Learning Instance Concepts from Multiple-Instance Data with Bags as Distributions

Gary Doran and **Soumya Ray**
Department of Electrical Engineering and Computer Science
Case Western Reserve University
Cleveland, OH 44106, USA
{gary.doran,sray}@case.edu

Abstract

We analyze and evaluate a generative process for multiple-instance learning (MIL) in which bags are distributions over instances. We show that our generative process contains as special cases generative models explored in prior work, while excluding scenarios known to be hard for MIL. Further, under the mild assumption that every negative instance is observed with nonzero probability in some negative bag, we show that it is possible to learn concepts that accurately label instances from MI data in this setting. Finally, we show that *standard supervised approaches* can learn concepts with low area-under-ROC error from MI data in this setting. We validate this surprising result with experiments using several synthetic and real-world MI datasets that have been annotated with instance labels.

Introduction

In the multiple-instance (MI) setting (Dietterich, Lathrop, and Lozano-Pérez 1997), an algorithm is given a training set of labeled *bags*, where each bag is a set of *instances*, described by feature vectors. Instances are presumed to have unobserved labels that obey a constraint: if a bag is labeled positive, then *at least one* instance in that bag is labeled positive; otherwise, *all* instances in the bag are labeled negative. An example of such a task is 3-Dimensional Quantitative Structure–Activity Relationship (3D-QSAR), where a positively labeled bag is a molecule that strongly binds to some target protein, while a negatively labeled molecule does not bind. Instances in this domain are *conformations* of the molecule. Given such data, there are two related learning tasks: learn a concept that can predict the labels of new bags (a "bag concept"), and/or learn a concept that can predict the labels of individual instances within bags (an "instance concept"). In 3D-QSAR, a bag concept would identify whether a new molecule binds to the target. An instance concept would identify whether a specific conformation of a new molecule binds.

In this work, we study a generative model for MIL in which bags are *distributions over instances*. To motivate our generative model, consider 3D-QSAR. In nature, a molecule exists in a dynamic equilibrium over its possible conformations. Conformations of a molecule are distributed according to their Gibbs free energy such that low-energy conformations are preferred. Thus, constructing a bag from low-energy conformations can be thought of as sampling instances from this distribution. We could ask of every conformation whether or not it is a *binding conformation* that activates the target, so there is a labeling function for instances corresponding to the instance concept. Similar arguments can be made for other MI problem domains as well.

Not only is this generative process more realistic, it is also less restrictive than other models proposed in the literature. For example, prior work has proposed a model where instances in all bags are drawn independently from the same distribution (Blum and Kalai 1998). However, in 3D-QSAR, this assumption obviously does not hold, since it requires that different molecules share the same conformations. A relaxation of this model assumes that bags are arbitrary tuples of some maximum size (Sabato and Tishby 2012). However, a molecule can transform dynamically from conformation to conformation, producing an infinite set of conformations. Other models attempt to cast bags as manifolds in the instance space (Babenko et al. 2011), but these models do not capture the distributional nature of bags in examples such as the one above in which low-energy conformations are more relevant to the prediction problem. Related generative models have been studied in prior work (Xu 2003; Behmardi et al. 2012), but learnability under these models is not explored. While some prior work assumes that bags can be modeled as parametric distributions (Xu 2003), we do not assume any specific parametric form for our data to prove our results. In a later section we describe how our model generalizes some of these alternatives, and the consequences for the theoretical results we derive.

Our analysis of this generative process yields two new theoretical results in MIL (under a mild assumption, described later). First, we affirmatively answer the question of whether instance concepts can be learned from MI data generated according to our model. Second, and more surprisingly, we show that by labeling all instances with their bags' labels (thus constructing a supervised dataset) and applying standard supervised learning approaches, *we can learn a good ranking over instances* (i.e., a high-area-under-ROC concept). In many applications, such a ranking of the data suffices as output. For example, in 3D-QSAR, the output of the computational affinity-prediction module is typically a ranked list for further testing by chemists. Our theoretical analysis suggests that a *supervised* approach applied to MI data will be successful at solving such problems. We support this theoretical result with an empirical evaluation using a number of synthetic and real-world MI datasets annotated with instance labels (Settles, Craven, and Ray 2008). The experiments reveal that although an instance concept learned with a supervised approach performs poorly with respect to accuracy, the same approach can often find a very good solution when evaluated using area under ROC (AUC)! We then discuss the implications of this counterintuitive result for future empirical work in MIL.

Bags as Distributions

We start by formalizing our proposed generative model. Let $\mathcal{X}$ be a space of instances, then the space of bags $\mathcal{B}$ is the set of probability distributions over $\mathcal{X}$. For each bag $B \in \mathcal{B}$, we refer to the corresponding bag-specific distribution over instances using the notation $\Pr(x \mid B)$. Let $D_{\mathcal{B}}$ be a distribution *over* bags. The labeling function $F : \mathcal{B} \rightarrow \{0, 1\}$ in this generative model operates at the level of *bags* so that $Y = F(B)$. Thus, we generate an MI dataset by first sampling from $D_{\mathcal{B}}$ a set of labeled bags $\{(B_i, Y_i)\}_{i=1}^{n}$. In special cases, we might directly observe B_i (e.g., we might sample the parameters of B_i when it is a parametric distribution). In the typical case, we only have access to samples, $X_i = \{x_{ij}\}_{j=1}^{m_i}$, each drawn according to the distribution corresponding to the bag B_i so that $\left\{\left(\{x_{ij}\}_{j=1}^{m_i}, Y_i\right)\right\}_{i=1}^{n}$ is the observed MI dataset.

