Multiple Instance Learning for Sparse Positive Bags

Razvan C. Bunescu RAZVAN@CS.UTEXAS.EDU
Raymond J. Mooney MOONEY@CS.UTEXAS.EDU
Department of Computer Sciences, University of Texas at Austin, 1 University Station C0500, TX 78712 USA

Abstract

We present a new approach to *multiple instance learning* (MIL) that is particularly effective when the positive bags are sparse (i.e. contain few positive instances). Unlike other SVM-based MIL methods, our approach more directly enforces the desired constraint that *at least one* of the instances in a positive bag is positive. Using both artificial and real-world data, we experimentally demonstrate that our approach achieves greater accuracy than state-of-the-art MIL methods when positive bags are sparse, and performs competitively when they are not. In particular, our approach is the best performing method for image region classification.

1. Introduction

In many applications of concept learning, unambiguously labeled positive and negative examples are not easily available; however, some weaker form of supervision is fairly obtainable. *Multiple instance learning* (MIL) considers a particular form of weak supervision in which the learner is given a set of *positive bags* which are sets of instances containing at least one positive example, and *negative bags* which are sets of instances all of which are negative. MIL was originally introduced to solve a problem in biochemistry (Dietterich et al., 1997); however, it it has since been applied to problems in other areas such as classifying image regions in computer vision (Zhang et al., 2002) and text categorization (Andrews et al., 2003; Ray & Craven, 2005).

A variety of MIL algorithms have been developed over the past ten years (Dietterich et al., 1997; Gartner et al., 2002; Andrews et al., 2003); however, the simplest approach is to transform the problem into a standard supervised learning problem by just labeling all instances from positive bags as positive. Despite the class noise in the resulting positive examples, this simple approach (Single Instance Learning, SIL) often achieves competitive results when compared with other more sophisticated MIL methods (Ray & Craven, 2005). However, SIL and many other MIL methods, rely on the positive bags being fairly rich in positive examples. In some MIL applications, positive bags can be quite "sparse" and contain only a small fraction of positive examples. In particular, for image-region classification, most regions do not contain the object of interest and therefore the positive bags are quite sparse (Zhang et al., 2002).

We present a new MIL method that is particularly effective when the positive bags are sparse. Like Gartner et al. (2002) and Andrews et al. (2003), we modify an SVM-based classifier; however, instead of simply using a multi-instance kernel, we change the constraints in the SVM objective function to make it more suitable for MIL. In order to enforce the desired constraint that *at least one* of the instances in a positive bag is positive, we further constrain all bag instances to be classified far away from the decision hyperplane, using the framework of transductive SVMs (Vapnik, 1995; Collobert et al., 2006). Using both artificial and real-world data, we experimentally demonstrate that our approach achieves significantly greater accuracy than state-of-the-art MIL methods when positive bags are sparse, and performs competitively when they are not. In particular, our approach is the best performing method for image region classification.

The paper is organized as follows: Section 2 gives an overview of SVM methods previously used for solving the MIL problem, followed in Section 3 by a description of transductive SVMs and their potential utility to MIL. In Section 4 we introduce an SVM approach to sparse MIL, which is augmented in Sections 5 and 6 with transductive constraints, and a global balancing constraint respectively. We then present the MIL

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007.

datasets used for evaluation, the experimental results, and end with a discussion of the results and conclusion.

2. SVM algorithms for MIL

A starting point for our work was the extensive study by Ray and Craven (2005), in which algorithms that had been specifically designed for the MIL problem were compared against their simpler, supervised (SIL) versions. The study included, for example, Diverse Density (Maron, 1998) and Multiple Instance Logistic Regression (Ray & Craven, 2005), which were evaluated against their SIL counterparts – a generative Gaussian model, and logistic regression respectively. Their experiments also included three basic SVM approaches:

■ **[SIL-SVM]** The Single Instance Learning approach to MIL transforms the MIL dataset into a standard supervised representation by applying the bag's label to all instances in the bag. A normal SVM is then trained on the resulting dataset (Figure 1).

■ **[NSK]** In the Normalized Set Kernel of Gartner et al. (2002) a bag is represented as the sum of all its instances, normalized by its 1 or 2-norm. The resulting representation is further used in training a traditional SVM (Figure 2).

