Reducing Label Complexity by Learning From Bags

Sivan Sabato
The Hebrew University

Nathan Srebro
TTI-Chicago

Naftali Tishby
The Hebrew University

Abstract

We consider a supervised learning setting in which the main cost of learning is the number of training labels and one can obtain a single label for a bag of examples, indicating only if a positive example exists in the bag, as in Multi-Instance Learning. We thus propose to create a training sample of bags, and to use the obtained labels to learn to classify individual examples. We provide a theoretical analysis showing how to select the bag size as a function of the problem parameters, and prove that if the original labels are distributed unevenly, the number of required labels drops considerably when learning from bags. We demonstrate that finding a low-error separating hyperplane from bags is feasible in this setting using a simple iterative procedure similar to latent SVM. Experiments on synthetic and real data sets demonstrate the success of the approach.

1 Introduction

Consider three applications from three different domains: In the first, you want to conduct market research using online ads, to identify which products are attractive. You can put up ads featuring products, but your only feedback is whether or not the ad was clicked. In the second application, consider some chemical or biological problem where the goal is to learn to classify chemical samples based on the result of a chemical experiment. Each experiment is costly, but is possible to conduct an experiment with numerous types of molecules at the same time, and to identify only if a reaction has occurred or not. In the third application, suppose the purpose is to learn a classifier that identifies images with faces, using a large set of labeled images. To obtain this labeled set, one introduces a large set of images to human labelers, who indicate whether the image contains a face or not. We would like to minimize the cost of the human work by reducing the labeling time to a minimum.

These examples come from different domains, but share a common feature: In all of them we have access to practically unlimited data which we can present to a teacher (a human labeler, or some experimental machinery for obtaining a label), but there is a high cost for each label obtained from the teacher. In addition, it is possible to obtain from the teacher a *single label* for *a set of examples* at essentially the same cost as a label for a single example. The single label indicates only if there exists a positive example in the examined set: In the market research application, it is possible to feature several products in one ad. In the chemical experiment task, it may be possible to conduct one large experiment testing several different samples, instead of several experiments, one for each sample. In the face recognition task, one can present test subjects an array of images instead of a single image (see Figure 1) and ask them to indicate whether there is a face anywhere in the array of images[1]. In these example application, the main cost of training is the number of labels, and not the total number of examples.

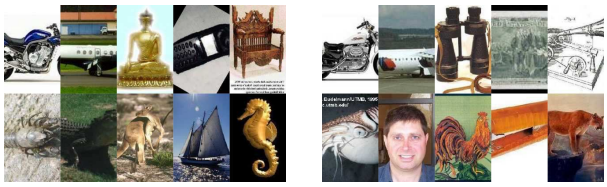

Figure 1: A person easily identifies whether there is a face in a bag of images. Left: Negative Label. Right: Positive Label. Images from CALTECH101 (L. Fei-Fei and Perona., 2004).

We consider learning in the setting illustrated by the three example applications, and investigate when it is worthwhile to present a teacher with sets of examples instead of individual examples in this setting. In our model we assume that the cost of obtaining a label does not depend on the size of the set for which the label was obtained, and that

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

[1] There might be other possibilities, such as asking the labeler to click the exact location of the face in the array, however this might produce a much slower labeling rate than if the labeler clicks only Yes or No buttons

obtaining examples to present to the teacher incurs no cost. Therefore, the cost of learning depends only on the number of obtained labels, and the goal is to reduce this number as much as possible using sets of examples of an optimal size.

The setting in which the teacher labels sets of examples using a single label is generally known as Multi-Instance Learning (MIL) Dietterich et al. (1997). In MIL, the training sample is composed of sets of examples. Sets of examples are termed *bags*, and examples in a bag are termed *instances*. Each bag in the training sample is labeled using the OR rule, so that the label is positive if and only if at least one of the examples in the bag is positive. The goal of the learner is to find the classification rule that classifies the instances or the bags, depending on the application. The general problem of learning from a MIL sample has been theoretically analyzed under various settings (Auer et al., 1998; Blum and Kalai, 1998; Sabato and Tishby, 2009). Many practical algorithms that were tested on real-life benchmarks (Dietterich et al., 1997; Andrews et al., 2002; Andrews and Hofmann, 2003; Maron and Lozano-Pérez, 1998; Maron and Ratan, 1998; Zhang and Goldman, 2001; Zhi-Hua Zhou, 2007; Weidmann et al., 2003, to name a few) have been provided for this problem, but in general it can be a computationally demanding problem for a non-separable sample. The setting presented in this paper is different from previous works on MIL, since here *the bags are created by the learner* for the purpose of classifying individual examples. Therefore the size of the bag may be chosen by the learner, and we also have some control over the distribution of instances within the bag.

Intuitively, there is an inherent trade-off when obtaining one label for a whole bag: On the one hand, this allows one label to provide information on a large number of examples. On the other hand, this information can be ambiguous, since if the label is positive we do not know which examples in the bag are the positive ones. In this work we investigate this trade-off, and show that it is possible to reduce the number of required labels by presenting bags of examples to the teacher instead of individual examples. After describing the formal setting (Section 2), we show, both analytically and experimentally, that using bags can indeed improve performance considerably, for a wide range of problem parameters. We show analytically (Section 3) how to *select the bag size* presented to the teacher for optimal performance. In addition, we propose (Section 4) a simple and practical algorithm along the lines of Felzenszwalb et al. (2008) for finding a separating hyperplane for individual examples from a training sample composed of labeled bags. Several types of experiments were performed (Section 5), on synthetic data sets and on real data sets. The experiments demonstrate the success of the proposed approach for an even wider range of parameters than guaranteed by the analysis.