Unlike previous models (e.g., Blum and Kalai 1998), our generative model allows for each bag to have its own distribution, which allows for bag-specific relationships (in a probabilistic sense) between instances. One key observation of our work is that, similar to previous work (Blum and Kalai 1998), we can still obtain a single bag-independent instance distribution $D_{\mathcal{X}}$ by *marginalizing out individual bag-specific distributions.* That is, given a probability measure $\Pr_{\mathcal{B}}$ over bags corresponding to $D_{\mathcal{B}}$, we can define a measure $\Pr_{\mathcal{X}}$ corresponding to $D_{\mathcal{X}}$ using: $\Pr_{\mathcal{X}}(x) = \int_{\mathcal{B}} \Pr(x \mid B) \,\mathrm{d} \Pr_{\mathcal{B}}(B)$. In our generative model, $D_{\mathcal{X}}$ is the distribution of the random variable observed if we first sample a bag B from $D_{\mathcal{B}}$, then sample a single instance from $\Pr(x \mid B)$.

Now, in addition to the bag-labeling function F, in a proper MI generative model there also exists an instance-labeling function $f : \mathcal{X} \rightarrow \{0, 1\}$. Traditionally, the MI assumption is stated with respect to a particular dataset so that $F(B_i) = \max_j f(x_{ij})$ for real-valued labels. That is, a bag is positive if *at least one* instance in it is positive, and negative if *all* instances are negative. Since we assume that F is defined *a priori* for bags, which are no longer finite sets of instances, we encode a relationship between F and f at the level of the generative process by enforcing that the probability of sampling a positive instance in any negative bag is zero. To formalize this relationship between labeling functions, suppose we sample instances from $D_{\mathcal{X}}$ according to the two-level sampling procedure described above, but we record the bag labels to obtain labeled singletons $\{(\{x_i\}, Y_i)\}_{i=1}^{n}$. The probability that $Y = 1$ is:

$$\Pr(Y = 1 \mid x) = \frac{\int_{\mathcal{B}_+} \Pr(x \mid B) \,\mathrm{d} \Pr(B)}{\int_{\mathcal{B}} \Pr(x \mid B) \,\mathrm{d} \Pr(B)} \triangleq c(x), \qquad (1)$$

where $\mathcal{B}_+$ is the set of positive bags and c is a probabilistic concept (p-concept) (Kearns and Schapire 1994). This is the probability of observing x within a positive bag. In the next section, we formally relate f and c and show that instance concepts are learnable in our generative model. Later, we explain the relationship between our generative process and those studied in prior work.

Learning Accurate Instance Concepts

We first extend the definition of PAC-learnability to instance concepts in the MI setting. Since we are interested in instance concepts, similar to the supervised case, we wish to place no restrictions on the instance distribution $D_{\mathcal{X}}$. However, in our case, the bag and instance distributions and labeling functions must jointly satisfy the relationships described in the section above. In particular, a bag must be negative only if there is a zero chance that it contains a positive instance. We will also need the following mild assumption: all negative instances appear with nonzero probability in negative bags.

To understand the intuition behind this condition, consider learning the instance concept in a content-based image retrieval (CBIR) problem where interesting images contain spoons and uninteresting images do not. However, suppose that every image you are shown that has a spoon also contains a fork. Even worse, suppose a fork never appears in an image that does not also contain a spoon. Clearly, it would be difficult to learn which of the fork or spoon was the actual object of interest. In the language of MI learning, since every positive instance appears only in positive bags, if some negative instance also appears only in positive bags, then it is impossible to know (in general) that this negative instance is not also positive. Although still weaker assumptions might be sufficient for learnability, we argue that this condition lies close to the boundary between "easy" and "hard" instance-concept learning scenarios in that it allows us to prove positive results in a more general setting than some prior work (Blum and Kalai 1998), while still excluding scenarios used to show the hardness of general MIL, as we describe later.

To formalize this condition, we assume that there is some constant $\gamma > 0$ such that every negative instance appears with probability at least γ in some negative bag(s). Given an instance-labeling function, this constraint limits the set of valid bag-labeling functions. The full set of constraints on valid MI generative processes (MI-GEN), consisting of bag distributions and labeling functions, are formalized below:

Definition 1 (MI-GEN)**.** *Given an instance distribution* $D_{\mathcal{X}}$ *(specified by* $\Pr(x)$*), instance-labeling function* f*, and* $0 < \gamma \leq 1$, $\text{MI-GEN}(D_{\mathcal{X}}, f, \gamma)$ *is the set of all bag distribution (*$D_{\mathcal{B}}$ *with measure* $\Pr(B)$*) and bag-labeling function (*F*) pairs satisfying the following conditions:*