■ **[STK]** The Statistic Kernel of Gartner et al. (2002) transforms every bag into a feature vector representation, in which every feature contributes two values: the maximum value and the minimum value across all instances in the bag.

The optimization formulations associated with the SIL-SVM and NSK approaches are illustrated in Figures 1 and 2 respectively. The notation is as follows:

- Let $\mathcal{X}$ be the set of bags used for training, $\mathcal{X}_p \subseteq \mathcal{X}$ the set of positive bags, and $\mathcal{X}_n \subseteq \mathcal{X}$ the set of negative bags.
- Let $\tilde{\mathcal{X}}_p = \{x | x \in X \in \mathcal{X}_p\}$ and $\tilde{\mathcal{X}}_n = \{x | x \in X \in \mathcal{X}_n\}$ be the set of instances from positive bags and negative bags, respectively.
- Let $L = L_p + L_n = |\tilde{\mathcal{X}}_p| + |\tilde{\mathcal{X}}_n|$ be the total number of instances.
- For any instance $x \in X$ from a bag $X \in \mathcal{X}$, let $\phi(x)$ be the feature vector representation of x.
- $\phi(X) = \sum_{x \in X} \phi(x)$ is the feature vector representation of bag X.
- $w\,\phi(x) + b$ is the SVM decision hyperplane, parameterized by a weight vector w and bias b.

In both formulations, the capacity control parameter C is normalized by the corresponding number of instances, so that it remains independent of the size of the dataset.

minimize:
$$\mathbf{J}(w, b, \xi) = \frac{1}{2}\|w\|^2 + \frac{C}{L} \sum_{X \in \mathcal{X}} \sum_{x \in X} \xi_x$$

subject to:
$$w\,\phi(x) + b \leq -1 + \xi_x, \;\; \forall x \in \tilde{\mathcal{X}}_n$$
$$w\,\phi(x) + b \geq +1 - \xi_x, \;\; \forall x \in \tilde{\mathcal{X}}_p \quad (*)$$
$$\xi_x \geq 0$$

Figure 1. SIL-SVM optimization problem.

minimize:
$$\mathbf{J}(w, b, \xi) = \frac{1}{2}\|w\|^2 + \frac{C}{|\mathcal{X}|} \sum_{X \in \mathcal{X}} \xi_X$$

subject to:
$$w\,\frac{\phi(X)}{|X|} + b \leq -1 + \xi_X, \;\; \forall X \in \mathcal{X}_n$$
$$w\,\frac{\phi(X)}{|X|} + b \geq +1 - \xi_X, \;\; \forall X \in \mathcal{X}_p$$
$$\xi_X \geq 0$$

Figure 2. NSK optimization problem.

By definition, all instances from negative bags are real negative instances. Therefore, a constraint can be created for every instance from a negative bag, leading to the tighter NSK formulation from Figure 3.

minimize:
$$\mathbf{J}(w, b, \xi) = \frac{1}{2}\|w\|^2 + \frac{C}{L_n} \sum_{x \in \tilde{\mathcal{X}}_n} \xi_x + \frac{C}{|\mathcal{X}_p|} \sum_{X \in \mathcal{X}_p} \xi_X$$

subject to:
$$w\,\phi(x) + b \;\; \leq -1 + \xi_x, \quad \forall x \in \tilde{\mathcal{X}}_n$$
$$w\,\frac{\phi(X)}{|X|} + b \geq +1 - \xi_X, \;\; \forall X \in \mathcal{X}_p \quad (*)$$
$$\xi_x \geq 0, \xi_X \geq 0$$

Figure 3. NSK optimization problem (tight).

Two other SVM approaches – mi-SVM and MI-SVM – were introduced by Andrews et al. (2003), and they can be seen as refinements of SIL-SVM and NSK respectively:

■ **[mi-SVM]** This is the maximum pattern margin formulation from Andrews et al. (2003). The associated

heuristic algorithm (Figure 2 in Andrews et al. (2003)) starts by training a SIL-SVM as explained above. This is followed by a relabeling of the instances in positive bags using the learned decision hyperplane. If a positive bag contains no instances labeled as positive, then the instance that gives the maximum value of the decision function for that bag is relabeled as positive. The SVM is then retrained with the new dataset, and the process of relabeling and retraining is repeated until no labels are changed.

