Good Learners for Evil Teachers

Ofer Dekel OFERD@MICROSOFT.COM
Microsoft Research, 1 Microsoft Way, Redmond, WA 98052, USA

Ohad Shamir OHADSH@CS.HUJI.AC.IL
The Hebrew University, Jerusalem 91904, Israel

Abstract

We consider a supervised machine learning scenario where labels are provided by a heterogeneous set of teachers, some of which are mediocre, incompetent, or perhaps even malicious. We present an algorithm, built on the SVM framework, that explicitly attempts to cope with low-quality and malicious teachers by decreasing their influence on the learning process. Our algorithm does not receive any prior information on the teachers, nor does it resort to repeated labeling (where each example is labeled by multiple teachers). We provide a theoretical analysis of our algorithm and demonstrate its merits empirically. Finally, we present a second algorithm with promising empirical results but without a formal analysis.

1. Introduction

A supervised machine learning algorithm receives a training set of labeled examples and returns a hypothesis that attempts to accurately predict the labels of new unseen instances. When designing learning algorithms, we typically overlook the data collection process and make the simplistic assumption that the training set is sampled i.i.d. from some fixed distribution. However, real datasets are often non-homogeneous. In particular, training examples may be labeled by various different teachers. Labels provided by different teachers can be of different quality, due to varying degrees of expertise, competence, and dedication among teachers. Moreover, some teachers may deliberately try to manipulate or confuse our learning algorithms by providing incorrect labels.

For example, imagine an algorithm that analyzes the logs generated by a search engine, with the intent of improving future search engine results. A search-engine log records which links were clicked-on by each user. We think of each user as a distinct teacher, and of each click as a label. The click-patterns of most users are informative and helpful, while the click-patterns of other users merely introduce noise. Search-engine optimizers (SEOs) are individuals who try to reverse-engineer the algorithms used to construct a search engine and to manipulate these algorithms in their favor. If a SEO knows that our search engine promotes pages that received many historic clicks, he may masquerade as numerous users and simulate fictitious clicks on the links that he wants to promote. Identifying this form of *click-spam* and attenuating its effect is an essential step in the learning process.

Another example of a multi-teacher scenario involves the use of *crowdsourcing* websites. Websites such as *Galaxy Zoo* (galaxy classification) and *Stardust@home* (interstellar dust particle detection) let members of the public volunteer their services by labeling astronomical images over the Internet. Amazon.com's *Mechanical Turk* is an online system on which any individual can publish a crowdsourcing task and offer a payment for its completion. Typically, a dataset labeling task is broken up into multiple subtasks, and each subtask is completed by a different worker. Occasionally, workers are tempted to cheat by building automated systems, known as *bots*, that appear to solve the tasks but actually provide worthless labels. Although these bots do not directly try to harm our learning algorithm, they do try to trick us into believing that their labels are genuine. This intentional deceitful input can be as detrimental to our learning algorithm as deliberate malice. Weeding out the bots from among the human teachers is an important and difficult feat.

To our knowledge, the existing literature on multi-teacher learning discusses two main approaches: using prior information and repeated labeling. For example, the work in (Blitzer et al., 2007; Crammer et al., 2008) assumes that labeled examples are obtained from multiple heterogeneous sources, and that we have explicit prior knowledge on the relationships between these sources. This approach is in-

Appearing in *Proceedings of the 26^{th} International Conference on Machine Learning*, Montreal, Canada, 2009. Copyright 2009 by the author(s)/owner(s).

adequate for the type of problems we are concerned with, since no such prior knowledge is available to us. Additionally, an adversarial teacher will behave in a way that contradicts any prior information we may rely on. Repeated labeling (see (Smyth et al., 1994; Sheng et al., 2008) and references therein) is the practice of having each example labeled by multiple teachers, and then aggregating these labels in a way that cleans noise and identifies bad teachers. Repeated labeling is a powerful and successful technique when it can be applied. However, we often have no control over the assignment of examples to teachers (as in the search engine example). Additionally, even when we do have control over the assignments, repeated labeling is wasteful and ultimately decreases the size of our training set. Yet another related approach is to design machine learning algorithms that withstand specific types of label-noise, either on the training set (Kearns, 1998) or on the test set (Teo et al., 2007; Dekel & Shamir, 2008). These approaches do not make use of teacher identities, and do not assume any heterogeneity in the data. We also note the related work in (Dekel et al., 2008), which addresses the multi-teacher learning problem from a mechanism design perspective, and incentivises teachers to be good.

In this paper, we address the problem of learning from heterogeneous, possibly malicious, teachers *without prior knowledge* on the teachers and *without repeated labeling*. For concreteness, we focus on the classic learning problem of binary classification. We present a new algorithm, based on the well-known support vector machine (SVM) framework, that explicitly attempts to identify low-quality and malicious teachers and to decrease their influence on the learning process. We exploit the fact that SVMs, like many other machine learning algorithms, explicitly reveal how important each training example is to the learning process. SVMs indicate which training examples are support vectors and which are not, and non-support vectors can be removed from the training set without changing the learned classifier. In the multi-teacher setting, we can measure the influence of each teacher by the cumulative effect of the examples he controls. Intuitively, if examples are assigned to teachers randomly and if all teachers are alike, we expect any two teachers to have a similar influence on the algorithm output. Specifically, for support vector machines, we expect each teacher to contribute roughly the same number of support vectors. Our algorithm essentially turns this observation into a constraint, namely, we require that all of the teachers have a similar contribution to the learned hypothesis. In our analysis, we show that this constraint is likely to affect only low-quality and malicious teachers.