1. $\Pr(x) = \int_{\mathcal{B}} \Pr(x \mid B) \,\mathrm{d} \Pr(B)$

2. $\forall x, B : f(x) = 1 \wedge F(B) = 0 \implies \Pr(x \mid B) = 0$

3. $\forall x : f(x) = 0 \implies c(x) \leq 1 - \gamma$.[1]

We could think of bag-labeled instances $\{(x_i, Y_i)\}$ sampled from such a generative process as being generated from the true instance concept $f(x)$ with one-sided label noise on negative instances (since positive instances will never be observed in negative bags by the second condition, they will always have the correct label). It is also worth noting that these conditions are not too restrictive, and $\text{MI-GEN}(D_{\mathcal{X}}, f, \gamma)$ is nonempty for any choice of $D_{\mathcal{X}}$, f, and $\gamma > 0$. In particular, for any $D_{\mathcal{X}}$ and f, we can represent supervised learning within our model if $D_{\mathcal{B}}$ is a distribution over Dirac measures δ_{x_i}, each distributed according to its corresponding instance x_i, and $F(\delta_{x_i}) = f(x_i)$. This scenario falls within $\text{MI-GEN}(D_{\mathcal{X}}, f, 1) \subset \text{MI-GEN}(D_{\mathcal{X}}, f, \gamma)$ for every $0 < \gamma \leq 1$.

Given such MI generative processes, we can now precisely state what it means to probably approximately correctly (PAC) learn accurate instance concepts:

Definition 2 (MI PAC-learning)**.** *We say that an algorithm $\mathcal{A}$ MI PAC-learns instance concept class $\mathcal{F}$ from MI data when for any distribution $D_{\mathcal{X}}$ over instances, $f \in \mathcal{F}$, $\gamma > 0$, $(D_{\mathcal{B}}, F) \in \text{MI-GEN}(D_{\mathcal{X}}, f, \gamma)$, and $\epsilon, \delta > 0$, $\mathcal{A}$ requires $\mathsf{poly}(\frac{1}{\gamma}, \frac{1}{\epsilon}, \frac{1}{\delta})$ bag-labeled instances sampled independently from the MI generative process $(D_{\mathcal{B}}, F)$ to produce an instance hypothesis h whose risk measured with respect to f is at most ϵ with probability at least $1 - \delta$ over independent and identically distributed (IID) samples drawn from $D_{\mathcal{X}}$.*

Now we show that instance concepts are MI PAC-learnable in the sense of Definition 2:

Theorem 1. *An instance concept class $\mathcal{F}$ with VC dimension $\text{VC}(\mathcal{F})$ is MI PAC-learnable using $O\left(\frac{1}{\epsilon\gamma}\left(\text{VC}(\mathcal{F}) \log \frac{1}{\epsilon\gamma} + \log \frac{1}{\delta}\right)\right)$ examples.*

Proof. By Condition 1 in Definition 1, we can treat bag-labeled instances as being drawn from the underlying instance distribution $D_{\mathcal{X}}$, for any $(D_{\mathcal{B}}, F) \in \text{MI-GEN}(D_{\mathcal{X}}, f, \gamma)$. Instances are observed with some label noise with respect to true labels given by f. Since positive instances never appear in negative bags (by Condition 2 of Definition 1), noise on instances is one-sided. If every negative instance appears in negative bags at least some γ fraction of the time (by Condition 3), then the maximum one-sided noise rate is $\eta = 1 - \gamma$. Since $\gamma > 0$, $\eta < 1$, which is required for learnability. Under our generative assumptions, the noise rate might vary across instances (but is bounded by $\eta < 1$). Recent results show that under this "semi-random" noise model, when a concept class $\mathcal{F}$ has Vapnik–Chervonenkis (VC) dimension $\text{VC}(\mathcal{F})$, $\mathcal{F}$ is PAC-learnable from $O\left(\frac{1}{\epsilon(1-\eta)}\left(\text{VC}(\mathcal{F}) \log \frac{1}{\epsilon(1-\eta)} + \log \frac{1}{\delta}\right)\right)$ examples using a "minimum one-sided disagreement" strategy (Simon 2012). This strategy entails choosing a classifier that minimizes the number of disagreements on positively-labeled examples while perfectly classifying all negatively-labeled examples. This strategy also works in the case that all instances and bags are positive ($\eta = 0$, or $\gamma = 1$, since there are no negative instances). Substituting $1 - \gamma$ for η in the bound above yields the bound in terms of γ. □

We note that MI-GEN and the proof above allow for noisy positive bags without positive instances, since the additional noise is essentially absorbed into η.

Relation to Previous Work

Some of the first work on instance concept learnability in the MI setting shows that axis-parallel rectangles (APRs) are learnable from MI data under the assumption that each bag contains r instances sampled independently from *the same* product distribution (Long and Tan 1998). Blum and Kalai (1998) show that whenever an instance concept is PAC-learnable in the presence of one-sided noise, it is also learnable from MI examples, again with every bag consisting of r instances that are identically distributed across bags.

