Support Vector Algorithms for Optimizing the Partial Area Under the ROC Curve

Harikrishna Narasimhan HARIKRISHNA@CSA.IISC.ERNET.IN
Shivani Agarwal SHIVANI@CSA.IISC.ERNET.IN
Department of Computer Science and Automation, Indian Insititute of Science, Bangalore, India.

Abstract

The area under the ROC curve (AUC) is a popular performance measure in machine learning. In an increasing number of applications, however, performance is measured in terms of the *partial* area under the ROC curve between two given false positive rates. In this work, we develop two SVM based methods for directly optimizing this partial AUC; the first method is a general structural SVM approach, while the second method uses an improved SVM formulation that optimizes a tighter convex upper bound on the partial AUC risk. We demonstrate the effectiveness of our methods on several biological data sets.

1. Introduction

The receiver operating characteristic (ROC) curve is an important evaluation tool in machine learning. In particular, the area under the ROC curve (AUC) is used to summarize the performance of a scoring function in binary classification, and is also used as a performance measure in bipartite ranking problems. In an increasing number of applications, however, the performance measure of interest is not the area under the full ROC curve, but instead, the *partial* area under the ROC curve between two specified false positive rates (see Figure 1). For example, in ranking applications where accuracy at the top is critical, one is often interested in the left-most part of the ROC curve; this corresponds to maximizing partial AUC in a false positive range $[0, \beta]$. In biometric screening, where false positives are intolerable, one is again interested in maximizing the partial AUC in a false positive range $[0, \beta]$ for some small β. In the KDD Cup 2008 challenge on breast cancer detection, performance was measured in terms of the partial AUC in a false positive range $[\alpha, \beta]$ deemed clinically relevant (Rao et al., 2008).

Figure 1. Partial AUC in the false positive range $[\alpha, \beta]$.

In this paper, we develop new support vector algorithms for directly optimizing the partial AUC between any two given false positive rates α and β. As a first cut, we develop a structural SVM based method for optimizing partial AUC; building on this, we propose an improved SVM formulation that optimizes a tighter convex upper bound on the partial AUC risk and has better run time guarantees. Both our algorithms make use of a cutting plane solver along the lines of the structural SVM based approach for optimizing the full AUC developed by (Joachims, 2005); one of our key technical contributions here is an efficient algorithm for solving the combinatorial optimization problem needed to find the most violated constraint in the cutting plane solver. We also develop an alternate primal projected subgradient solver, which offers computational savings in certain settings. We demonstrate on a wide variety of bioinformatics tasks that the proposed algorithms do indeed optimize partial AUC and in many cases, perform better than existing baseline techniques. In addition, we develop extensions of our method to learn sparse and group sparse models, often of interest in biological applications.

The paper is organized as follows. We describe our problem setup in Section 2, followed by brief descriptions of our new SVM algorithms in Sections 3 and 4 and highlight some experimental results in Section 5.

Parts of this work will be presented at ICML 2013 (Narasimhan & Agarwal, 2013a) and at KDD 2013 (Narasimhan & Agarwal, 2013b).

Appearing in Proceedings of the 1st Indian Workshop on Machine Learning, IIT Kanpur, India, 2013.

2. Problem Setup

Let X be an instance space. Given a training sample of positive examples $(x_1^+, \ldots, x_m^+) \in X^m$ and negative examples $(x_1^-, \ldots, x_n^-) \in X^n$, the goal is to learn a scoring function $f : X \to \mathbb{R}$ that maximizes the empirical partial AUC (pAUC) in a false positive range $[\alpha, \beta]$, or equivalently minimizes the following risk:

$$\widehat{R}(f) = \sum_{i=1}^{m} \sum_{j=j_\alpha+1}^{j_\beta} \mathbf{1}\big(f(x_i^+) < f(x_{(j)_f}^-)\big), \quad (1)$$

where $j_\alpha = \lceil n\alpha \rceil$, $j_\beta = \lfloor n\beta \rfloor$, and $x_{(j)}^-$ denotes the negative instance ranked in j-th position (among negatives, in descending order of scores) by f; see (Narasimhan & Agarwal, 2013a) for a more general definition. We next outline two new SVM based methods for (approximately) optimizing the above risk.