2. Setting and Notation

First, we review the typical setting of an SVM learning problem. Assume that each example is an instance-label pair $(\mathbf{x}, y)$, where $\mathbf{x}$ is a vector in $\mathcal{X} \subseteq \mathbb{R}^n$ and y takes values in $\{-1, +1\}$. Additionally, define a feature mapping ϕ, which maps instances from $\mathcal{X}$ to a reproducing kernel Hilbert space $\mathcal{H}$ (Shawe-Taylor & Cristianini, 2000). Our classifier is composed of a vector $\mathbf{w} \in \mathcal{H}$ and a bias term $b \in \mathbb{R}$. The *margin* of an instance $\mathbf{x}$ is defined as $\langle \phi(\mathbf{x}), \mathbf{w} \rangle + b$ and the predicted label for $\mathbf{x}$ is simply the sign of the margin. To simplify our presentation, assume that ϕ is the identity mapping and that $\mathcal{H} = \mathcal{X}$, which allows us to drop ϕ altogether. Additionally, we focus on unbiased classifiers, where $b = 0$. All of our results easily extend to the general setting.

The standard statistical learning paradigm assumes that a training set S is sampled i.i.d. from an unknown distribution $\mathcal{D}$ over the space of examples, $\mathcal{X} \times \{-1, +1\}$. The goal is to use S to find a classifier $\mathbf{w}$ such that $\Pr_{(\mathbf{x},y)\sim\mathcal{D}}\big(\text{sign}(\langle \mathbf{x}, \mathbf{w} \rangle \neq y)\big)$ is small. While our goal in this paper remains the same, we modify the training set generation process as follows: First, a set of m *unlabeled* instances, $\{\mathbf{x}_i\}_{i=1}^m$, is drawn i.i.d. from the marginal distribution of $\mathcal{D}$ on $\mathcal{X}$. This set is then randomly split into k disjoint subsets, and each subset is assigned to a different teacher. Each teacher labels his examples, resulting in a labeled training set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$. At this stage, we leave the exact random splitting mechanism unspecified.

For the sake of our theoretical analysis, we assume that each teacher is either *good* or *evil*. This is a harsh simplification of the real-world, but it is one that enables us to derive a rigorous theory, which inspires the design of our algorithms. An evil teacher may label his instances in any arbitrary, possibly malicious, manner. Evil teachers are even allowed to collude amongst themselves. The only assumption we make regarding the evil teachers is that they do not know the identity of the good teachers' examples. This happens, for instance, if good teachers never reveal their instances to evil teachers, or if teachers are simply isolated from each other. On the other hand, a good teacher labels each of its instances $\mathbf{x}$ by sampling a label from $p_{\mathcal{D}}(y|\mathbf{x})$, the marginal distribution over labels conditioned on $\mathbf{x}$. In other words, examples controlled by good teachers are essentially sampled directly from $\mathcal{D}$. In fact, this assumption is not really necessary, but the important condition that must be met is that when all of the teachers are good, a low error-rate classifier (in terms of $\mathcal{D}$) can be learned. To avoid making our problem trivially impossible, we assume that a majority of the data comes from good teachers.

Turning to more technical notation, we abbreviate the sets $\{1, \ldots, m\}$ and $\{1, \ldots, k\}$ by $[m]$ and $[k]$ respectively.

Also, we define the hinge function $[a]_+ \equiv \max\{a, 0\}$. We assume that the support of the marginal distribution of $\mathcal{D}$ on $\mathcal{X}$ is bounded in a ball of radius R around the origin. Teachers are recognized by an index $t \in [k]$, the set of evil teachers is denoted by $T^e \subset [k]$, and the set of good teachers is denoted by $T^g \subset [k]$. For any instance i, we let $t(i) \in [k]$ denote the teacher that labeled instance i. Finally, we let S_t denote the set of examples labeled by teacher t, and we abbreviate $S^g = \cup_{t \in T^g} S_t$ and $S^e = \cup_{t \in T^e} S_t$.

3. Facing Evil Teachers

We begin the derivation of our algorithm by recalling the plain vanilla 2-norm soft-margin SVM formulation. Given a training set $S = \{\mathbf{x}_i, y_i\}_{i=1}^m$ and a regularization parameter $\lambda > 0$, define

$$F(\mathbf{w}|S, \lambda) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m}[1 - y_i\langle \mathbf{x}_i, \mathbf{w}\rangle]_+ \ . \qquad (1)$$

The SVM classifier is the minimizer of $F(\mathbf{w}|S, \lambda)$. As discussed in the introduction, the philosophy behind our approach is to prevent any single teacher from disproportionally influencing the learned classifier. We find it convenient to enforce this constraint in the *dual* formulation of the SVM optimization problem,

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m_+} \ \sum_{i=1}^{m} \alpha_i - \frac{1}{2\lambda}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \qquad (2)$$
$$\text{s.t.} \ \ \forall i \in [m] \ \ 0 \le \alpha_i \le \tfrac{1}{m} \ .$$

We also know additional useful facts about the SVM optimization problem. First, the primal and dual variables are related by the equation $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$. Second, it holds that $\alpha_i > 0$ (namely, example i is a *support vector*), only if $y_i\langle \mathbf{x}_i, \mathbf{w}\rangle \le 1$, and that $\alpha_i = 1/m$ if $y_i\langle \mathbf{x}_i, \mathbf{w}\rangle < 1$. These results are thoroughly discussed in (Shawe-Taylor & Cristianini, 2000) and elsewhere.