Prior work by Blum and Kalai (1998) is a special case of Theorem 1. Intuitively, we can simulate their IID r-tuple model within our model by defining each bag to be a probability distribution parameterized by an r-tuple of instances $B_{(x_1,\ldots,x_r)}$. By appropriately defining the bag distribution, we can show that the instance distribution $\Pr(x)$ as defined in our generative model is identical to the underlying instance distribution as given in their prior work. Furthermore, we can show that the effective noise rate on negative instances ($1 - \gamma$ in our generative model) is less than one so that γ is strictly positive. Thus, Blum and Kalai's generative process is contained within MI-GEN.

More recent work by Sabato and Tishby (2012) describes conditions under which *bag* concepts are learnable from MI data by bounding the sample complexity of the bag hypothesis space in terms of the sample complexity of the instance hypothesis space. One aspect of that work is that it relaxes Blum and Kalai's model to allow for non-IID instances within bags while still requiring that bags be r-tuples of instances. We can also represent this relaxed generative model in a similar way as above by again parameterizing bag-specific distributions by tuples but allowing arbitrary distributions over the resulting bags. We still require that $\gamma > 0$ for instance concept learnability, while Sabato and Tishby analyze bag concept learnability without this assumption.

One might wonder whether separate results are needed for instance and bag concept learnability. In fact, learning an accurate concept on the bag-labeling task does not necessarily translate to high accuracy on the instance-labeling task (Tragante do O, Fierens, and Blockeel 2011). Because there is only a weak relationship between the bag- and instance-labeling tasks, using risk minimization strategies for learning a bag classifier might not guarantee good performance on the instance-labeling task.

Finally, learning instance concepts from MI data is known to be hard in the general case. How then does our general framework allow for a positive result? Interestingly, it turns out that our assumption that $\gamma > 0$ precludes scenarios used to demonstrate the known hardness results. For

[1] Note that the opposite direction of Condition 3 is implied by Condition 2. Combined, the two directions of Condition 3 describe the relationship between c and f.

example, Sabato and Tishby (2012) describe the impossibility of learning instance labels from a generative process that produces only positive bags, which is essentially an unsupervised learning problem. However, in our setting, if $\gamma > 0$, then the dataset can only contain all positive bags if all instances are also positive. From the perspective of computational complexity, Auer et al. reduce learning DNF formulae to learning APRs from MI data (1998), showing that efficiently PAC-learning these MI instance concepts is impossible (unless $\mathsf{NP} = \mathsf{RP}$) when arbitrary distributions over r-tuples are allowed. Similarly, finding classifying hyperplanes for MI data has been shown to be NP-complete (Diochnos, Sloan, and Turán 2012; Kundakcioglu, Seref, and Pardalos 2010). However, all of these reductions rely on generating bags such that certain negative instances only appear in positive bags. Therefore, we conjecture that the inherent hardness of general MI learning is related to the cases when there is no $\gamma > 0$ for the generative process of MI data (though performing minimum one-sided disagreement might still be hard for certain concept classes).

Learning High-AUC Instance Concepts

The results above show that the minimum one-sided disagreement approach (Simon 2012) can be used to learn instance concepts from MI data. Below, we show that the asymmetry of this approach is not required when learning under other performance metrics. In practice, it is often sufficient to rank instances according to the likelihood of being positive. In such cases, a metric such as AUC might be a better indicator of algorithm performance than accuracy. In order to achieve good performance with respect to AUC, we show that it suffices to learn a p-concept for c in Equation 1. Importantly, this can be done with more traditional empirical risk minimization (ERM) approaches instead of minimum one-sided disagreement.

The p-concept learning model (Kearns and Schapire 1994) assumes that instead of a function $f : \mathcal{X} \to \{0, 1\}$ assigning a binary label to each instance, a p-concept $c : \mathcal{X} \to [0, 1]$ assigns a "probabilistic" label to each instance. Then for each instance x, the label 1 is observed with probability $c(x)$, and 0 is observed with probability $1 - c(x)$. A p-concept class $\mathcal{C}$ is said to be learnable with a model of probability (analogous to PAC-learnability) when for $c \in \mathcal{C}$, we can find with probability $1 - \delta$ a hypothesis h such that $\mathrm{E}\left(|h(x) - c(x)|^2\right) \leq \epsilon$, where the expectation is taken with respect to the instance distribution. Since p-concepts produce a continuous-valued label, to characterize the capacity of a p-concept hypothesis class, the *pseudo-dimension* is used (Kearns and Schapire 1994). Similar to VC dimension, pseudo-dimension characterizes a class of functions by the maximal size of a set of points that can be labeled arbitrarily above and below some threshold by a function in the class. Kearns et al. show that standard ERM approaches (i.e., choosing a hypothesis from $\mathcal{C}$ that minimizes the quadratic loss on the $\{0, 1\}$-labeled training sample) can learn p-concept classes with a bound on the number of examples in terms of the pseudo-dimension. Further, some relationships exist between the pseudo-dimension of a p-concept class and the VC dimension of concept classes obtained by thresholding to produce a $\{0, 1\}$ label. For example, for the class of hyperplanes (producing a real-valued output), the pseudo-dimension is equal to the VC dimension of the corresponding class of separating hyperplanes.