■ **[MI-SVM]** This is the maximum bag margin formulation from Andrews et al. (2003). As in the mi-SVM above, the optimization problem is hard to solve for the global optimum, and an approximation algorithm (Figure 3 in Andrews et al. (2003)) is initialized by training an NSK. For every positive bag, the learned decision function is used to select the bag instance that gives the maximum value, and the bag representation (initially an average of all bag instances) is replaced with this instance. The SVM is then retrained with the new dataset, and the process is repeated until no bag representation is changed.

3. Transductive SVMs

By using only one inequality constraint for every positive bag, multi-instance kernels (e.g. NSK, STK) ignore individual instances from positive bags. Alternatively, instances from positive bags can be treated as unlabeled data, with the potential of further improving the generalization accuracy when used in the framework of transductive support vector machines (TSVMs) (Vapnik, 1995). TSVMs, like SVMs, learn a large margin classification hyperplane using labeled training data, but additionally force this hyperplane to be as far as possible from the unlabeled data.

One way to achieve this is by augmenting the optimization formulation with a constraint $|w\,\phi(x)+b| \geq 1-\xi_x$ on the value of the decision function for every unlabeled instance x. Recently, Collobert et al. (2006) have given an approximation method for solving the resulting non-convex optimization problem, in which the objective function is rewritten as a difference of convex functions, and then solved using the Concave Convex Procedure introduced by Yuille and Rangarajan (2002). They also noticed a potential problem with this formulation: all unlabeled examples might be classified as belonging to only one of the classes with a very large margin, especially in high dimensions and with little training data. To ensure that unlabeled examples are assigned to both classes, they further constrained the solution by introducing a *balancing constraint*, similar to (Chapelle & Zien, 2005). More exactly, if L is the labeled training data and U is the unlabeled dataset, and if $y(x) = \pm 1$ denotes the label of x, then the balancing constraint has the form shown in Equation 1 below:

$$\frac{1}{|U|}\sum_{x\in U} w\,\phi(x) + b = \frac{1}{|L|}\sum_{x\in L} y(x) \qquad (1)$$

Intuitively, the average value of the decision function on the unlabeled data is constrained to be equal to the "average label" in the unlabeled data, which, assuming unlabeled and labeled examples come from the same distribution, is equal to the average label in the labeled dataset.

In Section 4, we show that the positive bag constraint $(*)$ from the NSK formulation in Figure 3 can be interpreted as a balancing constraint, and modify it to better suit the case of sparse positive bags. Then, in Section 5, we use additional transductive constraints, so that we get closer to enforcing the desired constraint that *at least one* of the instances in a positive bag is positive.

4. An SVM approach to sparse MIL

In the SIL-SVM approach, all instances from positive bags are treated as real positive instances. The NSK approach can be construed as a relaxed version of SIL-SVM: for any positive bag, the corresponding inequality constraint $(*)$ in NSK is satisfied whenever the inequality constraints $(*)$ from the SIL-SVM formulation are satisfied. Thus, any feasible point for SIL-SVM is also a feasible point for NSK. The inequality constraint on positive bags in NSK can also be seen as a *balancing constraint*. Let $y(x) = \pm 1$ be the hidden (i.e. unknown) label of an instance x from a positive bag X. Then the inequality constraint $(*)$ in NSK can be rewritten as follows:

$$\sum_{x\in X}\frac{w\,\phi(x)+b}{|X|} \geq \sum_{x\in X}\frac{y(x)}{|X|} - \xi_X \qquad (2)$$
$$y(x) = 1,\ \forall x \in X$$

We believe that the balancing constraint above is too strong, since it implicitly assumes that all instances inside the bag are positive. This is especially problematic when the bag X is sparse in positive instances. Instead, we would like the constraint to express the requirement that *at least one* instance $\hat{x}$ from the bag is positive, as shown below:

$$\sum_{x\in X}\frac{w\,\phi(x)+b}{|X|} \geq \sum_{x\in X}\frac{y(x)}{|X|} - \xi_X \qquad (3)$$
$$y(x) = -1,\ \forall x \in X\backslash\{\hat{x}\}$$
$$y(\hat{x}) = +1$$

Replacing the inequality constraint (*) from Figure 3 with the new balancing constraint (derived from Equation 3 by summing up the hidden labels) leads to the optimization problem in Figure 4 (sMIL).