2 Problem Setting

Let $\mathcal{X}$ be a domain of examples, and let D be an arbitrary and unknown distribution over $\mathcal{X}$ and $c : \mathcal{X} \to \{0, 1\}$ an unknown labeling of the examples. The goal of the learner is to learn c, i.e. to find a function $h : \mathcal{X} \to \{0, 1\}$ with low (true) labeling error $\mathbb{P}_{X\sim D}[h(X) \neq c(X)]$. As in the PAC framework, we will focus on learners that choose h from some fixed hypothesis class $\mathcal{H}$.

We diverge from the classical supervised learning setting in our assumptions on the training set. We assume the learner has unlimited access to samples x drawn from D. We consider the case where the main cost incurred in the learning procedure is that of obtaining labels from the teacher, while the cost of presenting examples to the teacher is negligible. We assume that one can ask the teacher to label *bags* of examples using a single label. The teacher's label indicates whether at least one of the examples in the bag is positive. Formally, denote a bag of size r by $\overline{\mathbf{x}} = (x^1, \ldots, x^r)$, where each $x^i \in \mathcal{X}$ is a single example. Any hypothesis h that labels individual examples can be converted to a bag-labeling hypothesis $\overline{h}$ using the OR rule, so that $\overline{h}(\overline{\mathbf{x}}) = \mathrm{OR}(h(x^1), \ldots, h(x^r))$. We denote $\overline{\mathcal{H}} \triangleq \{\overline{h} \mid h \in \mathcal{H}\}$. For every bag $\overline{\mathbf{x}}$ presented to the teacher, the teacher returns a single binary label $\overline{c}(\overline{\mathbf{x}})$. We wish to get low error over *individual examples*, using the smallest possible number of labels. Note that unlike active learning, here the entire sample is generated in advance, with no feedback from the teacher. The following procedure is proposed:

1. Select a bag size r and a sample size m_r;
2. Create m_r bags of size r from $r \cdot m_r$ examples drawn independently from D;
3. Present the bags $\{\overline{\mathbf{x}}_i\}_{i=1}^{m_r}$ to the teacher, and receive m_r labels $\{y_i\}_{i=1}^{m_r}$ such that $y_i = \overline{\mathbf{c}}(\overline{\mathbf{x}}_i)$.
4. Return the hypothesis $\hat{h} \in \mathcal{H}$ such that $\overline{\hat{h}}$ minimizes the training error over bags:

$$\hat{h} = \operatorname*{argmin}_{h\in\mathcal{H}} \sum_{i=1}^{m_r} |\overline{h}(\overline{\mathbf{x}}_i) - y_i|. \tag{1}$$

This procedure is a generalization of the classical empirical risk minimization (ERM) strategy, where the learner finds the hypothesis with minimal training error: For $r = 1$ this procedure is exactly ERM over an i.i.d. sample drawn from the distribution D. For a general r, we use an i.i.d. sample drawn from the distribution D^r. Importantly, regardless of the chosen r, our goal is to minimize $\mathbb{P}[\hat{h}(X) \neq c(X)]$, the error on *individual examples* drawn from D, and we will measure success based on this goal.

We denote by α the probability of a single example having a positive label, i.e. the frequency of positive examples in

D. As we will see, the methods we describe are relevant when α is substantially smaller then half. That is, when positive examples are relatively rare. When the frequency α of positive examples is small, measuring the error becomes tricky: a hypothesis which labels everything as negative has error α, but we typically want a hypothesis that better balances type I and type II errors. In our analysis we assume for simplicity that all the hypotheses in $\mathcal{H}$ have the same probability for a positive label:

$$\forall h \in \mathcal{H}, \mathbb{P}_{X\sim D}[h(X)=1] = \mathbb{P}_{X\sim D}[c(X)=1] = \alpha. \quad (2)$$

That is, all hypothesis are calibrated by the known positive example rate. This assumption implies that the probability of type I errors is identical to the probability of type II errors, and allows us to use the overall error as a single objective even for very small α. In particular, if the learner balances type I and type II errors, or in the realizable case, if the learner seeks a zero empirical error hypothesis, then the hypotheses chosen by the learner satisfies this condition at least approximately. This assumption also implies that the true error of h is in the range $[0, 2\alpha]$.

3 Theoretical Analysis

In this section we analyze the procedure described above, and show how it can reduce the required number of labels. For simplicity we focus on the realizable case, where $c \in \mathcal{H}$. We start by analyzing the relationship between the bag size, the sample size, and the resulting true error over individual examples, based on theoretical error bounds. We then use these bounds to choose a bag size r and study the reduction in sample size achieved by the proposed procedure.

3.1 The Sample Complexity of Training on Bags

We will base our analysis on standard results, that bound the true error when using ERM on a training sample with a given sample size. These bounds do not suffice by themselves, since they refer to the true error over examples drawn from the same distribution as the training sample. In our case, these results will bound the error over *bags* drawn from D^r, while we wish to bound the true error over *individual examples* drawn from D. We thus start with the following theorem, which provides the relationship between the true error on bags and the true error on individual examples. The proof is provided in the appendix.

Theorem 1. *For any $h : \mathcal{X} \to \{0,1\}$ such that Eq. (2) holds, we have*

$$\mathbb{P}[\overline{h}(\overline{\mathbf{X}}) \neq \overline{c}(\overline{\mathbf{X}})] = \kappa_r^\alpha(\mathbb{P}[c(X) \neq h(X)]) \quad (3)$$

where $\kappa_r^\alpha(\epsilon) \triangleq 2((1-\alpha)^r - (1-\alpha-\epsilon/2)^r)$.