We first define the notion of learnability of an instance concept with respect to AUC rather than accuracy. The AUC of a p-concept hypothesis h is the probability that a randomly chosen negative instance will be ranked lower by h than a randomly chosen positive example. We express this as: $\mathrm{AUC}_f(h) = \Pr(h(x_-) < h(x_+) \mid f(x_-) = 0, f(x_+) = 1)$. So that this quantity is well-defined, we consider only nondegenerate cases when $\min\{p_{\mathrm{neg}}, 1 - p_{\mathrm{neg}}\} = p > 0$, where $p_{\mathrm{neg}} = \Pr(f(x) = 0)$. The "error" with respect to AUC, written $\mathrm{err}_{\mathrm{AUC}_f}(h)$, is $\Pr(h(x_-) > h(x_+) \mid f(x_-) = 0, f(x_+) = 1)$.

Definition 3 (MI AUC-learnability)**.** *For a p-concept class $\mathcal{C}$, let $\mathcal{F}_\gamma = \{\mathbb{1}\left[c(\cdot) > 1 - \gamma\right] \mid c \in \mathcal{C}\}$ be an instance concept class for some $\gamma > 0$. An algorithm $\mathcal{A}$ MI AUC-learns $\mathcal{F}_\gamma$ with $\mathcal{C}$ when for any instance distribution $D_\mathcal{X}$, $c \in \mathcal{C}$, $f(\cdot) = \mathbb{1}\left[c(\cdot) > 1 - \gamma\right]$, and $\epsilon, \delta > 0$, given $\mathsf{poly}(\frac{1}{\gamma}, \frac{1}{\epsilon}, \frac{1}{\delta})$ examples labeled by c, $\mathcal{A}$ returns a hypothesis $h \in \mathcal{C}$ such that $\mathrm{err}_{\mathrm{AUC}_f}(h) \leq \epsilon$ with probability at least $1 - \delta$ over IID samples drawn from $D_\mathcal{X}$.*

We now show that instance concepts are MI AUC-learnable using an ERM approach. Intuitively, given the p-concept for bag-labeled instances, we know that positive instances have a positive label with probability 1 and negative instances have a positive label with probability at most $1-\gamma$. Learning an *accurate* labeling function from the p-concept requires knowing a threshold based on γ to distinguish positive and negative instances. On the other hand, if we only care about *ranking* instances from the two classes, then because $1 - \gamma < 1$ even if $1 - \gamma$ is close to 1, an accurate p-concept can be used to rank instances.

Theorem 2. *Let $\mathcal{C}$ be a p-concept class with pseudo dimension $\mathrm{PD}(\mathcal{C})$ corresponding to p-concepts of bag-labeled instances, and $\mathcal{F}_\gamma$ as defined above for some $\gamma > 0$ containing the corresponding MI instance-labeling functions. Then standard ERM can MI AUC-learn $\mathcal{F}_\gamma$ using $O\left(\frac{1}{(\epsilon\gamma p)^4}\left(\mathrm{PD}(\mathcal{C}) \log \frac{1}{\epsilon\gamma p} + \log \frac{1}{\delta}\right)\right)$ examples, where $p = \min\{p_{\mathrm{neg}}, 1 - p_{\mathrm{neg}}\}$, from bag-labeled instances labeled according to some $c \in \mathcal{C}$.*

Proof. For any underlying p-concept c corresponding to bag-labeled instances, and hypothesis $h \in \mathcal{C}$, suppose that $|h(x) - c(x)| < \frac{\gamma}{2}$. Then for a positive (with respect to labels given by f) instance $x_+ \in \mathcal{X}_+$, $c(x_+) = 1$ since positive instances must appear in positive bags by Definition 1, condition 2. Thus, $|h(x_+) - c(x_+)| = 1 - h(x_+) < \frac{\gamma}{2}$, so $h(x_+) > 1 - \frac{\gamma}{2}$. For a negative instance $x_- \in \mathcal{X}_-$, if $h(x_-) \leq c(x_-)$, then $h(x_-) \leq c(x_-) \leq 1 - \gamma < 1 - \frac{\gamma}{2} < h(x_+)$. If $h(x_-) > c(x_-)$:

$$
\begin{aligned}
|h(x_-) - c(x_-)| = h(x_-) - c(x_-) &< \tfrac{\gamma}{2} \\
h(x_-) < \tfrac{\gamma}{2} + c(x_-) &\leq \tfrac{\gamma}{2} + (1 - \gamma) = 1 - \tfrac{\gamma}{2}.
\end{aligned}
$$