To motivate our next step, imagine a situation where all but one of the teachers are good. Now assume that despite the presence of the evil teacher, we manage to find a good classifier $\mathbf{w}$ with respect to $\mathcal{D}$. We expect this classifier to disagree with many of the labels provided by the evil teacher. As a result, many of the examples controlled by the evil teacher will have a margin less than 1 ($y\langle \mathbf{x}, \mathbf{w}\rangle < 1$). Using the facts stated above, we expect the average value of the dual variables associated with the evil teacher, $\frac{1}{|S^e|}\sum_{i \in S^e} \alpha_i$, to be unusually high compared to the overall average dual variable, $\frac{1}{m}\sum_i \alpha_i$. On the other hand, we intuitively expect a good teacher to have an average dual variable value close to $\frac{1}{m}\sum_i \alpha_i$.

This motivates us to add the following constraint to the dual optimization problem:

$$\forall t \in [k] \ \ \frac{1}{|S_t|}\sum_{i \in S_t} \alpha_i \ \le \ \frac{1}{m}\sum_{i=1}^{m} \alpha_i \ + \ \frac{\epsilon}{m\sqrt{|S_t|}}, \qquad (3)$$

where $\epsilon > 0$ is a parameter. In the scenario described above, this constraint is likely to affect only the evil teacher, and to reduce his influence on $\mathbf{w}$. The form of the slack term, $\epsilon/(m\sqrt{|S_t|})$, comes from large deviation considerations: If instances are assigned randomly to teachers and the sample size increases, we expect the random variable $\frac{1}{|S_t|}\sum_{i \in S_t} \alpha_i$ to be concentrated about its expected value of $\frac{1}{|S^g|}\sum_{i \in S^g} \alpha_i$. If most of the examples are controlled by good teachers, then $\frac{1}{|S^g|}\sum_{i \in S^g} \alpha_i$ and $\frac{1}{m}\sum_{i=1}^{m} \alpha_i$ will be close. These informal statements are made more precisely in our theoretical analysis in Sec. 4.

Adding the constraints in Eq. (3) to the optimization problem in Eq. (2), and using standard tools from convex analysis to convert the problem back into its primal form, we obtain:

$$\min_{\mathbf{w} \in \mathbb{R}^n, \boldsymbol{\xi} \in \mathbb{R}^m_+, \boldsymbol{\nu} \in \mathbb{R}^k_+} \ \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \xi_i + \sum_{t=1}^{k} \frac{\epsilon \nu_t}{\sqrt{|S_t|}}$$
$$\text{s.t.} \ \ \forall i \in [m] \ \ y_i\langle \mathbf{x}_i, \mathbf{w}\rangle \ge 1 - \left(\tfrac{m\nu_{t(i)}}{|S_{t(i)}|} - \textstyle\sum_{t=1}^{k} \nu_t\right) - \xi_i$$

This problem is convex, and there exist various methods of solving it, either in the primal or in the dual formulation. Our proposed learning algorithm calculates the solution to this optimization problem and outputs the resulting classifier $\mathbf{w}$.

4. Theoretical Analysis

Our new optimization problem can be written more compactly as the minimization over $\mathbf{w}$ of

$$G(\mathbf{w}|S, \lambda) = \min_{\boldsymbol{\nu} \in \mathbb{R}^m_+} \ \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{t=1}^{k} \frac{\epsilon \nu_t}{\sqrt{|S_t|}} \qquad (4)$$
$$+ \frac{1}{m}\sum_{i=1}^{m}\left[1 - \left(\tfrac{m\nu_{t(i)}}{|S_{t(i)}|} - \textstyle\sum_{t=1}^{k} \nu_t\right) - y_i\langle \mathbf{x}_i, \mathbf{w}\rangle\right]_+$$

The basic idea behind our analysis is that instead of a hinge loss defined with a fixed required-margin of 1, as in Eq. (1), the required-margin in Eq. (4) is teacher-dependent. This flexibility allows us to trade-off the margin requirement for different teachers. When $\nu_t = 0$ for all t, our formulation reduces to the standard SVM formulation. However, setting ν_t to a high value for each evil teacher and to zero for each good teacher decreases the margin requirement for evil teachers and compensates by requiring a higher margin for good teachers. The end result is that if we can find a high-margin classifier for the good teachers' instances, then Eq. (4) automatically becomes less sensitive to the loss obtained on the evil teachers' instances, and is thus less influenced by them.

There are many ways to provide this intuitive observation with a more solid theoretical grounding. Here, we take an optimization-based approach. Suppose we somehow knew beforehand the identity of S^g, the subset of examples labeled by good teachers. By assumption, S^g is essentially sampled i.i.d. from $\mathcal{D}$. Therefore, with high probability, the standard SVM solution $\mathbf{w}^\star = \arg\min_{\mathbf{w}} F(\mathbf{w}|S^g, \lambda)$ has a low error-rate over the entire distribution (Shawe-Taylor & Cristianini, 2000). However, we do not know S^g in advance and we cannot calculate $\mathbf{w}^\star$. We therefore pose the question: How large can $F(\hat{\mathbf{w}}|S^g, \lambda) - F(\mathbf{w}^\star|S^g, \lambda)$ be when $\hat{\mathbf{w}} = \arg\min G(\mathbf{w}|S, \lambda)$? In other words, how sub-optimal is the classifier trained by our algorithm on S compared to a classifier trained by a standard SVM on S^g. We also ask the same question when $\hat{\mathbf{w}} = \arg\min F(\mathbf{w}|S, \lambda)$, namely, how sub-optimal is the SVM classifier trained on all of S?

Although this is not the focus of our paper, it is straightforward to derive a generalization analysis for our learned classifier based on the sub-optimality bound described above. The idea is to use a uniform convergence argument to relate both $F(\hat{\mathbf{w}}|S^g, \lambda)$ and $F(\mathbf{w}^\star|S^g, \lambda)$ to their expectations with respect to the underlying distribution. Thus, the sub-optimality of our learned classifier with respect to S^g translates to a similar sub-optimality with respect to the underlying distribution, and generalization guarantees for $\mathbf{w}^\star$ can be converted into similar guarantees for $\hat{\mathbf{w}}$. We refer the interested reader to standard references such as (Shawe-Taylor & Cristianini, 2000).