minimize:

$$\mathbf{J}(w,b,\xi) = \frac{1}{2}\|w\|^2 + \frac{C}{L_n}\sum_{x\in\tilde{\mathcal{X}}_n}\xi_x + \frac{C}{|\mathcal{X}_p|}\sum_{X\in\mathcal{X}_p}\xi_X$$

subject to:

$$\begin{aligned} & w\,\phi(x) + b \;\leq -1+\xi_x, && \forall x\in\tilde{\mathcal{X}}_n \\ & w\,\frac{\phi(X)}{|X|} + b \geq \frac{2-|X|}{|X|} - \xi_X, && \forall X\in\mathcal{X}_p \quad (*) \\ & \xi_x \geq 0, \xi_X \geq 0 \end{aligned}$$

Figure 4. Sparse MIL (sMIL).

Notice that the right hand side of the new inequality constraint $(*)$ is larger for smaller bags. Intuitively, this implies that small positive bags are more informative than large positive bags. The sMIL problem is still convex, and it can be shown that its dual formulation is obtained from the dual formulation of NSK by modifying the linear term in the objective – the dual variable corresponding to constraint $(*)$ is now multiplied by $-1+2/|X|$ (instead of multiplying it by 1). Therefore, with this modification, the QP solver used for finding the solution to NSK can also be used for sMIL.

Also notice that the three formulations – SIL-SVM, NSK and sMIL – are all equivalent when the positive bags are of size 1. However, we believe that the sMIL approach is more appropriate for cases when the bags are sparse in positive instances, or when the distribution of the positive instances inside a bag has a high variance across positive bags. SIL-SVM, and to a lesser extent NSK, assume that all instances from positive bags are actual positive instances, therefore they are expected to incur a loss in performance when applied to sparse bags. The mi-SVM and MI-SVM methods from (Andrews et al., 2003) are based on an initial labeling of the bag instances, which is acquired from applying SIL-SVM and NSK respectively. Consequently, they inherit the aforementioned drawbacks of these initial algorithms.

5. A transductive SVM approach to sparse MIL

Even though the balancing constraint $(*)$ from the sMIL formulation is closer to expressing the requirement that at least one instance from a positive bag is positive, there may be cases when all instances from a bag have negative scores, yet the bag satisfies the balancing constraint. This can happen for instance when the negative scores are very close to 0. On the other hand, if all negative instances inside a bag X were constrained to have scores less than or equal to $-1+\xi_X$, then the balancing constraint $w\,\phi(X) + b\,|X| \;\geq\; (2-|X|)(1-\xi_X)$ would guarantee that at least one instance x had a score $w\,\phi(x) + b \;\geq\; 1-\xi_X$. The upper bound $-1+\xi_X$ on the scores of negative instances inside a positive bag could be enforced by adding the transductive constraints $|w\,\phi(x)+b| \geq 1-\xi_X$ to the sMIL formulation, where ξ_X is the slack used in the balancing constraint. However, sharing the same slack between the balancing constraint and the transductive constraints leads to a mixed integer programming problem which cannot be solved efficiently using current state-of-the-art optimization tools. A more manageable problem is obtained by decoupling the transductive slacks from the slack used in the balancing constraint, as shown in Figure 5.

minimize:

$$\mathbf{J}(\cdot) = \frac{\|w\|^2}{2} + \frac{C}{L_n}\sum_{x\in\tilde{\mathcal{X}}_n}\xi_x + \frac{C^*}{L_p}\sum_{x\in\tilde{\mathcal{X}}_p}\xi_x + \frac{C}{|\mathcal{X}_p|}\sum_{X\in\mathcal{X}_p}\xi_X$$

subject to:

$$\begin{aligned} & w\,\phi(x) + b \;\leq -1+\xi_x, && \forall x\in\tilde{\mathcal{X}}_n \\ & |w\,\phi(x) + b| \geq +1-\xi_x, && \forall x\in\tilde{\mathcal{X}}_p \\ & w\,\frac{\phi(X)}{|X|} + b \geq \frac{2-|X|}{|X|} - \xi_X, && \forall X\in\mathcal{X}_p \quad (*) \\ & \xi_x \geq 0, \xi_X \geq 0 \end{aligned}$$

Figure 5. Sparse transductive MIL (stMIL).