So when $|h(x) - c(x)| < \frac{\gamma}{2}$, $h(x_-) < h(x_+)$. Given this, AUC error can only occur when $|h(x) - c(x)| \geq \frac{\gamma}{2}$ for either x_+ or x_-:

$$\begin{aligned}
&\Pr\left(h(x_-) > h(x_+) \mid f(x_-) = 0, f(x_+) = 1\right) \\
&\leq \Pr\Big(|h(x_-) - c(x_-)| \geq \tfrac{\gamma}{2} \\
&\quad \vee |h(x_+) - c(x_+)| \geq \tfrac{\gamma}{2} \mid f(x_-) = 0, f(x_+) = 1\Big) \\
&\leq \Pr\Big(|h(x_-) - c(x_-)| \geq \tfrac{\gamma}{2} \mid f(x_-) = 0\Big) \\
&\quad + \Pr\Big(|h(x_+) - c(x_+)| \geq \tfrac{\gamma}{2} \mid f(x_+) = 1\Big) \\
&\leq \Pr\Big(|h(x_-) - c(x_-)| \geq \tfrac{\gamma}{2}\Big) / \Pr(f(x_-) = 0) \\
&\quad + \Pr\Big(|h(x_+) - c(x_+)| \geq \tfrac{\gamma}{2}\Big) / \Pr(f(x_+) = 1)
\end{aligned}$$

By Markov's inequality $\Pr\left(|h(x) - c(x)| \geq \frac{\gamma}{2}\right) \leq \frac{2\,\mathrm{E}(|h(x)-c(x)|)}{\gamma}$, and by Jensen's inequality, $\mathrm{E}\left(|h(x) - c(x)|\right) \leq \sqrt{\mathrm{E}\left(|h(x) - c(x)|^2\right)}$. If we use an ERM strategy to learn h so that $\mathrm{E}\left(|h(x) - c(x)|^2\right) \leq \epsilon'$ with probability $1 - \delta$, then we have: $\mathrm{err}_{\mathrm{AUC}_f} \leq \frac{2\sqrt{\epsilon'}}{\gamma p} + \frac{2\sqrt{\epsilon'}}{\gamma p} = \frac{4\sqrt{\epsilon'}}{\gamma p}$. Therefore, if we want $\mathrm{err}_{\mathrm{AUC}_f} \leq \epsilon$, it is sufficient to choose $\epsilon' \leq \frac{\epsilon^2\gamma^2 p^2}{16}$. Substituting this ϵ' into existing bounds (Kearns and Schapire 1994) gives us learnability of h with $O\left(\frac{1}{(\epsilon\gamma p)^4}\left(\mathrm{PD}(\mathcal{C}) \log\frac{1}{\epsilon\gamma p} + \log\frac{1}{\delta}\right)\right)$ examples. □

A surprising consequence of this result is that a naïve supervised learning strategy can be expected to learn instance concepts that have low AUC error from bag-labeled instances, given enough data.

Empirical Evaluation

Our theoretical results indicate that a supervised approach might learn concepts that perform well with respect to AUC from data generated according to our generative process. We empirically evaluate this hypothesis with experiments using both synthetic and real datasets.

For our synthetic experiments, we generate two-dimensional data from a mixture of Gaussians centered at $(0, 0)$ and $(1, 1)$ with covariance matrices both $\frac{1}{2}\mathbf{I}$, where $\mathbf{I}$ is a two-by-two identity matrix. The true instance concept is given by the hyperplane $h(\mathbf{x}) = \langle(1, 1), \mathbf{x}\rangle - 1$, with $f(\mathbf{x}) = \mathbb{1}\left[h(\mathbf{x}) > 0\right]$. Instances are drawn independently from this generative process and the labels of negative instances are flipped with probability $\eta(\mathbf{x})$. In one scenario, $\eta(\mathbf{x}) = 1 - \gamma$ for all negative instances, which corresponds to the IID r-tuple model (Blum and Kalai 1998) with the appropriate choice of r. In the other scenario, we allow the noise level $\eta(\mathbf{x})$ to vary linearly (as a function of $h(\mathbf{x})$) between $1 - \gamma$ and $1 - \sqrt{\gamma}$. Here, the noise is bounded by $\eta(\mathbf{x}) \leq 1 - \gamma$, but is less for most instances. Each case corresponds to an entire family of two-level sampling processes that are subsets of MI-GEN$(D_{\mathcal{X}}, f, \gamma)$, but for simplicity we simulate draws directly from the instance distribution labeled according to the p-concept $c(x) = \Pr(F(B) = 1 \mid x)$.

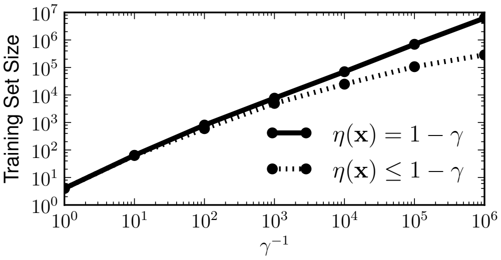


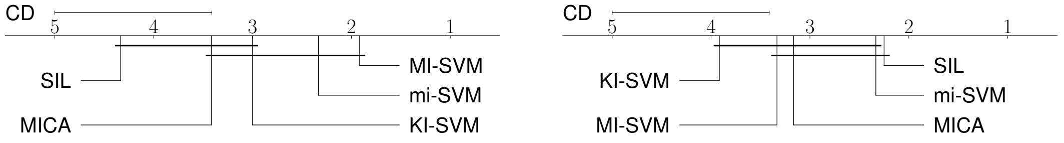
Figure 1: Training set size needed to obtain a test set performance of $\mathrm{err}_{\mathrm{AUC}} = 0.05$ with probability at least 0.5 across 100 training samples from two synthetic datasets corresponding to different two-level sampling procedures.