For technical reasons, we actually prove bounds on $F(\hat{\mathbf{w}}|S^g, \frac{m}{|S^g|}\lambda) - F(\mathbf{w}^\star|S^g, \frac{m}{|S^g|}\lambda)$ rather than on $F(\hat{\mathbf{w}}|S^g, \lambda) - F(\mathbf{w}^\star|S^g, \lambda)$. Namely, we compare the SVM objective value using a slightly different regularization parameter. Since $m/|S^g|$ is assumed to be small, this does not materially affect the conclusions.

We begin with a simple theorem that bounds the effect evil teachers may have on a standard SVM. Proofs are given at the end of the section.

Theorem 1. *Let a training set S of size m be fixed. Let $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} F(\mathbf{w}|S, \lambda)$ and let $\mathbf{w}^\star = \arg\min_{\mathbf{w}} F(\mathbf{w}|S^g, \frac{m}{|S^g|}\lambda)$. Then $F(\hat{\mathbf{w}}|S^g, \frac{m}{|S^g|}\lambda) - F(\mathbf{w}^\star|S^g, \frac{m}{|S^g|}\lambda)$ is at most*

$$\frac{|S^e|}{|S^g|}(1 + R\|\mathbf{w}^\star\|) \ .$$

Similarly, the next theorem bounds the effect evil teachers may have on our algorithm. Compared to Thm. 1, the next theorem has an additional non-trivial condition. We see shortly when this condition holds.

Theorem 2. *Let a training set S of size m be fixed. Let $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} G(\mathbf{w}|S, \lambda)$ (with optimal auxiliary parameters $\hat{\nu}_1, \ldots, \hat{\nu}_k$), and let $\mathbf{w}^\star = \arg\min_{\mathbf{w}} F(\mathbf{w}|S^g, \frac{m}{|S^g|}\lambda)$. Furthermore, assume that for any $t \in T^g$, $\hat{\nu}_t \leq |S_t| \sum_t \hat{\nu}_t / m$. Then $F(\hat{\mathbf{w}}|S^g, \frac{m}{|S^g|}\lambda) - F(\mathbf{w}^\star|S^g, \frac{m}{|S^g|}\lambda)$ is at most*

$$\frac{|S^e|}{|S^g|}(1 + R\|\mathbf{w}^\star\|)\left(\frac{\#V(\mathbf{w}^\star)}{|S^g|} + \frac{\epsilon m}{|S^g|}\frac{\sum_{t \in T^e}\sqrt{|S_t|}}{|S^e|}\right),$$

where $\#V(\mathbf{w}^\star)$ is the number of instances $(\mathbf{x}, y) \in S^g$ such that $y\langle \mathbf{x}, \mathbf{w}^\star \rangle \leq 1 + (1 + R\|\mathbf{w}^\star\|)|S^e|/|S^g|$.

Comparing the two bounds, we see that they both include the term $(1 + R\|\mathbf{w}^\star\|)|S^e|/|S^g|$. In Thm. 2, this term is multiplied by an additional expression

$$\frac{\#V(\mathbf{w}^\star)}{|S^g|} + \frac{\epsilon m}{|S^g|}\frac{\sum_{t \in T^e}\sqrt{|S_t|}}{|S^e|} \ . \tag{5}$$

We now explain why this expression can be much smaller than 1, leading to a tighter bound in Thm. 2 compared to Thm. 1. Eq. (5) is the sum of two terms: the second term is generally $O(\epsilon/\sqrt{m})$, assuming that the set of teachers remains fixed, and that the fraction of examples controlled by each teacher remains roughly constant as the training set grows. The first term is the fraction of examples controlled by good teachers that attain a high margin with respect to our learned classifier. This definition matches the intuitive explanation given earlier about how large margins over good teachers' instances can reduce our sensitivity to the evil teachers' instances. In a certain sense, if the original data is easy to classify (in terms of having large margins), it is easy to identify teachers who are misbehaving. These observations should be taken with a grain of salt, since we are comparing theoretical upper-bounds. However, we believe that our analysis supports our algorithmic design choices and that it complements the empirical study presented later on.

To complete the analysis, it remains to justify the technical condition in Thm. 2, namely that that for all $t \in T^g$, $\hat{\nu}_t \leq |S_t| \sum_t \hat{\nu}_t / m$, where $\hat{\nu}_1, \ldots, \hat{\nu}_k$ are the optimal parameters with respect to our learned classifier $\hat{\mathbf{w}}$. To understand this more clearly, consider the important special case where $|S_t| = m/k$ for all t. The condition now reduces to $\hat{\nu}_t \leq \sum_t \hat{\nu}_t / k$. Namely, for any good teacher t, $\hat{\nu}_t$ is at most the average value of $\hat{\nu}_t$ over all the teachers. This is intuitively plausible, since we expect $\hat{\nu}_t$ to be large for the evil teachers and small for the good teachers. Below, we prove a stronger assertion, provided ϵ is not too small.

Proposition 1. *Using the notation of Thm. 2, assume that the evil teachers do not have access to S^g, that* *instances are split randomly between the teachers (with $|S_1|, \ldots, |S_k|$ being fixed in advance), and that $\epsilon > |S^e|\sqrt{|S_t|}/m$ for any good teacher $t \in T^g$. Then with probability of at least*

$$1 - \sum_{t \in T^g} \exp\left(-2|S_t|\left(\frac{\epsilon}{\sqrt{|S_t|}} - \frac{|S^e|}{m}\right)^2\right),$$

over the random assignment of instances to teachers, we have that $\hat{\nu}_t = 0$ for all $t \in T^g$.

