Carnegie Mellon University
Research Showcase @ CMU

Computer Science Department | School of Computer Science

6-2007

Open Problems in Efficient Semi-Supervised PAC Learning

Avrim Blum
Carnegie Mellon University

Maria-Florina Balcan
Carnegie Mellon University

Follow this and additional works at: http://repository.cmu.edu/compsci

This Article is brought to you for free and open access by the School of Computer Science at Research Showcase @ CMU. It has been accepted for inclusion in Computer Science Department by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.

Open Problems in Efficient Semi-Supervised PAC Learning

Avrim Blum* and Maria-Florina Balcan*
Carnegie Mellon University

1 Introduction

The standard PAC model focuses on learning a class of functions from labeled examples, where the two critical resources are the number of examples needed and running time. In many natural learning problems, however, unlabeled data can be obtained much more cheaply than labeled data. This has motivated the notion of semi-supervised learning, in which algorithms attempt to use this cheap unlabeled data in a way that (hopefully) reduces the number of labeled examples needed for learning [4]. For instance, semi-supervised and transductive SVM [2, 5] and co-training [3] are two examples of semi-supervised learning algorithms. In [1], a semi-supervised PAC model is introduced that provides a common framework for the kinds of assumptions these algorithms make; however, most of the results in [1] deal with sample complexity rather than computational efficiency, or are only computationally efficient under strong assumptions on the underlying distribution. This note poses several questions related to developing computationally efficient algorithms in this semi-supervised PAC model.

2 The model

The high-level idea of the semi-supervised PAC model of [1] is that rather than talking of learning a concept class C, one talks of learning a class C under a compatibility notion χ. Given a hypothesis h and distribution D, $\chi(h, D)$ is a score in $[0, 1]$ indicating how compatible h is with D. For example, if we believe data should be separable by a large margin, then χ would give a low score to separators that slice through dense regions under D and high score to those that do not. Or, if data has two "views" and one believes that either view should be sufficient for classification (as in co-training) then χ can give a low score to hypothesis pairs that disagree on a large probability mass of examples and a high score to those that tend to agree. Formally, in order to ensure that compatibility can be estimated from a finite sample, one requires that (overloading notation) $\chi(h, D) \equiv \mathbf{E}_{x\sim D}[\chi(h, x)]$ where $\chi(h, x) \in [0, 1]$. The quantity $1 - \chi(h, D)$ can be viewed as a notion of unlabeled error rate. For example, if we define $\chi(h, x) = 0$ if x is within distance γ of hyperplane h and $\chi(h, x) = 1$ otherwise, then the unlabeled error rate $1-\chi(h, D)$ is the probability

* Supported in part by National Science Foundation grant CCF-0514922 and a Google Research Grant.

mass within distance γ of h. The analog to the standard PAC "realizable case" assumption that the target function lies in C is an assumption that furthermore the target is perfectly compatible (i.e., it has both zero true error and zero unlabeled error). In such a case, unlabeled data from D can allow one to reduce the space of plausible functions from the set of all functions in C (which are all potential candidates before any unlabeled data is seen) to just those that happen to be highly compatible with the distribution D (once enough unlabeled data has been seen to uniformly estimate compatibilities of all functions in C).

3 The question

For a given class C, compatibility notion χ, and distribution D, define $C_{D,\chi}(\epsilon) = \{h \in C : 1 - \chi(h, D) \leq \epsilon\}$. Under the assumption that the target belongs to C and is fully compatible, then given enough unlabeled data we can in principle reduce our search space from C down to $C_{D,\chi}(\epsilon)$. Thus, we should in principle need at most $O(\frac{1}{\epsilon}(\log |C_{D,\chi}(\epsilon)| + \log \frac{1}{\delta}))$ labeled examples to learn well.[1] Furthermore, if the distribution D is helpful, then $|C_{D,\chi}(\epsilon)|$ may be much smaller than $|C|$. The high-level question is whether for interesting classes C and notions of compatibility χ, one can learn with this many (or polynomial in this many) labeled examples by *efficient* algorithms. If so, we say that such an algorithm is an efficient semi-supervised learning algorithm for the pair (C, χ). We now instantiate this high-level question with a few specific classes and compatibility notions.