For each selected value of γ, binary search is used to find the smallest training set size such that AUC performance of a standard linear support vector machine (SVM) on a held-out test set of 10^4 examples is at least 0.95 across at least half of 100 independent training sets. In terms of Theorem 2, this corresponds to a fixed $\epsilon = 0.05$ and $\delta = 0.5$. Figure 1 shows a log-log plot of the number of training samples needed to accomplish this accuracy for increasingly smaller values of γ. As Theorem 2 suggests, the required number of samples appears to grow polylogarithmically with γ^{-1} in the worst case (constant noise) scenario (Pearson's $r = 0.999$ for a linear fit of the log-transformed data), and even more slowly when the noise only achieves a rate of $1 - \gamma$ on some instances.

This result is specific to AUC. If we take the classifiers learned here and measure their accuracy, we observe that they never achieve 95% accuracy when γ is sufficiently small (e.g., when $\gamma = 10^{-1}$ in the uniform noise scenario, the accuracy is always less than 50.2% as sample size increases). It is possible to learn accurate classifiers in this setting by using the algorithm described in Blum and Kalai's work for the IID case or minimum one-sided disagreement in our more general scenario. However, the counterintuitive result we observe here is that a standard SVM can learn arbitrarily well w.r.t. AUC in our setting even though the *same* classifier has poor performance w.r.t. accuracy.

To evaluate whether our theoretical results apply to real-world MI datasets, we compare a "supervised" approach to standard MI SVMs and measure both accuracy and AUC. We call the supervised approach single-instance learning (SIL). This approach applies a bag's label to each of its instances, then learns a concept from the resulting supervised dataset. Using all instances within each bag (rather than sampling a single instance from every bag) might introduce correlation in the dataset. However, we observe below that this seems to have little effect in practice.

We hypothesize that, given our theoretical results, SIL will perform well for instance-labeling under the AUC metric. To test this hypothesis, we use the Spatially Independent, Variable Area, and Lighting (SIVAL) dataset from the CBIR domain, which has been annotated with both bag and instance labels (Settles, Craven, and Ray 2008). We cre-



(a) Accuracy

(b) AUC

Figure 2: Critical difference diagrams comparing the average ranks (lower is better) of the MI SVM approaches across datasets with respect to instance-labeling accuracy and AUC. Horizontal lines connect approaches that do not differ with $\alpha = 0.1$ significance using the Nemenyi test.

ate 12 one-vs-one datasets by randomly pairing up image classes from the 25 original classes. The SIVAL images contain objects of each class taken on the same set of backgrounds. Therefore, each negative background segment appears both in positive and negative bags under the generative process corresponding to one-vs-one class pairing, so we expect that $\gamma > 0$ for this task.

We compare SIL used with a standard supervised SVM to four other MI SVM approaches: mi-SVM, MI-SVM (Andrews, Tsochantaridis, and Hofmann 2003), MICA (Mangasarian and Wild 2008), and the "instance" variant of KI-SVM (Li et al. 2009), which are specifically designed to learn instance concepts from MI data. We evaluate algorithms using 10-fold cross-validation, with 5-fold inner-validation used to select parameters using random search (Bergstra and Bengio 2012). We use the radial basis function (RBF) kernel in all cases, with scale parameter $\gamma \in [10^{-6}, 10^{1}]$, and regularization–loss trade-off parameter $C \in [10^{-2}, 10^{5}]$. We implement KI-SVM using published code[2] and the other approaches using Python with NumPy (Ascher et al. 2001) for general matrix computations, and the CVXOPT library (Dahl and Vandenberghe 2009) for optimization. We use L_2 regularization with MICA for a more direct comparison with the other approaches. We use only bag labels during training and parameter selection. We use the instance labels to evaluate the accuracy and AUC of predictions pooled across the 10 outer folds.

We compare algorithms using critical difference diagrams (Demšar 2006). We rank algorithms according to performance (with "1" being the best rank) and average the ranks across the 12 datasets. Then we use the Friedman test to reject the null hypothesis that the algorithms perform similarly at an $\alpha = 0.1$ significance level. We construct a critical difference diagram using the Nemenyi test to determine when algorithms differ with $\alpha = 0.1$ significance. We connect algorithms that are not significantly different under this test with a horizontal line. Figure 2 shows the results with ranks computed with respect to accuracy and AUC.[3]

The results show that under the accuracy metric, the performance of SIL is poor with respect to "proper" MI approaches. This may be because in some cases, γ is small and the SIL classifier will falsely classify many negative instances as positive. On the other hand, when AUC is used to measure performance, SIL is as effective as the top MI algorithms on the instance-labeling task, as suggested by our theoretical results. There is some change in the relative rankings of the other MI SVM approaches, but they remain statistically indistinguishable from each other in either case.