For example, say that $|S_t|$ is the same for all t, a quarter of the teachers are evil, and $m = 1000, k = 40$. Then the bound in Proposition 1 is greater than 0.93 for a very reasonable $\epsilon = 3$. The bound in Proposition 1 depends somewhat on the exact mechanism used to assign instances to teachers. However, we note that a somewhat different expression can be obtained if we choose a teacher for each instance uniformly at random, independently and without fixing $|S_1|, \ldots, |S_k|$ in advance. In both cases, the bottom line remains the same.

We conclude this section with proofs of the results stated above.

Proof of Thm. 1. By the definition of $F(\mathbf{w}|S,\lambda)$ in Eq. (1), we have for any $\mathbf{w}$ that

$$\begin{aligned}
&\frac{m}{S^g}F(\mathbf{w}|S,\lambda) - F\Big(\mathbf{w}|S^g, \frac{m}{|S^g|}\lambda\Big) \qquad (6)\\
&= \frac{1}{|S^g|}\sum_{(\mathbf{x},y)\in S^e}[1 - y\langle \mathbf{x}, \mathbf{w}\rangle]_+\\
&\le \frac{1}{|S^g|}\sum_{(\mathbf{x},y)\in S^e}[1 + \|\mathbf{x}\|\|\mathbf{w}\|]_+ \le \frac{|S^e|}{|S^g|}(1 + R\|\mathbf{w}\|).
\end{aligned}$$

In a similar manner, for any $\mathbf{w}$,

$$\begin{aligned}
&F\Big(\mathbf{w}|S^g, \frac{m}{|S^g|}\lambda\Big) - \frac{m}{|S^g|}F(\mathbf{w}|S,\lambda)\\
&= -\frac{1}{|S^g|}\sum_{(\mathbf{x},y)\in S^e}[1 - y\langle \mathbf{x}, \mathbf{w}\rangle]_+ \le 0. \qquad (7)
\end{aligned}$$

Finally, by the definition of $\hat{\mathbf{w}}$, $F(\hat{\mathbf{w}}|S,\lambda) \le F(\mathbf{w}^\star|S,\lambda)$. Chaining this with Eq. (6) (for $\mathbf{w} = \mathbf{w}^\star$) and Eq. (7) (for $\mathbf{w} = \hat{\mathbf{w}}$), the theorem follows. □

Proof of Thm. 2. The proof has a similar structure to the proof of Thm. 1, but is more involved. The first part of the proof consists of showing that for any $\mathbf{w}$,

$$\begin{aligned}
&\frac{m}{|S^g|}G(\mathbf{w}|S,\lambda) - F\Big(\mathbf{w}|S^g, \frac{m}{|S^g|}\lambda\Big) \qquad (8)\\
&\le \frac{|S^e|}{|S^g|}(1 + R\|\mathbf{w}\|)\left(\frac{\#V(\mathbf{w})}{|S^g|} + \frac{\epsilon m}{|S^g|}\frac{\sum_{t\in T^e}\sqrt{|S_t|}}{|S^e|}\right).
\end{aligned}$$

Recall that $G(\mathbf{w}|S,\lambda)$ (defined in Eq. (4)) for any $\mathbf{w}$ is in fact a minimum over the non-negative variables $\nu_1, \ldots, \nu_k$. Therefore, we can upper bound it for any given $\mathbf{w}$ by fixing $\nu_t = 0$ for all $t \in T^g$, and

$$\nu_t = \frac{|S_t|}{|S^g|}(1 + R\|\mathbf{w}\|)$$

for all $t \in T^e$. Note that for this choice, $\sum_t \nu_t = |S^e|(1 + R\|\mathbf{w}\|)/(|S^g|)$. By definition,

$$\begin{aligned}
\frac{m}{|S^g|}G(\mathbf{w}|S,\lambda) \;&\le\; \frac{m\lambda}{2|S^g|}\|\mathbf{w}\|^2 + \frac{\epsilon m}{|S^g|}\sum_{t=1}^{k}\frac{\nu_t}{\sqrt{|S_t|}} \qquad (9)\\
&+ \frac{1}{|S^g|}\sum_{(\mathbf{x},y)\in S^g}\left[1 + \sum_{t=1}^{k}\nu_t - y_i\langle \mathbf{x}, \mathbf{w}\rangle\right]_+ \qquad (10)\\
&+ \frac{1}{|S^g|}\sum_{(\mathbf{x},y)\in S^e}\left[1 - \frac{m\nu_{t(i)}}{|S_{t(i)}|} + \sum_{t=1}^{k}\nu_t - y\langle \mathbf{x}_{\mathbf{w}}\rangle\right]_+ \qquad (11)
\end{aligned}$$

Line (10) can be upper bounded by

$$F\Big(\mathbf{w}|S^g, \frac{m}{|S^g|}\lambda\Big) - \frac{m\lambda}{2|S^g|}\|\mathbf{w}\|^2 + \frac{\#V(\mathbf{w})}{|S^g|}\sum_{t=1}^{k}\nu_t \;. \qquad (12)$$

Leaving this aside for a minute, it is easy to verify that with our choice of $\nu_1, \ldots, \nu_k$, it holds for all $t \in T^e$ that $\frac{m}{|S_t|}\nu_t - \sum_{t=1}^{k}\nu_t = 1 + R\|\mathbf{w}\|$. This implies that line (11) can be upper bounded by:

$$\begin{aligned}
&\frac{1}{m}\sum_{(\mathbf{x},y)\in S^e}[1 - (1 + R\|\mathbf{w}\|) - y\langle \mathbf{x}, \mathbf{w}\rangle]_+\\
&\le \frac{1}{m}\sum_{(\mathbf{x},y)\in S^e}[-R\|\mathbf{w}\| + R\|\mathbf{w}\|]_+ = 0 \;. \qquad (13)
\end{aligned}$$

Substituting Eq. (12), Eq. (13) and our choice of $\nu_1, \ldots, \nu_k$ into the decomposition of $G(\mathbf{w}|S,\lambda)$ in Eq. (9), we get Eq. (8).