Unfortunately, the new optimization problem is still non-convex. However, Collobert et al. (2006) have recently shown that adding independent transductive constraints to the SVM formulation leads to an objective function that can be decomposed into a convex and a concave part. This means that the non-convex problem from Figure 5 can be optimized using the Concave Convex Procedure (CCCP) (Yuille & Rangarajan, 2002). The algorithm is initialized by ignoring the transductive constraints, which corresponds to the sMIL formulation in Figure 4. At each subsequent iteration of the CCCP procedure, the concave part (caused by the transductive constraints) is replaced with its first order Taylor approximation. The resulting convex function is then minimized using QP solvers specifically designed for the SVM formulation, such as the SMO algorithm (Platt, 1999). A useful property of

the CCCP procedure is that it usually converges after only a few iterations, with results that are competitive with other transductive approaches (Collobert et al., 2006).

By incorporating the transductive constraints into the sMIL formulation, we expect that instances from positive bags are pushed far away from the decision hyperplane (i.e. they are kept out of the margin). Coupled with the balancing constraint (∗), this will make it more likely that the positive bag contains at least one instance whose score is greater than $1 - \xi_X$.

6. A balanced SVM approach to MIL

The classification accuracy of MIL methods depends on many factors, in particular on how well their assumptions match the real distribution of positive instances inside positive bags. Among the approaches mentioned in this paper, SIL-SVM and sMIL stand at the opposite ends of this spectrum. When the bags are very rich in positive instances, SIL-SVM is expected to be the best method. On the other hand, if the bags are very sparse in positive instances, then we expect sMIL to outperform SIL-SVM. If the expected density of positive instances is known, a reasonable question is whether an MIL method could use it as a balance hyperparameter, so that when the density is high the method would converge to SIL-SVM. Conversely, if the density is low, the same method would converge to sMIL. The hyperparameter could be estimated from a small set of labeled instances, or by optimization on a separate development dataset.

One way of incorporating the new balance parameter into an SVM based MIL method is to use it transductively, as in Joachims (1999). There, the expected percentage η of positive instances is inferred from a small labeled dataset L. An SVM is initially trained only on the labeled dataset, and then used to score all instances from a significantly larger unlabeled dataset U. The top $\eta|U|$ ranked instances are labeled as positive, the rest are labeled as negative, after which all the newly labeled instances are added to the initial training dataset. Then the algorithm proceeds in an iterative fashion: at each iteration, it finds a pair of instances with different labels that were classified on the wrong side of the hyperplane, switches their labels and retrains the SVM. The impact of making mistakes on instances from the unlabeled dataset is also gradually increased, starting from a very small number, until it reaches a predefined capacity parameter. Applied on top of sMIL, this approach would have the advantage of incorporating both the transductive constraint (alternatively covered by stMIL) and the global balancing constraint controlled by η. However, experiments with standard MIL datasets have revealed that the tight coupling between the balancing constraint and transduction often results in decreased performance, a behavior that is also consistent with the results obtained using the transductive stMIL, as described in Section 8. Instead, we decided to ignore the transductive constraints and used only the initialization part of the TSVM algorithm by Joachims (1999), as illustrated in Figure 6.

Input:
- training bags $\mathcal{X}_n$ and $\mathcal{X}_p$
- feature representation $\phi(x)$
- capacity parameter C from sMIL
- balance parameter $\eta \in (0, 1]$

Output:
- decision function $f(x) = w\,\phi(x) + b$

Procedure:
- $(w, b) = \text{solve\_sMIL}(\mathcal{X}_n, \mathcal{X}_p, \phi, C)$
- order instances $x \in \tilde{\mathcal{X}}_p$ using $f(x)$
- label instances in $\tilde{\mathcal{X}}_p$:
  - the top $\eta|\tilde{\mathcal{X}}_p|$ as positive
  - the rest $(1 - \eta)|\tilde{\mathcal{X}}_p|$ as negative
- $(w, b) = \text{solve\_SIL}(\tilde{\mathcal{X}}_n, \tilde{\mathcal{X}}_p, \phi, C)$
- return (w, b)

Figure 6. Balanced MIL (sbMIL).

