Ellipsoidal Multiple Instance Learning

Gabriel Krummenacher GABRIEL.KRUMMENACHER@INF.ETHZ.CH
Department of Computer Science, ETH Zurich, Switzerland

Cheng Soon Ong CHENGSOON.ONG@UNIMELB.COM.AU
NICTA, Victoria Research Laboratory, Melbourne, Australia

Joachim M. Buhmann JBUHMANN@INF.ETHZ.CH
Department of Computer Science, ETH Zurich, Switzerland

Abstract

We propose a large margin method for asymmetric learning with ellipsoids, called eMIL, suited to multiple instance learning (MIL). We derive the distance between ellipsoids and the hyperplane, generalising the standard support vector machine. Negative bags in MIL contain only negative instances, and we treat them akin to uncertain observations in the robust optimisation framework. However, our method allows positive bags to cross the margin, since it is not known which instances within are positive.

We show that representing bags as ellipsoids under the introduced distance is the most robust solution when treating a bag as a random variable with finite mean and covariance. Two algorithms are derived to solve the resulting non-convex optimization problem: a concave-convex procedure and a quasi-Newton method. Our method achieves competitive results on benchmark datasets. We introduce a MIL dataset from a real world application of detecting wheel defects from multiple partial observations, and show that eMIL outperforms competing approaches.

1. Introduction

In many applications of supervised learning, the cost of obtaining ground truth labels is a significant bottleneck. This has led to research on weakly labeled data, among which the framework of multiple instance learning (MIL) has shown promising results. In parallel, there has been developments in robust optimization, where data uncertainty is taken into account. Motivated by a real world application of defect detection based on multiple partial observations, we propose a novel approach based on both MIL and robust optimization. In this paper, we consider the binary classification problem (labels $y \in \{-1, +1\}$) in the MIL setting (Dietterich et al., 1997). Instead of having one label per example $\mathbf{x}_j$, we are given B bags of examples where each bag $\{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ij}, \ldots, \mathbf{x}_{in_i}\}_{i=1}^{B}$ consists of n_i instances. Unlike standard supervised learning, labels are provided only at the bag level such that $y_i = +1$ if at least one of $y_{i1}, \ldots, y_{in_i}$ is positive, and the bag is negative ($y_i = -1$) only if all $y_{i1} = \ldots = y_{in_i} = -1$. Unfortunately, due to the weak labeling, it is unclear during training time how to allocate the positive label. Since any number of examples in a positive bag may be positive, one would naively have to look at all possible labelings that include at least one positive label. This results in potentially expensive computations involving the solution of a combinatorial optimization problem. See Kim & la Torre (2010) for an overview of recent MIL methods.

We propose ellipsoidal multiple instance learning (eMIL) where we relax the function over the set of instance labels and approximate a bag by the first and second moment of the empirical distribution. I.e., we take the arithmetic mean and empirical covariance matrix of the within bag instances. Neglecting higher order moments gets rid of the combinatorial optimisation problem and enforces regularisation.

Our proposed method (eMIL) results in a modular two stage algorithm: (1) estimate the ellipsoids, and (2) optimise the generalised large margin algorithm. An additional benefit is that our approach gives an instance level classifier (instead of a bag level classi-

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

fier), which may be important in some applications. We first derive the optimization problem to find a maximum-margin type classifier for ellipsoids, where one class of ellipsoids can overlap with the decision boundary (Section 2). Two different ways of scaling the empirical covariance matrix for different distributional assumptions on the bags are presented in Section 2.3 and Section 2.4. We show in Section 2.4 that solving this optimisation problem is equivalent to treating each bag as a random variable and robustly maximizing the margin between instances distributed according to this random variable under asymmetric probabilistic constraints over all distributions with finite mean and covariance. To solve the resulting nonconvex optimisation problem a quasi Newton method and a decomposition of the objective into the difference of two convex functions is presented in Section 3. The method compares favourably to state of the art MIL methods with respect to accuracy on benchmark datasets (Section 4). Finally we introduce our motivating application: a safety critical real world problem of detecting wheel defects, and show that eMIL has better accuracy than recent methods (Section 5).

2. Detection with ellipsoids

We derive a maximum-margin type classifier for the problem of learning with positive and negative ellipsoidal examples. Our aim is to exploit the structure of a bag in the MIL setting and not just treat the instances as individual separate points. We capture this bag structure by the empirical mean and the empirical covariance matrix of all the instances in a bag. This naturally leads to the interpretation of a bag as an ellipsoid. The notion that a positive bag label only guarantees one instance to be positive, is represented by letting ellipsoids with positive label overlap the negative half space. Negatively labeled ellipsoids on the other hand are required to be maximally distant from the decision surface, since we know that all instances in a bag with negative label are indeed negative.