Now, using the assumptions in the theorem statement, we have that

$$\begin{aligned}
&\frac{m}{|S^g|}G(\hat{\mathbf{w}}|S,\lambda) \ge \frac{m\lambda}{2|S^g|}\|\hat{\mathbf{w}}\|^2\\
&\quad + \frac{1}{|S^g|}\sum_{(\mathbf{x},y)\in S^g}\left[1 - \frac{m\hat{\nu}_{t(i)}}{|S_{t(i)}|} + \sum_{t=1}^{k}\hat{\nu}_t - y\langle \mathbf{x}, \mathbf{w}\rangle\right]_+\\
&\ge \frac{m\lambda}{2|S^g|}\|\hat{\mathbf{w}}\|^2 + \frac{1}{|S^g|}\sum_{(\mathbf{x},y)\in S^g}[1 - y\langle \mathbf{x}, \mathbf{w}\rangle]_+\\
&= F\big(\hat{\mathbf{w}}|S^g, \frac{m}{|S^g|}\lambda\big). \qquad (14)
\end{aligned}$$

Also, since $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} G(\mathbf{w}|S,\lambda)$, we have that $G(\hat{\mathbf{w}}|S,\lambda) \le G(\mathbf{w}^\star|S,\lambda)$. Chaining this with Eq. (14), and Eq. (8) (for $\mathbf{w} = \mathbf{w}^\star$), the theorem follows. □

Proof of Proposition 1. We begin by recalling that $\hat{\mathbf{w}}$ is the global minimum of $G(\mathbf{w}|S,\lambda)$ (Eq. (4)). Also let $\hat{\boldsymbol{\nu}}$ be the optimal value of the auxiliary vector $\boldsymbol{\nu}$ in Eq. (4) and let $\hat{\boldsymbol{\alpha}}$ be the corresponding optimizer of the dual problem. Using the KKT optimality conditions, a sufficient condition for $\hat{\nu}_t = 0$ is that the corresponding inequality constraint in the dual problem is *strictly* satisfied. Thus, it suffices to show that $\hat{\boldsymbol{\alpha}}$ satisfies

$$\frac{1}{|S_t|}\sum_{i\in S_t}\hat{\alpha}_i < \frac{1}{m}\sum_{i=1}^{m}\hat{\alpha}_i + \frac{\epsilon}{m\sqrt{|S_t|}} \tag{15}$$

for all $t \in T^g$. But since $\hat{\alpha}_i \in [0, 1/m]$, it is not hard to see that

$$\frac{1}{m}\sum_{i=1}^{m}\hat{\alpha}_i \ge \frac{1}{m}\sum_{i\in S^g}\hat{\alpha}_i \ge \frac{1}{|S^g|}\sum_{i\in S^g}\hat{\alpha}_i - \frac{|S^e|}{m^2}.$$

Therefore, for Eq. (15) to hold, it is sufficient to show that for any $t \in T^g$,

$$\frac{1}{|S_t|}\sum_{i\in S_t}\hat{\alpha}_i < \frac{1}{|S^g|}\sum_{i\in S^g}\hat{\alpha}_i + \frac{\epsilon}{m\sqrt{|S_t|}} - \frac{|S^e|}{m^2}. \tag{16}$$

S^g is labeled by good teachers, all of whom draw labels according to $p_{\mathcal{D}}(y|\mathbf{x})$, and S^g is unknown to the evil teachers. Therefore, the labeling and learning process is statistically equivalent to the following: First split S into S^e and S^g, distribute S^e to the evil teachers and have them generate labels, draw labels for S^g according to $p_{\mathcal{D}}(y|\mathbf{x})$ (hence fixing the optimal $\hat{\alpha}_1, \ldots, \hat{\alpha}_m$) and *only then* assign S^g to the different teachers in T^g. As a result, we can think of $\sum_{i\in S_t}\hat{\alpha}_i/|S_t|$ in Eq. (16) simply as the average of a random subset of α's from $\{\hat{\alpha}_i\}_{i\in S^g}$. The condition in Eq. (16) is then simply the event (over splitting the α's) that for each good teacher, the average of its α's is not significantly larger than the average of all α's. Since the α's were split at random, we can apply Hoeffding's bound plus a union bound to get that with probability at least

$$1 - \sum_{t\in T^g}\exp\left(-2|S_t|\left(\frac{\epsilon}{\sqrt{|S_t|}} - \frac{|S^e|}{m}\right)^2\right),$$

conditioned on S^g, Eq. (16) holds for all $t \in T^g$. Since the bound holds for any S^g, we can remove the conditioning to get a bound on the unconditional probability of Eq. (16) holding for all $t \in T^g$. □

5. Experiments

We empirically evaluated our new algorithm with a set of text categorization experiments using *Reuters Corpus Vol. 1* (RCV1) (Lewis et al., 2004), a collection of $800K$ news articles collected by Reuters. A typical article in the corpus contains around 240 words, and the entire corpus contains over half a million distinct tokens (not including numbers and dates). Each article in the corpus is associated with one or more *high-level categories*, which are: Corporate/Industrial (CCAT), Economics (ECAT), Government/Social (GCAT), and Markets (MCAT). We represented each article in the corpus by a vector of TF-IDF values, and considered the 6 binary classification problems of distinguishing between each pair of two high-level categories. Specifically, for high-level categories A and B, we considered the problem of distinguishing articles of category A from articles of category B, while ignoring articles associated with both A and B, or with neither A nor B. Each of these 6 problems has different characteristics, due to the non-uniform category sizes, the varying degree of category similarity, and the varying degree of homogeneity within each category.