7. MIL Datasets

In order to evaluate the behavior of the various SVM methods for MIL in conditions of maximal sparsity, we use the procedure shown in Figure 7 to create an artificial dataset where each positive bag contains only one positive instance. The actual supervised dataset $\mathcal{D}$ is based on the AIMed corpus of human protein-protein interactions, which has been used before by Bunescu and Mooney (2006) in conjunction with a relation extraction kernel. It consists of 225 Medline abstracts, of which 200 are known to describe interactions between human proteins, while the other 25 do not refer to any interaction. There are around 4,000 protein references and 1,000 tagged interactions in this dataset. Many protein pairs co-occur multiple times in the corpus, leading naturally to bags of candidate interaction pairs. Considering only bags of size two or more, we set the parameter ρ to be equal to the ratio of negative to positive bags in this corpus. The bag instances were provided as input to the algorithm in Figure 7 to create the maximally sparse dataset, henceforth referred to as AIMed.

Input:
- a traditional supervised dataset $\mathcal{D} = \mathcal{D}_p \cup \mathcal{D}_n$
- the minimum bag size S_{min}
- the maximum bag size S_{max}
- the bag class ratio $\rho = |\mathcal{X}_n|/|\mathcal{X}_p|$

Output:
- an MIL dataset $\mathcal{X} = \mathcal{X}_p \cup \mathcal{X}_n$

Procedure:
- ▷ initialize $\mathcal{X}_p = \emptyset$ and $\mathcal{X}_n = \emptyset$
- ▷ for every positive instance $x \in \mathcal{D}_p$:
  - ▷ create a positive bag $X = \{x\}$
  - ▷ pick a random size S between S_{min} and S_{max}
  - ▷ fill bag X with $S-1$ negative instances randomly sampled from $\mathcal{D}_n$
  - ▷ update $\mathcal{X}_p \leftarrow \mathcal{X}_p \cup \{X\}$
- ▷ repeat until $|\mathcal{X}_n| = \rho|\mathcal{X}_p|$:
  - ▷ pick a random size S between S_{min} and S_{max}
  - ▷ create a negative bag X by randomly sampling S negative instances from $\mathcal{D}_n$
  - ▷ update $\mathcal{X}_n \leftarrow \mathcal{X}_n \cup \{X\}$
- ▷ return $\mathcal{X} = \mathcal{X}_p \cup \mathcal{X}_n$

Figure 7. Creation of maximally sparse dataset.

The MIL datasets used in the experimental evaluation are as follows:

■ **[AIMed]** This is the maximally sparse dataset created from a corpus of protein-protein interactions using the algorithm from Figure 7. The resulting dataset has 670 positive and 1,040 negative bags. We also include a smaller dataset AIMED$\frac{1}{2}$, created by using around half of the positive and negative bags from AIMED.

■ **[CBIR]** In the Content Based Image Retrieval (CBIR) domain, the task is to categorize images as to whether they contain an object of interest. An image is represented as a bag of image regions that are characterized by color, texture and shape descriptors. The underlying assumption is that the object of interest is contained in at least one region. In the experimental evaluation we use the TIGER, ELEPHANT and FOX datasets from (Andrews et al., 2003), the task being to separate the three types of animals from background images. Each dataset contains 100 positive and 100 negative example images. We expect that most positive images contain only one instance of the target object. Since the number of regions varies widely from one image to another, this means that the density of positive instances inside a bag has a high variance across the bags. Therefore, we believe sMIL to be the most appropriate SVM method for this task.

■ **[MUSK]** This is the original dataset used in the drug activity prediction task from (Dietterich et al., 1997). The bags correspond to molecules, while bag instances correspond to three dimensional conformations of the same molecule. For positive bags, it is assumed that at least one of these low energy conformations binds to a predefined target. Equivalently, for the MUSK dataset, a bag is considered positive if the molecule smells "musky". We conduct experimental evaluation on both versions of the dataset: MUSK1, with an average bag size of 6, and MUSK2, with an average bag size of 60.

■ **[TST]** Andrews et al. (2003) have built a text categorization dataset in which MEDLINE articles are represented as bags of overlapping text passages. Every document in the MEDLINE corpus comes annotated with a set of MeSH terms, each representing a binary category. Given a category, a bag is considered positive if the corresponding document belongs to that category. In the experimental evaluation, we use the first two datasets, TST1 and TST2.