To bound the true error on bags achieved by $\hat{\overline{h}}$, we invoke the VC-bound for the realizable case (Vapnik and Chervonenkis, 1971). With probability at least $1-\delta$ over a sample of m_r bags:

$$\mathbb{P}[\hat{\overline{h}}(\overline{\mathbf{X}}) \neq \overline{c}(\overline{\mathbf{X}})] \leq \quad (4)$$
$$\leq 2\frac{d_r}{m_r}(\log\frac{2em_r}{d_r} + \log\frac{2}{\delta}) \triangleq \text{VC-BOUND}(m_r, d_r),$$

where d_r denotes the VC-dimension of $\overline{\mathcal{H}}$, the class of hypotheses over bags of size r.

Combining Theorem 1 with Eq. (4), and taking the inverse of κ_r^α, yields the following learning bound for the proposed procedure:

Corollary 1. *If Eq. (2) holds and $c \in \mathcal{H}$, and the procedure described in Section 2 is used, then with probability $1-\delta$ over the samples of bags,*

$$\mathbb{P}[\hat{h}(X) \neq c(X)] \leq$$
$$2(1-\alpha) - 2((1-\alpha)^r - \text{VC-BOUND}(m_r, d_r)/2)^{1/r}.$$

In order to understand the effect of using bags, it will be useful to study the relationship between the bag size and the sample complexity, based on the bound in Corollary 1 (Note that the sample size is equal to the number of labels, which is the cost we wish to minimize). We will thus fix a target error rate, and ask how the sample complexity for this error rate changes as a function of the bag size. To this end, define $\tilde{m}_r(\epsilon)$ as the number of bags of size r required to obtain a bound of ϵ on the true error of individual examples, based on Corollary 1. This is an upper bound on the sample complexity when using bags of size r. In particular, $\tilde{m}_1(\epsilon)$ is the "standard" VC-bound sample complexity, when using a regular sample with individual examples. The following theorem bounds the reduction in sample complexity when bags of size r are used instead of a regular sample:

Theorem 2. *Let d be the VC-dimension of $\mathcal{H}$, and let d_r be the VC-dimension of the class $\overline{\mathcal{H}}$ of hypotheses over bags of size r. We have:*

$$\frac{\tilde{m}_r(\epsilon)}{\tilde{m}_1(\epsilon)} \leq \frac{\epsilon}{\kappa_r^\alpha(\epsilon)} \cdot \frac{d_r}{d}. \quad (5)$$

Proof. Let $m_r = \min\{\tilde{m}_1(\epsilon)\frac{\epsilon}{\kappa_r^\alpha(\epsilon)} \cdot \frac{d_r}{d}, \tilde{m}_1(\epsilon)\}$. We have

$$\mathbb{P}[\hat{\overline{h}}(\overline{\mathbf{X}}) \neq \overline{c}(\overline{\mathbf{X}})] \leq \text{VC-BOUND}(m_r, d_r) =$$
$$\frac{d_r\tilde{m}_1(\epsilon)}{dm_r} \cdot 2\frac{d}{\tilde{m}_1(\epsilon)}(\log\frac{2em_r}{d_r} + \log\frac{2}{\delta})$$
$$\leq \frac{d_r\tilde{m}_1(\epsilon)}{dm_r}\text{VC-BOUND}(\tilde{m}_1(\epsilon), d) = \frac{d_r\tilde{m}_1(\epsilon)\cdot\epsilon}{dm_r} \leq \kappa_r^\alpha(\epsilon).$$

From Theorem 1 it follows that $\mathbb{P}[\hat{h}(X) \neq c(X)] \leq \epsilon$. Therefore the minimal sample size to achieve ϵ using bags of size r is no more than m_r, and Eq. (5) follows. □

Examining Eq. (5), it is obvious that $\frac{d_r}{d} \geq 1$, since the hypotheses class over bags cannot have a lower VC-dimension then the hypotheses class over individual examples. Therefore a reduction in sample complexity will only be attained if $\kappa_r^\alpha(\epsilon) > \epsilon$. That is, only if the error rate on bags is *higher* then the error rate on individual examples. This may seem counterintuitive—why would we gain from using bags if it causes an *increase* in the error rate? The key point is that we are interested in the implied error rate on individual examples, and so we can allow ourselves a higher error rate on bags, if it implies a lower error on individual examples. Note, however, that any reduction in the sample complexity due to $\kappa_r^\alpha(\epsilon) > \epsilon$ might be canceled if the VC-dimension d_r grows very fast with r. Fortunately, this is not the case, as the following theorem shows:

Theorem 3.

$$d_r \leq -\frac{d}{\ln(2)} \cdot W_{-1}\left(-\tfrac{\ln(2)}{er}\right) = O(d \log r). \quad (6)$$

Where W_{-1} denotes the negative branch of the Lambert W function, $x = W(x)e^{W(x)}$.

This result is derived from the implicit inequality on d_r (Sabato and Tishby, 2009): $d_r \leq d \log_2(e \cdot r d_r / d)$. The full derivation is provided in the appendix. Equipped with Theorems 2 and 3, we can now study the optimal bag size and the reduction in sample complexity it affords.

3.2 Choosing the Bag Size

We now turn to the question of how to choose a bag size r so as to minimize the sample complexity $\tilde{m}_r(\epsilon)$. The two important parameters here are the positive example rate α and the desired error guarantee ϵ. Intuitively, it can be speculated that a good size for a bag is such that the labels on bags are distributed more or less evenly, such that every label received from the teacher conveys a large amount of information to the learner. Thus r should be larger for smaller α. The bag size r should also grow as ϵ is reduced, since larger bags imply a higher sensitivity to error. The following analysis corroborates this intuition, and quantifies the dependence on both ϵ and α.