For each binary problem, we took 40 random splits of the corpus into equally sized training and test sets. On each split, we trained a standard well-tuned linear SVM classifier (Shalev-Shwartz et al., 2007) on the training set, and evaluated the resulting classifier on the test set. This test error-rate represents the performance of SVM when all of the teachers are good, and serves as a baseline for measuring the effect of label noise. Next, for each train/test split, we randomly assigned each training example to one of 100 different teachers. For each k in the set $\{5, 10, \ldots, 40\}$, we selected k of the 100 teachers, designated them as malicious teachers, and flipped all of the labels under their control. It is likely that there exists a more sophisticated and harmful way of simulating a malicious teacher, but we decided to choose the simplest and most obvious candidate for the job. No manipulation was applied to any of the test sets. The 40 different train/test splits and the 8 different choices of k led to a total of 320 different noisy variations of each of our 6 binary problems.

For each of the noisy variations described above, we trained a classifier using standard linear SVM and using our algorithm (with $\epsilon = 1$), and we evaluated both classifiers on the test data. We then compared the results using the following metric: let e_1 be the test error-rate attained by the SVM that was trained on noise-free training data, let e_2 be the test error-rate attained by SVM with noisy training data, and let e_3 be the test error-rate attained by our algorithm with noisy training data. Define the *excess-error* sustained by SVM as $e_2 - e_1$ and the excess-error sustained by our algorithm as $e_3 - e_1$. Finally, define the *excess-error ratio* of the two algorithms to be $(e_3 - e_1)/(e_2 - e_1)$. This number compares the resistance of the two algorithms to the evil teachers. Specifically, if this ratio is less than 1 then our algorithm outperforms SVM. The main advantage of reporting our results in this way is that it allows us to fairly compare our algorithm to a standard SVM across a

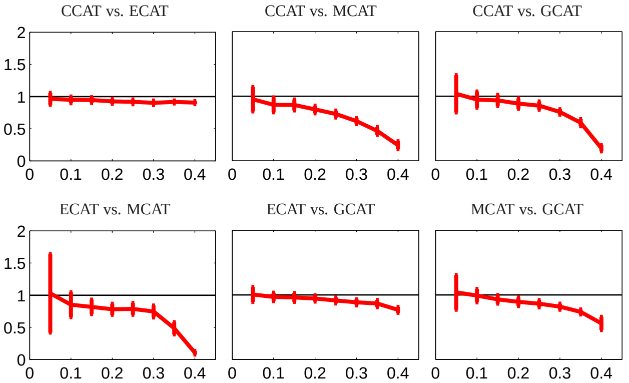

Figure 1. Damage ratio of the algorithm presented in Sec. 3 vs. SVM, as a function of the fraction of evil teachers, examples are assigned to teachers randomly.

wide range of noise levels and on 6 binary problems, each with a different inherent baseline difficulty.

The plots in Fig. 1 show the distribution of damage ratios as a function of k, the number of evil teachers. The effectiveness of our algorithm varies on different binary problems and on different noise-levels, but it consistently performs no worse than SVM. As noise levels increase, the advantage of our algorithm over the naive SVM becomes more profound. On three of the 6 binary problems, when k takes its highest values, the excess-error of our algorithm is a mere 20% of the excess-error of SVM.

6. A Second Algorithm

Drawing intuition from the preceding approach, we derive a second algorithm, which also attempts to limit the influence of any single teacher. Despite its close similarity to our first algorithm, we currently have no theoretical analysis for this second algorithm. We present it here because its empirical behavior is surprisingly different from that of our first algorithm.

The idea behind this algorithm is to apply a constraint similar to the one in Eq. (3) directly to the primal SVM problem, rather than to its dual. Our starting point is a stochastic gradient-descent approach for primal SVM training (Shalev-Shwartz et al., 2007). This algorithm repeatedly draws a random example and performs a gradient-descent step with a decreasing step size. At each step of this process, the current classifier is defined as the linear combination $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$. We modified this algorithm as follows: at each step and for each teacher, we keep track of the average coefficient across all of the examples controlled by that teacher. Namely, for teacher t we keep track of $\frac{1}{|S_t|}\sum_{i \in S_t} \alpha_i$. Before performing the stochastic gradient-descent step, we check if this step will cause the constraint

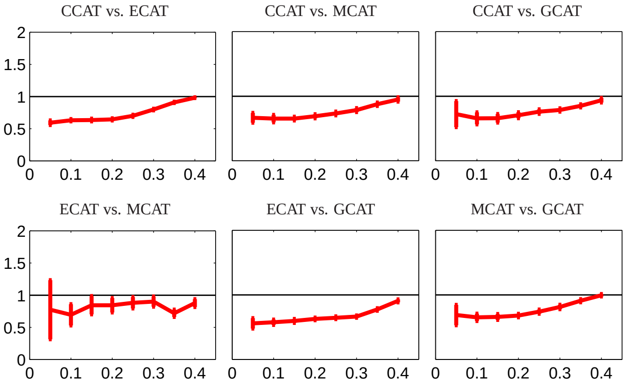

Figure 2. Damage ratio of the algorithm presented in Sec. 6 vs. SVM, as a function of the fraction of evil teachers, examples are assigned to teachers randomly.

$$\frac{1}{|S_t|}\sum_{i \in S_t} \alpha_i \le \frac{1}{m}\sum_{i=1}^{m} \alpha_i + \frac{\epsilon}{m\sqrt{|S_t|}}$$

to be violated. If so, we reduce the update step-size, and set it to the largest non-negative value that still satisfies this constraint, which may even be zero. As a result, no teacher can have a disproportionate influence on our classifier: if the examples of teacher t have already received more than their fair share of updates in the past, the algorithm will compensate by performing smaller updates on the examples of teacher t in the future.