8. Experimental Evaluation

We evaluated the following SVM approaches to MIL:

- the supervised SVM in Figure 1 (SIL-SVM);
- the normalized set kernel in Figure 3 (NSK);
- the statistic kernel (STK);
- the sparse MIL approach in Figure 4 (sMIL);
- the balanced MIL approach initialized with sMIL, as in Figure 6 (sbMIL);
- the sparse MIL approach augmented with transductive constraints in Figure 5 (stMIL).

All methods were implemented by modifying Fabian Sinz' UNIVERSVM[1] package to appropriately reflect the corresponding optimization formulations. The parameters were set to their default values. Also, for stMIL, since every positive bag has its own balancing constraint, there is no need to use the global balancing constraint (Equation 1). Ray and Craven (2005) observed that the quadratic kernel generally results in more accurate models, therefore we used a quadratic kernel with all datasets, except for AIMED where we used the subsequence kernel approach from (Bunescu & Mooney, 2006).

[1] http://www.kyb.mpg.de/bs/people/fabee/universvm.html

Table 1. Average area under ROC curve for each SVM method on each dataset.

Dataset	**SIL-SVM**	**NSK**	**STK**	**sMIL**	**sbMIL**	**stMIL**
AIMED	57.44	87.11	N/A	87.19	87.99	**92.11**
AIMED$\frac{1}{2}$	45.86	54.06	N/A	54.08	67.66	**72.94**
TIGER	76.65	79.07	80.80	81.12	**82.95**	74.48
ELEPHANT	85.08	82.94	85.22	87.98	**88.58**	81.64
FOX	52.72	64.01	62.14	66.13	**69.78**	60.67
MUSK1	87.82	85.61	69.44	86.91	**91.78**	79.46
MUSK2	87.33	**90.78**	61.01	81.19	87.74	68.41
TST1	96.25	97.16	96.19	97.29	**97.41**	96.81
TST2	85.37	**90.60**	86.87	87.97	90.57	88.55

We test the algorithms using 10-fold cross validation. For each fold, the test performance is summarized using the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve measures the true-positive rate versus the false-positive rate of a classifier as a threshold is varied across a measure of confidence in its predictions. The prediction confidence for each bag is computed as the maximum over the prediction confidence of each instance in the bag. At instance level, the confidence is set to the value of the decision function $f(x) = w\,\phi(x) + b$ on that instance (i.e. the signed margin). For each method, we report the area under the ROC curve averaged over the ten folds.

For lack of labeled instances, the density parameter η from sbMIL in Figure 6 is estimated as follows:

1. Each outer training fold is split into 9 inner folds; the sbMIL is trained on 8 inner folds and tested on the remaining inner fold with η set to ten different values between 0 and 1, using an increment of 0.1.

2. The procedure above is repeated with each of the 9 inner folds used as a test dataset, while the other 8 are used for training.

3. For each η value we compute the area under the ROC curve averaged over the 9 test inner folds; the η value with the maximum average ROC area is then associated with the outer training fold.

The overall experimental results are shown in Table 1. The STK method cannot be applied to AIMED since it would require an explicit bag representation where each feature corresponded to a subsequence of words observed in one of the bag instances. This representation is not feasible in terms of space.

9. Discussion of Results

On the maximally sparse datasets AIMED and AIMED$\frac{1}{2}$, the sMIL approach gives the same performance as NSK. Further augmenting sMIL with transductive constraints leads to a statistically significant increase in accuracy (paired t-test with $p < 0.01$), as illustrated by the last column (stMIL) in Table 1 . However, this is in stark contrast to the rest of the datasets, for which adding transduction to sMIL consistently hurts accuracy. A possible explanation for the non-utility of using transduction with the real-world MIL datasets may come from the observation made by Ray and Craven (2005): "the nature of the negative instances in the positive bags may be different from the nature of the negative instances in the negative bags". Consequently, if negative instances in the positive bags come from a different distribution than negative instances in the negative bags, then treating them the same and forcing them out of the margin may not be the most appropriate thing to do. This would also explain why transduction is helping for the AIMED datasets: there, the negative instances in the positive bags are sampled from the same set as the negative instances in the negative bags. It is also possible that, in some of the real-world datasets used in our experiments (e.g. MUSK), the negative and positive instances in the positive bags are very similar. The behavior is also consistent with a set of similar results that we obtained when applying the transductive approach from (Joachims, 1999) on top of sMIL.