Following Theorem 2, we would like to choose r such that $\kappa_r^\alpha(\epsilon)/d_r$ is maximal. However, since $d \leq d_r \leq O(d \log r)$, i.e. d_r grows relatively slowly with r, we ignore the exact value of d_r, and define our choice for the bag size as the value of r that maximizes $\kappa_r^\alpha(\epsilon)$:

$$\begin{aligned} r^*(\alpha,\epsilon) &\triangleq \operatorname*{argmax}_r \kappa_r^\alpha(\epsilon) \\ &\equiv \operatorname*{argmax}_r \left[(1-\alpha)^r - (1-\alpha-\epsilon/2)^r\right]. \end{aligned}$$

We shall see that though this choice is not necessarily optimal, it provides a substantial reduction in sample size. Numerical calculations show that using the upper bound for d_r does not change the resulting sample size significantly.

Differentiating $\kappa_r^\alpha(\epsilon)$ we obtain a single maximum in r for all $0 < \alpha < 0.5, 0 < \epsilon < 2\alpha$:

$$r^*(\alpha,\epsilon) = \ln\left(\frac{\ln(1-\alpha-\epsilon/2)}{\ln(1-\alpha)}\right) \Big/ \ln\left(\frac{1-\alpha}{1-\alpha-\epsilon/2}\right). \quad (7)$$

As our preliminary intuition implied, $r^*(\alpha,\epsilon)$ is monotonic decreasing in α and in ϵ. We also speculated that the labels on bags of an optimal size should be balanced. Defining $r^*(\alpha,0) \triangleq \lim_{\epsilon\to 0^+} r^*(\alpha,\epsilon)$, we have $r^*(\alpha,0) = -1/\ln(1-\alpha) \approx 1/\alpha$. For this value of r^*, $\mathbb{P}[\overline{c}(\overline{\mathbf{X}}) = 1] = 1 - 1/e$ and the expected number of positive examples in each bag is approximately one. Figure 2 plots $\mathbb{P}[\overline{c}(\overline{\mathbf{X}}) = 1]$ as a function of α. The gray area between the two boundaries corresponds to different values of ϵ, in the range $(0, 2\alpha]$. This plot shows that choosing the bag size to be $r^*(\alpha,\epsilon)$ results in an almost constant probability of obtaining positive labels, confirming our intuition.

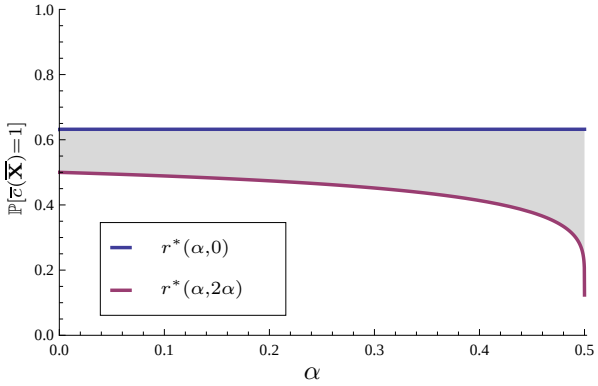

Figure 2: The probability for a positive bag.

3.3 The Sample Size Reduction Factor

We can now ask whether our choice of r^* leads to a reduction in the sample size, and how large is this reduction. Substituting Eq. (7) and Eq. (6) in Eq. (5), yields an upper bound on $\tilde{m}_{r^*}(\epsilon)/\tilde{m}_1(\epsilon)$, the sample size reduction factor when using a bag of size r^*. For $\epsilon \to 0$ we have a simplified form:

Corollary 2.

$$\lim_{\epsilon\to 0^+} \frac{\tilde{m}_{r^*(\alpha,\epsilon)}(\epsilon)}{\tilde{m}_1(\epsilon)} \leq (1-\alpha)\ln(1-\alpha) \cdot W_{-1}\big(\ln(2)\ln(1-\alpha)/e\big) \cdot \frac{e}{\ln(2)}.$$

The bound for $\epsilon \in [0, 2\alpha]$ is plotted in Figure 3. Whenever the bound is smaller than 1, using bags of size r^* results in a guaranteed sample size reduction. From the figure it can be seen that this holds for $\alpha < 0.04$. This result is only a worst case bound; The experiments described in Section 5

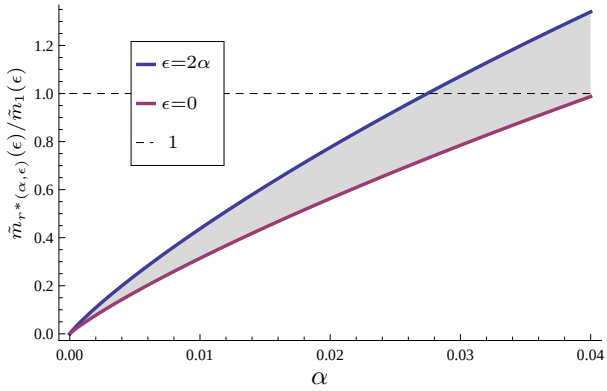

Figure 3: Sample size reduction factor. Anything below 1 implies a multiplicative reduction.

show that in practice an even larger reduction is achieved, and that it is achieved for larger α as well.