3. $\text{SVM}_{\text{pAUC}}^{\text{struct}}$: A Structural SVM Approach for Optimizing pAUC

Our first algorithm is a general structural SVM based method along the lines of Joachims' structural SVM based approach for optimizing the full AUC (Joachims, 2005); our method essentially minimizes a (convex) hinge relaxation of Eq. (1), where the resulting quadratic program is solved using a cutting plane method. Each iteration of the cutting plane method involves a combinatorial search over an exponential number of orderings of the positive vs. negative training instances – with each ordering represented as a binary matrix – to find the currently most violated constraint. In the case of the full AUC, this combinatorial optimization problem decomposes neatly into one in which each matrix entry can be chosen independently (Joachims, 2005). Unfortunately, for the partial AUC, such a straightforward decomposition is no longer possible; instead, we formulate an equivalent optimization problem with a restricted search space, which can then be broken down into smaller tractable optimization problems. Following this reformulation, each *row* of the matrix can be optimized separately – and efficiently; the resulting algorithm has the same computational complexity as in the case of Joachims' algorithm for optimizing the usual AUC.

4. $\text{SVM}_{\text{pAUC}}^{\text{tight}}$: SVM Formulation with a Tighter Convex Upper Bound

Building on our first algorithm, we develop a new support vector method, $\text{SVM}_{\text{pAUC}}^{\text{tight}}$, that optimizes a tighter convex upper bound on the partial AUC risk, which leads to both improved accuracy and reduced computational complexity. In particular, by rewriting the partial AUC risk as a maximum of a certain quantity over subsets of negative instances, we derive a new formulation, where a truncated form of the earlier optimization objective is evaluated on each of these subsets, leading to a tighter hinge relaxation on the partial AUC risk. As with $\text{SVM}_{\text{pAUC}}^{\text{struct}}$, the resulting optimization problem can be solved using a cutting plane algorithm, but the new method requires smaller computation time for finding the most-violated constraint and was observed to converge faster. We also develop an alternate primal projected subgradient solver for $\text{SVM}_{\text{pAUC}}^{\text{tight}}$, which offers additional computational savings in certain settings; this in turn allows us to extend our method to learn sparse and group sparse models by incorporating different sparsity inducing regularizers in the projection step of the solver.

5. Experimental Results

We evaluated our methods on a variety of bioinformatics tasks, ranging from protein-protein interaction prediction (PPI) to cancer diagnosis (see Table 1). In most cases, the proposed methods gave improvements in partial AUC over existing methods; between $\text{SVM}_{\text{pAUC}}^{\text{struct}}$ and $\text{SVM}_{\text{pAUC}}^{\text{tight}}$, the latter performed better.

	$\text{SVM}_{\text{pAUC}}^{\text{tight}}[0, 0.1]$	$\text{SVM}_{\text{pAUC}}^{\text{struct}}[0, 0.1]$	SVM_{AUC}
PPI	**52.95**	51.96	39.72
Cheminformatics	**65.30**	65.28	62.78
KDD Cup 2001	69.91	**70.12**	62.23
Leukemia	**30.44**	24.64	28.83

Table 1. Partial AUC in $[0, 0.1]$ for the two proposed support vector methods and the AUC optimizing method due to (Joachims, 2005) on different bioinformatics data sets.

6. Conclusions

We have outlined two support vector algorithms for optimizing the partial AUC, a performance measure of interest in several real-world applications. Experimental results confirm the effectiveness of our methods.

References

Joachims, T. A support vector method for multivariate performance measures. In *ICML*, 2005.

Narasimhan, H. and Agarwal, S. A structural SVM based approach for optimizing partial AUC. In *ICML*, 2013a.

Narasimhan, H. and Agarwal, S. $\text{SVM}_{\text{pAUC}}^{\text{struct}}$: A new support vector method for optimizing partial AUC based on a tight convex upper bound. In *KDD*, 2013b.

Rao, R. B., Yakhnenko, O., and Krishnapuram, B. KDD Cup 2008 and the workshop on mining medical data. *SIGKDD Explor. Newsletter*, 10(2):34–38, 2008.