We repeated the experiment outlined in Sec. 5 using the heuristic algorithm and obtained the plots presented in Fig. 2. While the performance of our first algorithm improved with higher levels of label-noise, our second algorithm seems to perform well on low to moderate levels of noise. When k, the number of malicious teachers, equals 10, the excess-error of our second algorithm is $60\% - 70\%$ of the excess-error attained by the standard SVM. However, as the number of evil teachers increases, the advantage of our algorithm deteriorates.

In our experiments so far, each teacher controlled roughly the same number of examples. Moreover, the set of examples controlled by a teacher was chosen randomly. Either of these assumptions may not always hold in practice. Therefore, we also conducted another set of experiments, where we assumed that each teacher has a distinct topic of expertise, and is required to contribute labeled examples from his own topic. In this setting, the examples controlled by two teachers are statistically different, and the number of examples contributed by each teacher may vary greatly.

In addition to the high-level categories mentioned in Sec. 5 above, each article in RCV1 is also associated with one or more *low-level categories*, with 99 different low-level

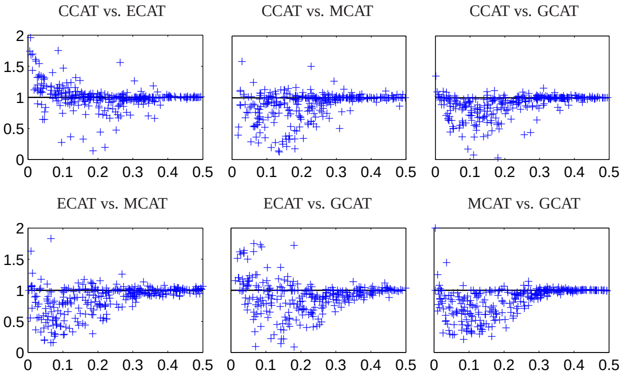

Figure 3. Damage ratio of the algorithm presented in Sec. 6 vs. SVM, as a function of the fraction of evil examples, examples are assigned to teachers by topic.

categories overall. Instead of using the low-level categories to define additional classification problems, we used them to define 99 different expert-teachers. Namely, we assumed that each low-level category corresponds to a different teacher, who controls all of the articles that belong to that low-level category. If an article has more than one low-level category, we assigned it to one of the relevant teachers at random.

As with the previous set of experiments, we choose k random teachers, designated them as malicious, and flipped all of their labels. Since the sizes of the low-level categories vary greatly, we observed cases where 5 evil teachers controlled a large portion of the training set while in other cases 40 evil teachers only controlled a small set of examples. As before, we calculated the damage ratio for 8 values of k, with 40 different train/test splits and random choices of the evil teacher set, for each of our 6 binary problems. For each of these random variations of the noisy classification problem, letting ν denote the fraction of flipped labels, we marked a "+" at location $(\nu, (e_3 - e_1)/(e_2 - e_1))$ in the respective plot in Fig. 3. Since each repetition of the experiment introduced a different amount of noise, it is unclear how to report average results.

Overall, the results in Fig. 2 resemble the results in Fig. 3. In all but the CCAT-ECAT binary problem, our algorithm outperformed the standard SVM a majority of the time. When a low to moderate noise was applied, our algorithm often attained an excess-error that was 40% of the SVM excess-error, or better. On the CCAT-ECAT problem, our algorithm actually performed slightly worse than SVM when a very low noise level was applied, and performed no worse than SVM when a moderate to high noise level was applied. In all 6 binary problems, the advantage of our algorithm could no longer be noticed when the noise level exceeded 35%.

References

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Wortman, J. (2007). Learning bounds for domain adaptation. *Proc. of NIPS 21*.

Crammer, K., Kearns, M., & Wortman, J. (2008). Learning from multiple sources. *Journal of Machine Learning Research*, *5*, 1757–1774.

Dekel, O., Fischer, F., & Procaccia, A. D. (2008). Incentive compatible regression learning. *SODA* (pp. 884–893).

Dekel, O., & Shamir, O. (2008). Learning to classify with missing and corrupted features. *Proc. of ICML 2008* (pp. 216–223).

Kearns, M. J. (1998). Efficient noise-tolerant learning from statistical queries. *J. ACM*, *45*, 983–1006.

Lewis, D., Yang, Y., Rose, T., & Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *J. Machine Learning Research*, *5*, 361–397.

Shalev-Shwartz, S., Singer, Y., & Srebro, N. (2007). Pegasos: Primal estimated sub-gradient solver for svm. *Proc. of ICML 24* (pp. 807–814).

Shawe-Taylor, J., & Cristianini, N. (2000). *Support vector machines and other kernel-based learning methods*. Cambridge University Press.

Sheng, V., Provost, F., & Ipeirotis, P. (2008). Get another label? Improving data quality using multiple, noisy labelers. *Proc. of KDD-08* (pp. 614–622).

Smyth, P., Fayyad, U., Burl, M., Perona, P., & Baldi, P. (1994). Inferring ground truth from subjective labeling of Venus images. *Proc. of NIPS 8* (pp. 1085–1092). MIT Press.

Teo, C. H., Globerson, A., Roweis, S. T., & Smola, A. (2007). Convex learning with invariances. *Proc. of NIPS 21* (pp. 1489–1496).