4 Finding a Separating Hyperplane using Bags: The PMIL Algorithm

The analysis above provides bounds on the required sample size under the assumption that it is possible to find the hypothesis with the lowest training error on samples of bags of an arbitrary size. We now turn to show how one might find the correct hypothesis efficiently. This problem is not trivial, since it is not known which are the positive examples in a positive bag. Learning from bags with arbitrary distribution is theoretically solvable in the almost realizable case (Sabato and Tishby, 2009), however there is no algorithm that is guaranteed to work with the small sample size that our learning bounds allow. Many heuristic algorithms have also been proposed for MIL (Andrews et al., 2002; Andrews and Hofmann, 2003; Dietterich et al., 1997; Zhi-Hua Zhou, 2007, and others). These algorithms are typically quite involved, as they must deal with samples of bags with arbitrary dependence between instances. Luckily, though the MIL problem is hard in general, our setting only employs bags with statistically independent instances, which can be expected to be a much easier problem. This case is also provably solvable (Blum and Kalai, 1998), but again only by using a large sample size.

We propose PMIL (Table 1), a simple iterative algorithm for finding a separating hyperplane from samples of bags, following ideas from Felzenszwalb et al. (2008). PMIL executes the basic perceptron algorithm several times on different input samples, using parameters T and L. Though PMIL is a local-search algorithm for a non-convex objective and so might potentially find only a local minimum, it was very successful in our experiments (see Section 5), and has almost always found the separating hyperplane with zero or close to zero mistakes. This indicates that it is practically feasible to reduce the number of required labels using bags of independent examples. We defer the comparison of PMIL to other possible heuristics to future work.

Table 1: The PMIL algorithm

1. Initialize a separator w randomly;
2. Repeat until T time has passed, or until w classifies the bags with zero training error:
   (a) For each bag $\overline{\mathbf{x}}_k = (x_k^1, \ldots, x_k^r)$, select a representative example from the bag with index $i_k = \mathrm{argmax}_i (w \cdot x_k^i)$,
   (b) Run L epochs of the perceptron algorithm on the sample of individual examples $\{(x_k^{i_k}, y_k)\}_{i=1}^m$.

5 Experiments

In this section we present the results of experiments done on several types of learning problems. In the first batch of experiments, presented in Section 5.1, the procedure is tested on a finite hypothesis class, using an exhaustive search for the hypothesis with the lowest training error. This allows us to inspect the learning curves of the true $\hat{h}$, without needing to worry about the possible sub-optimality of the PMIL algorithm. Then, in Section 5.2, we show that the PMIL algorithm is indeed successful on both synthetic and real data sets. The experiments demonstrate a significant sample size reduction that is even better than the one promised by the analysis. They further demonstrate that using bags improves performance even when the simplifying assumption that $c \in \mathcal{H}$ does not hold. Moreover, it is shown that even using a small bag size yields a significant improvement.

5.1 Finite hypothesis class

We start by examining the actual sample complexity behavior, with experiments on a finite hypothesis class, where the hypothesis with lowest training error is found using exhaustive search. We generated random examples from the domain $\mathcal{X} = \{0,1\}^{1000}$, with each of the 1000 features drawn independently with a positive example rate of α, for various values of α. The examples were labeled with a hypothesis from the class $\mathcal{H} = \{h_1, \ldots, h_{1000}\}$, where $h_i(x)$ is the value of the i'th coordinate of x. Each experiment reported was repeated either 100 or 1000 times. The plots show the average true error that was achieved.

First, we wanted to check the effect of the proposed bagging strategy on the output error on individual examples: If we fix the sample size, is there an optimal bag size $r > 1$ that achieves the lowest error? How close is the empirical optimal r to our $r^*(\alpha, \epsilon)$? Figure 4 shows the average true error of the learned hypothesis as a function of the bag size, for different sample sizes, and for two values of α. Even for α as large as 0.2, using bags reduces the achieved error with

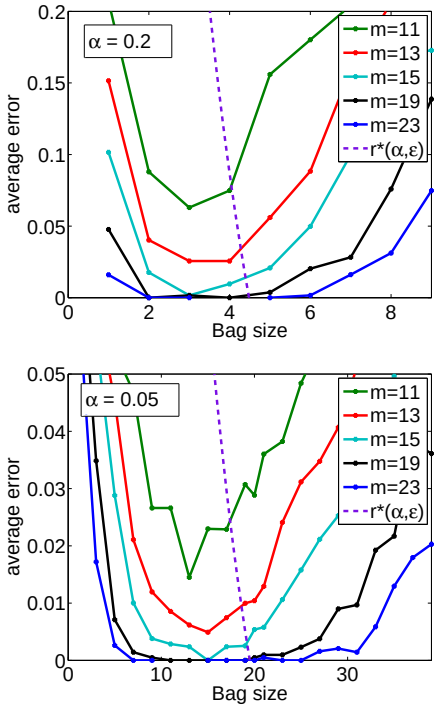

Figure 4: Experiments on a finite hypothesis class for two different α. Plots show the error as a function of the bag size, for several sample sizes m.

a fixed sample size. The dips in the plot lines indicate the existence of an optimal bag size, as predicted by the theoretical analysis. The calculated $r^*(\alpha, \epsilon)$, indicated with the dashed line, is quite close to the empirical optimum in both plots, and yields almost optimal performance.

To visualize the improvement in learning performance compared to regular supervised learning, we plotted the learning curves for selected bag sizes. The plots in Figure 6 compare the achieved error as a function of the sample size, for three bag sizes: one, two, and $r^*(\alpha, 0)$ (rounded). The left and middle plots show results for two values of α, with no label noise. We see a sharp improvement in performance for $r \sim r^*(\alpha, 0)$. The improvement is sharper for the smaller α. Note also, that even a bag with only two examples delivers a much better result than when using no bags. This means that a considerable improvement can be achieved even in an application that allows only small bag sizes.