Early work on MIL used accuracy as a performance measure, and found the SIL approach to be inaccurate in the MI setting for labeling bags (Dietterich, Lathrop, and Lozano-Pérez 1997). As a result, subsequent studies rarely used it as a baseline when evaluating new MI techniques. However, our results suggest that SIL can learn high-AUC instance concepts, which in turn suggests that (at least for some domains) it can also learn bag concepts that are good in this sense. The fact that SIL appears to perform well at bag-labeling with respect to AUC has also been observed in prior work (Ray and Craven 2005), and we believe our work provides further support for this observation. Thus we recommend that future work should not only compare proposed MI techniques to their supervised counterparts, but also evaluate using multiple performance metrics (including AUC).

Though SIL clearly works well for the tasks we have investigated here, we conjecture that there may be other domains for which $\gamma \not> 0$, requiring stronger generative assumptions to be made for instance concept learnability. Further, although our results show that SIL works given enough data, the inductive bias of specialized MI algorithms might be useful when limited data is available. To test such conjectures, we suggest obtaining instance-labeled datasets from other domains as an important direction for MIL.

Conclusion

We have presented a new generative process for MI data that requires every negative instance to appear in a negative bag with some probability. This generative process is more general than previously studied models, included as special cases, and still excludes scenarios currently used to show hardness results for MIL. Even under our more general assumptions, we are able to show that instance concepts are PAC-learnable from MI data generated according to our model. Further, we show that supervised approaches can be employed to learn concepts with low AUC error from MI data. Our empirical results support the theory, and may partly explain previously reported behavior of supervised approaches on MI tasks. In future work, we plan to extend our analysis to learning bag concepts under our generative model, extending the work of Sabato and Tishby (2012).

[2] `http://lamda.nju.edu.cn/code_KISVM.ashx`

[3] Tables of numerical results are available at `http://engr.case.edu/doran_gary/publications.html`.

Acknowledgments

We thank the anonymous reviewers for their comments and suggestions. Gary Doran was supported by GAANN grant P200A090265 from the US Department of Education and NSF grant CNS-1035602. Soumya Ray was partially supported by CWRU award OSA110264.

References

Andrews, S.; Tsochantaridis, I.; and Hofmann, T. 2003. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems*, 561–568.

Ascher, D.; Dubois, P. F.; Hinsen, K.; Hugunin, J.; and Oliphant, T. 2001. *Numerical Python*. Lawrence Livermore National Laboratory, Livermore, CA.

Auer, P.; Long, P. M.; and Srinivasan, A. 1998. Approximating hyper-rectangles: learning and pseudorandom sets. *Journal of Computer and System Sciences* 57(3):376–388.

Babenko, B.; Verma, N.; Dollár, P.; and Belongie, S. 2011. Multiple instance learning with manifold bags. In *Proceedings of the International Conference on Machine Learning*, 81–88.

Behmardi, B.; Briggs, F.; Fern, X.; and Raich, R. 2012. Regularized joint density estimation for multi-instance learning. In *IEEE Statistical Signal Processing Workshop*, 740–743.

Bergstra, J., and Bengio, Y. 2012. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research* 13:281–305.

Blum, A., and Kalai, A. 1998. A note on learning from multiple-instance examples. *Machine Learning Journal* 30:23–29.

Dahl, J., and Vandenberghe, L. 2009. CVXOPT: A python package for convex optimization.

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research* 7:1–30.

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

Diochnos, D.; Sloan, R.; and Turán, G. 2012. On multiple-instance learning of halfspaces. *Information Processing Letters*.

Kearns, M. J., and Schapire, R. E. 1994. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences* 48(3):464–497.

Kundakcioglu, O.; Seref, O.; and Pardalos, P. 2010. Multiple instance learning via margin maximization. *Applied Numerical Mathematics* 60(4):358–369.

Li, Y.-F.; Kwok, J. T.; Tsang, I. W.; and Zhou, Z.-H. 2009. A convex method for locating regions of interest with multi-instance learning. In *Machine Learning and Knowledge Discovery in Databases*. Springer. 15–30.

Long, P., and Tan, L. 1998. PAC learning axis-aligned rectangles with respect to product distributions from multiple-instance examples. *Machine Learning* 30(1):7–21.

Mangasarian, O., and Wild, E. 2008. Multiple instance classification via successive linear programming. *Journal of Optimization Theory and Applications* 137:555–568.

Ray, S., and Craven, M. 2005. Supervised versus multiple instance learning: an empirical comparison. In *Proceedings of the 26th International Conference on Machine Learning*, 697–704.

Sabato, S., and Tishby, N. 2012. Multi-instance learning with any hypothesis class. *Journal of Machine Learning Research* 13:2999–3039.

Settles, B.; Craven, M.; and Ray, S. 2008. Multiple-instance active learning. In *Advances in Neural Information Processing Systems*, 1289–1296.

Simon, H. U. 2012. Pac-learning in the presence of one-sided classification noise. *Annals of Mathematics and Artificial Intelligence* 1–18.

Tragante do O, V.; Fierens, D.; and Blockeel, H. 2011. Instance-level accuracy versus bag-level accuracy in multi-instance learning. In *Proceedings of the 23rd Benelux Conference on Artificial Intelligence*.

Xu, X. 2003. Statistical learning in multiple instance problems. Master's thesis, The University of Waikato.