning repository," available at http://archive.ics.uci.edu/ml, 2010.