3.1 A simple non-open problem

Before presenting open problems, here is a simple example from [1] of a (C, χ) pair for which efficient semi-supervised learning is easy. Let C be the class of monotone disjunctions over $\{0, 1\}^n$. Now, suppose we say an example x is compatible with function h if either all variables set to 1 in x are relevant variables of h or none of them are. This is a very strong notion of "margin": it says, in essence, that every variable is either a positive indicator or a negative indicator, and no example should contain both positive and negative indicators.

In this case efficient semi-supervised learning is easy. Just draw a large set of unlabeled examples and create a graph with n vertices, one for each variable. Put an edge between two vertices if any example has both variables set to 1. Under the compatibility assumption, all variables in the same connected component of this graph must either all be positive indicators or all be negative indicators. So, if we have k components, we only need $O(\frac{1}{\epsilon}[k + \log \frac{1}{\delta}])$ labeled examples to achieve a PAC guarantee. Furthermore, as long as we created the graph using enough unlabeled data we can be confident that $k \leq \lg |C_{D,\chi}(\epsilon)|$. Note that in this context, a "helpful" distribution is one that produces a small number of components.

[1] Or even less depending on the structure of C. For example, we would ideally use an ϵ-cover bound here. Note that we have overloaded "ϵ" for both labeled and unlabeled error bounds for simplicity.

3.2 Specific open problems

Two-sided disjunctions: This is a generalization of the example above where we now allow variables to be positive indicators, negative indicators, or irrelevant. Specifically, define a "two-sided disjunction" h to be a pair of disjunctions (h_+, h_-) where only h_+ is used for classification, but h is compatible with D iff for all examples x, $h_+(x) = -h_-(x)$. That is, D is such that both the positive and negative classes can be described by OR-functions.

Two-sided majority with margins: As a different generalization of the problem from Section 3.1, suppose that again every variable is either a positive or negative indicator, but we relax the margin condition a bit. In particular, say we require that x either contain at least 60% of the positive indicators and at most 40% of the negative indicators (for positive examples) or vice versa (for negative examples).

Co-training with disjunctions: This is the "inverse" of the two-sided disjunction problem. Let C be the class of disjunctions, but an example x is a *pair* of points (x_1, x_2) in $\{0,1\}^n$. Define $h(x) = h(x_1)$ but say that h is compatible with x iff $h(x_1) = h(x_2)$. That is, under our compatibility assumption, each unlabeled example is either a pair of positive examples or a pair of negative examples. Note that D is now a distribution over pairs.

Co-training with linear separators: A generalization of the above problem is the case that h is a linear separator. It is known that the consistency problem is NP-hard [A. Flaxman, personal communication], however efficient algorithms are known for the special case that the elements x_1 and x_2 of the pair are drawn independently given their label [3, 1]. Even if one cannot solve the problem efficiently in general, a natural question is whether one can at least weaken the independence-given-the-label assumption in a nontrivial way and still get an efficient algorithm for this class.

Monetary rewards: $300 for a positive solution to any of the above questions. More generally, it would be interesting to consider other classes and notions of compatibility as well.

References

1. M.-F. Balcan and A. Blum. A PAC-style model for learning from labeled and unlabeled data. In *Proc. 18th Annual Conference on Learning Theory*, 2005.
2. Kristin P. Bennett and Ayhan Demiriz. Semi-supervised support vector machines. In *NIPS*, pages 368–374, 1998.
3. A. Blum and T. M. Mitchell. Combining labeled and unlabeled data with co-training. In *Proc. 11th Annual Conference on Learning Theory*, pages 92–100, 1998.
4. O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.
5. T. Joachims. Transductive inference for text classification using support vector machines. In *Proc. ICML*, pages 200–209, 1999.