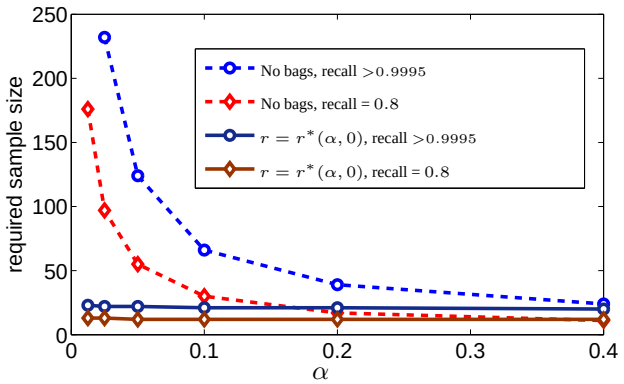

Figure 5: The sample size to achieve a fixed recall. Compare dashed lines (no bags) to solid lines ($r = r^*$).

One of the assumptions in our theoretical analysis was that $c \in \mathcal{H}$. We now deviate from this assumption by adding randomly flipping some of the labels creating a situation where the optimal hypothesis has error $0.017 = \alpha/3$. The right plot in Figure 6 shows that even when label noise is high compared to α, bagging improves the achieved error rate considerably.

Finally, we show a striking comparison between the required sample size when learning with no bags, to the required sample size when bags of optimal size are used. We have seen in the analysis that the positive example rate α is a significant parameter affecting optimal bag size and expected improvement when using bags. As α decreases, labels on single examples become less balanced. In regular learning, this means that more examples are required for effective learning. Since it is less informative to compare absolute error for varying α, Figure 5 examines the effect of α on the outcome *recall* (the fraction of positive examples which are identified by the output hypothesis; Note that by Eq. (2), the precision is also controlled). When learning without bags (dashed lines), the required sample size for a fixed recall value grows fast as α decreases. In contrast, when bags of size $r^*(\alpha, 0)$ are used (solid lines), the effect of α disappears completely. Thus, the use of bags almost eliminates the effects of unbalanced labels, by changing the bag size according to α.

5.2 Experiments Using `PMIL`

Having investigated the sample complexity effects of the use of bags, we now turn to more realistic experiments, where $\mathcal{H}$ is the set of separating hyperplanes, and `PMIL` is used to find a separator. In each setting we applied the procedure in Table 1 several times, until a separator with perfect classification on the sample of bags was found, or one second of runtime had passed. If a second had passed, we selected the separator that produced the lowest number of errors. L was set to 10.

The first set of experiments was on synthetic examples with no label noise, drawn uniformly from $\mathcal{X} = [0, 1]^{10}$. A positive label was a assigned to a fraction of size α of the cube. We performed the experiments with different sample sizes, bag sizes, and values of α. `PMIL` usually succeeded in achieving zero or almost zero error on the training set. Even for a bag size of 19, the algorithm usually finished with a negligible number of errors. Figure 7 compares the learning curves when using bags and without the use of bags for two values of α. Each dot in is the average of 1000 exper-

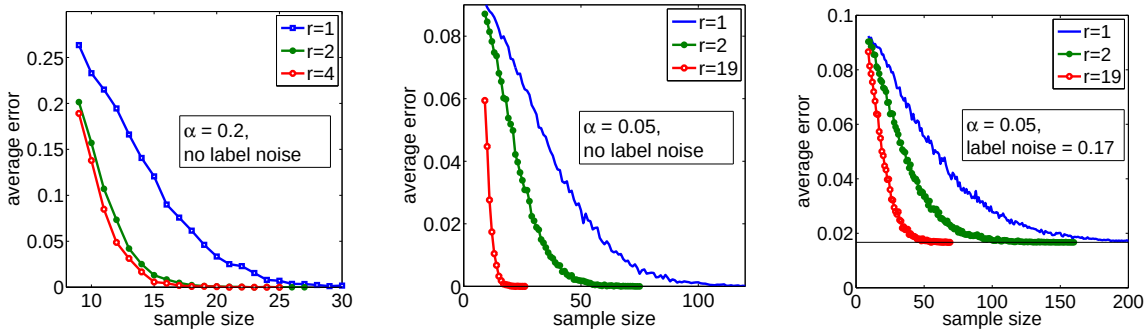

Figure 6: Learning curves for the finite hypothesis class, with different values of α: comparing no use of bags, bags of size 2, and bags of size $r^*(0, \alpha)$. In the right plot, some of the labels were randomly flipped.

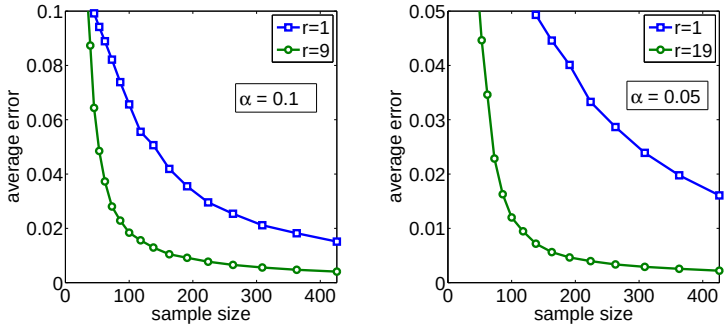

Figure 7: Learning synthetic data using PMIL, For two different α. The optimal bag size produces a significant improvement over $r = 1$

iments. Here too the improvement in performance when using bags is clearly visible.