As expected, in all three image datasets, sMIL is performing better than the previous SVM approaches, and the performance is further improved by using the global balancing constraint in sbMIL. The differences between sMIL/sbMIL and the best performing SVM method for each image dataset are not statistically significant. However, sMIL has a more consistent per-

formance on all of these datasets – when compared only among themselves, none of the first three SVM approaches (SIL-SVM, NSK and STK) is better than the others on all image datasets, therefore there is no clear second to sMIL.

The last section of Table 1 shows that sMIL and its balanced version sbMIL also perform competitively on the MUSK and TST datasets, for which the bags are expected to be significantly less sparse than in the image datasets. Therefore, given that the three initial SVM methods already compare well with other non-SVM approaches to MIL (Ray & Craven, 2005), we can infer with high confidence that sMIL and its balanced version sbMIL offer a very competitive alternative, especially for datasets with sparse positive bags.

10. Future Work and Conclusion

One property of the MIL problem that is not captured by the SVM methods explored in this paper is the fact that, for many real-world datasets, instances belonging to the same bag are, in general, more similar than instances belonging to different bags. Modeling this type of distribution imbalance may lead to further improvements in accuracy. Kuck and de Freitas (2005) use the number of annotations associated with each image to estimate the fraction of positively-labeled instances per bag. Such estimates could be exploited in our SVM formulations too, by incorporating them in the bag-level balancing constraints.

We have presented a new SVM approach to multiple instance learning that is particularly effective when the positive bags are sparse in positive instances. Our approach more directly enforces the constraint that *at least one* instance in a positive bag is positive. Experimental results demonstrate that the new approach outperforms previous SVM methods on image datasets, and performs competitively on other types of MIL data. We have also shown that treating instances from positive bags as unlabeled data in a transductive setting leads to a significant increase in accuracy in the case when negative instances in positive bags come from the same distribution as negative instances in negative bags.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful remarks. This work was supported by grant IIS-0325116 from the NSF, and a gift from Google Inc. The experiments were run on the Mastodon cluster, provided by NSF grant EIA-0303609.

References

Andrews, S., Tsochantaridis, I., & Hofmann, T. (2003). Support vector machines for multiple-instance learning. *Advances in Neural Information Processing Systems 15* (pp. 561–568). Vancouver, BC: MIT Press.

Bunescu, R. C., & Mooney, R. J. (2006). Subsequence kernels for relation extraction. *Advances in Neural Information Processing Systems 18*. Vancouver, BC.

Chapelle, O., & Zien, A. (2005). Semi-supervised classification by low density separation. *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*.

Collobert, R., Sinz, F., Weston, J., & Bottou, L. (2006). Large scale transductive SVMs. *Journal of Machine Learning Research*, *7(Aug)*, 1687–1712.

Dietterich, T. G., Lathrop, R. H., & Lozano-Perez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, *89*, 31–71.

Gartner, T., Flach, P., Kowalczyk, A., & Smola, A. (2002). Multi-instance kernels. *In Proceedings of the 19th International Conference on Machine Learning* (pp. 179–186). Sydney, Australia: Morgan Kaufmann.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. *Proceedings of the Sixteenth International Conference on Machine Learning (ICML-99)* (pp. 200–209). Bled, Slovenia.

Kuck, H., & de Freitas, N. (2005). Learning about individuals from group statistics. *Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)* (p. 332). Edinburgh, Scotland.

Maron, O. (1998). *Learning from ambiguity* (Technical Report Doctoral dissertation). Dept. of EECS, MIT, Cambridge, MA.

Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization. In B. Scholkopf, C. J. C. Burges and A. J. Smola (Eds.), *Advances in kernel methods - support vector learning*, 185–208. Cambridge, MA: MIT Press.

Ray, S., & Craven, M. (2005). Supervised versus multiple instance learning: An empirical comparison. *Proceedings of 22nd International Conference on Machine Learning (ICML-2005)* (pp. 697–704). Bonn, Germany.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Berlin: Springer-Verlag.

Yuille, A. L., & Rangarajan, A. (2002). The concave-convex procedure (CCCP). *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press.

Zhang, Q., Goldman, S. A., Yu, W., & Fritts, J. (2002). Content-based image retrieval using multiple-instance learning. *Proceedings of 19th International Conference on Machine Learning (ICML-2002)* (pp. 682–689).