Next, we tested our learning procedure on real data sets, using samples of bags created from the original labeled examples. The first data set is the **Statlog (Shuttle) dataset** (Asuncion and Newman, 2007). It was chosen due to the relative ease of classification using regular supervised learning, which allowed us to investigate the results of using bags in multiple experiments. To make the original multi-class problem into a binary classification problem, we selected from the training set and from the test set only examples with class 1 and 5. Class 5 was mapped to a positive label. Its occurrence in the data set is $\alpha = 0.067$, thus $r^*(\alpha, 0) \sim 14.5$. The results are plotted in Figure 8. On the left is the error as a function of the bag size for different sample sizes, showing that the lowest error is achieved, as expected, around $r = 14$. In the middle we compare the learning curve between learning with no bags, with bags of size 2, and with $r = 14$. Here too even a bag size of 2 provides a large improvement in the error.

The second real data set we learned with PMIL was the **Caltech101 image data set** (L. Fei-Fei and Perona., 2004), exemplified in Figure 1. The positive class was the Faces_easy category. The negative class was all the categories except for Faces and BACKGROUND_Google, since they contain images of faces. We built a random training set of 3850 images and a random holdout set of 500 images. In both sets the we set the fraction of faces to $\alpha = 0.1$. We extracted 1000 features from the training images using k-means clustering on interest points detected as in Mikolajczyk and Schmid (2004), with default parameters. PMIL was applied to the resulting feature vectors with several bag sizes and sample sizes. Because of the default feature extraction methodology and the relatively small number of examples of faces, the best error rate that could be reached using individual examples was quite high compared to α, and only small bag sizes could be tested. Figure 8 (Right) compares the learning curves for $r = 1$, $r = 2$ and $r = 5$, which are lower than $r^*(\alpha, 0) \sim 9.5$. An interesting effect can be seen: When the sample size is small, it is better to use bags of a smaller size. As the sample grows, larger bags become more beneficial.

6 Summary

We studied a novel paradigm for learning from a labeled sample using a teacher that can provide OR-labels, when the cost of obtaining labels from the teacher is high, while the cost of presenting examples to the teacher is negligible. We demonstrated that a significant improvement in the error can be achieved with a fixed amount of labels, by presenting to the teacher bags of examples instead of individ-

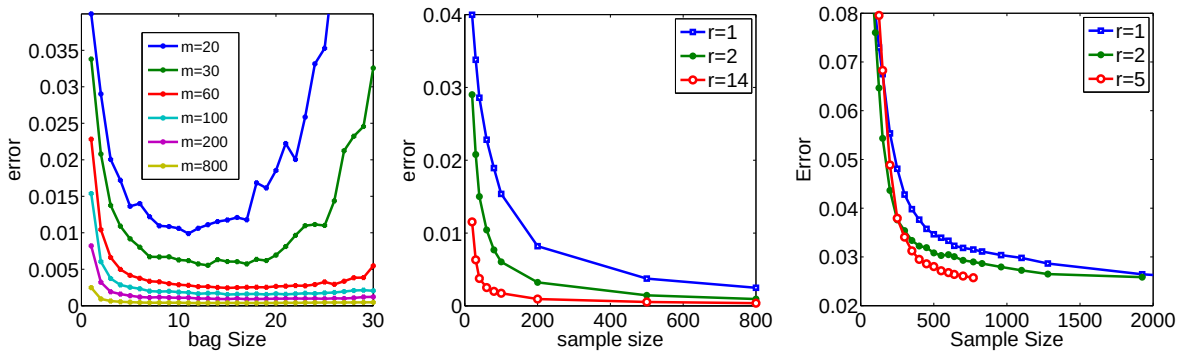

Figure 8: Left and Middle: Experiments on the Statlog data set ($\alpha = 0.067$). Left: the error as a function of the bag size. Each line is a sample size. Middle: Learning curves, comparing bag sizes of 1 (no bags), 2, and 14. Right: Classifying images with faces ($\alpha = 0.1$) – learning curves, comparing three bag sizes.

ual examples. We have shown that the size of the bag that should be used has an optimum and that an almost optimal bag size can be analytically found. The `PMIL` algorithm was proposed for finding a separating hyperplane with low training error from a sample of bags. Experiments on various types of data sets demonstrate that the proposed method and learning algorithm work well in practice, and that the method can be used even if the exact problem parameters are not known. Many aspects in this new paradigm call for further work. One important issue is a more general analysis of the problem that would dispense with the simplifying assumptions, and allow for a non-negligible cost of presenting examples to the teacher, and a labeling error that depends on the bag size.

Acknowledgments

We thank Alon Zweig and Amit Gruber for their help with image data analysis. Sivan Sabato is supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities.

References

S. Andrews and T. Hofmann. Multiple-instance learning via disjunctive programming boosting. In *NIPS*, 2003.

S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, pages 561–568, 2002.

A. Asuncion and D. Newman. UCI machine learning repository, 2007.

P. Auer, P. M. Long, and A. Srinivasan. Approximating hyper-rectangles: learning and pseudorandom sets. *J. Comput. Syst. Sci.*, 57(3):376–388, 1998.

A. Blum and A. Kalai. A note on learning from multiple-instance examples. *Mach. Learn.*, 30(1):23–29, 1998.

T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89(1-2):31–71, 1997.

P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8, 2008.

R. F. L. Fei-Fei and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *CVPR 2004, Workshop on Generative-Model Based Vision*. IEEE, 2004.

O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *NIPS '97: Proceedings of the 1997 conference on Advances in neural information processing systems 10*, pages 570–576, Cambridge, MA, USA, 1998. MIT Press.

O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, pages 341–349, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1): 63–86, 2004.

S. Sabato and N. Tishby. Homogeneous multi-instance learning with arbitrary dependence. In *COLT*, 2009.

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its applications*, XVI (2):264–280, 1971.

N. Weidmann, E. Frank, and B. Pfahringer. A two-level learning method for generalized multi-instance problems, 2003.

Q. Zhang and S. Goldman. Em-dd: An improved multiple-instance learning technique. In *Advances in Neural Information Processing Systems 14*, 2001.

M.-L. Z. Zhi-Hua Zhou. Solving multi-instance problems with classifier ensemble based on constructive cluster-

ing. *Knowledge and Information Systems*, 11(2):155–170, 2007.

A Proofs omitted from the text

Proof of Theorem 1. Let $X \sim D$ be a random variable over individual examples and $\overline{\mathbf{X}} \sim D^r$ be a random variable over bags. We have

$$\begin{aligned}
&\mathbb{P}[\overline{h}(\overline{\mathbf{X}}) \neq \overline{c}(\overline{\mathbf{X}})] = \qquad (8)\\
&= \mathbb{P}[\overline{h}(\overline{\mathbf{X}}) = 0 \wedge \overline{c}(\overline{\mathbf{X}}) = 1] + \mathbb{P}[\overline{h}(\overline{\mathbf{X}}) = 1 \wedge \overline{c}(\overline{\mathbf{X}}) = 0]\\
&= \mathbb{P}[\overline{h}(\overline{\mathbf{X}}) = 0] - \mathbb{P}[\overline{h}(\overline{\mathbf{X}}) = 0 \wedge \overline{c}(\overline{\mathbf{X}}) = 0] +\\
&\quad \mathbb{P}[\overline{c}(\overline{\mathbf{X}}) = 0] - \mathbb{P}[\overline{c}(\overline{\mathbf{X}}) = 0 \wedge \overline{h}(\overline{\mathbf{X}}) = 0]\\
&= \mathbb{P}[\overline{h}(\overline{\mathbf{X}}) = 0] + \mathbb{P}[\overline{c}(\overline{\mathbf{X}}) = 0] - 2\mathbb{P}[\overline{c}(\overline{\mathbf{X}}) = 0 \wedge \overline{h}(\overline{\mathbf{X}}) = 0].
\end{aligned}$$

Since the instances in $\overline{\mathbf{X}}$ are statistically independent we have $\mathbb{P}[\overline{c}(\overline{\mathbf{X}}) = 0] = (\mathbb{P}[c(X) = 0])^r = (1-\alpha)^r$.. From Eq. (2) we also have

$$\mathbb{P}[\overline{h}(\overline{\mathbf{X}}) = 0] = \mathbb{P}[\overline{c}(\overline{\mathbf{X}}) = 0] = (1-\alpha)^r.$$

In addition,

$$\begin{aligned}
&\mathbb{P}[\overline{c}(\overline{\mathbf{X}}) = 0 \wedge \overline{h}(\overline{\mathbf{X}}) = 0] = (\mathbb{P}[c(X) = 0 \wedge h(X) = 0])^r =\\
&\quad (1 - \alpha - \mathbb{P}[c(X) \neq h(X)]/2)^r.
\end{aligned}$$

The last equality above follows from Eq. (2) using some simple calculations. Using the three equations above in Eq. (8), we get

$$\mathbb{P}[\overline{h}(\overline{\mathbf{X}}) \neq \overline{c}(\overline{\mathbf{X}})] = 2((1-\alpha)^r - (1-\alpha-\mathbb{P}[c(X) \neq h(X)]/2)^r).$$

Eq. (3) follows from this equality by setting $\epsilon = \mathbb{P}[c(X) \neq h(X)]$. □

Proof of Theorem 3. We start with the following bound from the proof of theorem 5 in Sabato and Tishby (2009):

$$d_r \leq d(\log_2(erd_r/d)).$$

We reorganize this bound to find an upper bound for d_r:

$$\begin{aligned}
d_r &\leq d(\log_2(erd_r/d)) &\Rightarrow\\
d_r \ln(2) &\leq d(\ln(er/d) + \ln(d_r)) &\Rightarrow\\
d_r \ln(2) - d\ln(d_r) &\leq d\ln(er/d) &\Rightarrow\\
d\ln(d_r) - d_r\ln(2) &\geq -d\ln(er/d) &\Rightarrow\\
\ln(d_r) - \tfrac{\ln(2)}{d} d_r &\geq -\ln(er/d) &\Rightarrow\\
d_r \exp(-\tfrac{\ln(2)}{d} d_r) &\leq d/er &\Rightarrow\\
-\tfrac{\ln(2)}{d} d_r \exp(-\tfrac{\ln(2)}{d} d_r) &\leq -\ln(2)/er.
\end{aligned}$$

Since $r \geq 1$, we have that $-\frac{1}{e} \leq -\ln(2)/er \leq 0$. From the properties of the Lambert function we have that for $-\frac{1}{e} \leq x \leq 0$, $we^w \leq x \Rightarrow w \geq W_{-1}(x)$. Therefore

$$\begin{aligned}
-\tfrac{\ln(2)}{d} d_r &\geq W_{-1}(-\ln(2)/er) &\Rightarrow\\
d_r &\leq -\tfrac{d}{\ln(2)} W_{-1}(-\ln(2)/er).
\end{aligned}$$

□