Recall that an ellipsoid in $\mathbf{x} \in \mathbb{R}^d$ is given by a positive semidefinite covariance matrix $\mathbf{P} \in \mathbb{S}^d_+$, and a central vector $\mathbf{q} \in \mathbb{R}^d$ by

$$(\mathbf{x}-\mathbf{q})^\top \mathbf{P}^{-1}(\mathbf{x}-\mathbf{q}) = 1. \tag{1}$$

Given a set of examples and corresponding labels $\{(\mathbf{P}_i, \mathbf{q}_i), y_i\}_{i=1}^B$, we would like to find a linear separating hyperplane with $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$

$$\mathbf{w}^\top \mathbf{x} + b = 0 \tag{2}$$

which follows the maximum margin principle. Therefore for the predictor based on ellipsoid i given by $f(\mathbf{P}_i; \mathbf{q}_i)$, we would like to solve the following regularized empirical risk problem

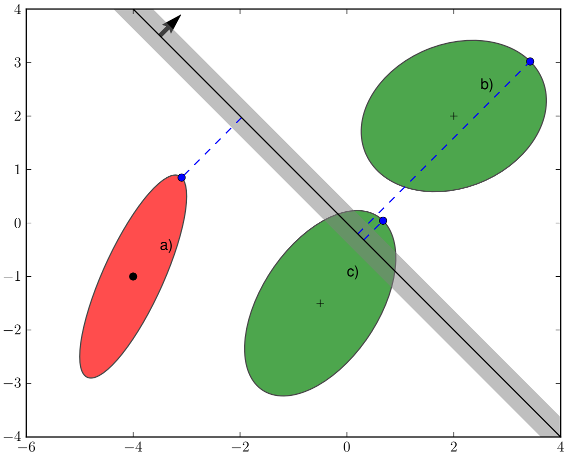


Figure 1. Derivation of the prediction function. The following three cases are distinguished: (a) ellipsoid is fully in the negative-half space, then the distance is the minimal distance of a point on the ellipsoid to the hyperplane; (b) ellipsoid is fully in the positive half-space, then the distance is the maximal distance from the hyperplane and (c) ellipsoid intersects the hyperplane, then the distance is to the point maximally in the positive half-space. This means that for all three cases the distance is always to the point on the ellipse, that is maximally in the direction of the hyperplane normal vector $\mathbf{w}$ and the sign is the sign of the half-space where that point lies in.

$$\min_{\mathbf{w},b} \sum_{i=1}^{B} \ell(y_i f(\mathbf{P}_i; \mathbf{q}_i)) + \frac{\lambda}{2}\|\mathbf{w}\|^2 \tag{3}$$

where λ is the regularisation parameter and $\ell(t) = \max(0, 1-t)$ is the hinge loss.

2.1. Optimisation problem

The prediction function $f(\mathbf{P}_i, \mathbf{q}_i)$ should give the signed distance of the ellipsoid to the hyperplane. By reasoning about the geometry of the problem (refer to Figure 1), we get the following distance for any ellipsoid to the hyperplane.

Proposition 1. *Given an ellipsoid, Equation (1), and a hyperplane, Equation (2), and taking the asymmetry of positive and negative ellipsoids into account, the signed distance from the ellipsoid to the hyperplane is given by*

$$\frac{1}{\|\mathbf{w}\|}\left(\sqrt{\mathbf{w}^\top \mathbf{P} \mathbf{w}} + \mathbf{w}^\top \mathbf{q} + b\right). \tag{4}$$

See Appendix C in the supplementary file for the full derivation of this distance. Therefore the prediction function is given by

$$f(\mathbf{P};\mathbf{q}) = \sqrt{\mathbf{w}^\top \mathbf{P}\mathbf{w}} + \mathbf{w}^\top \mathbf{q} + b. \qquad (5)$$

Substituting Equation (5) into Equation (3) we obtain the following optimisation problem, which we call ellipsoidal multiple instance learning (eMIL).

$$\min_{\mathbf{w},b} \sum_{i=1}^{B} \ell\left(y_i\left(\sqrt{\mathbf{w}^\top \mathbf{P}_i\mathbf{w}} + \mathbf{w}^\top \mathbf{q}_i + b\right)\right) + \frac{\lambda}{2}\|\mathbf{w}\|^2. \qquad (6)$$

Equation (6) is subtly different from robust optimization problems. We discuss this further in Section 2.5. Note that the optimization problem given by Equation (6) is non-convex. This is due to the term $-\sqrt{\mathbf{w}^\top \mathbf{P}_i\mathbf{w}}$ in the hinge loss for positive bags $\left(\max\left(0, 1 - \left(\sqrt{\mathbf{w}^\top \mathbf{P}_i\mathbf{w}} + \mathbf{w}^\top \mathbf{q}_i + b\right)\right)\right)$, which is a concave function in $\mathbf{w}$. It can also be observed that the problem is not a second order cone program by decomposing the hinge loss:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{B} \xi_i \\ \text{s.t.} \quad & \|\mathbf{A}_i\mathbf{w}\| \geq -\xi_i - \mathbf{w}^\top \mathbf{q}_i - b + 1, \forall i\colon y_i = +1 \\ & \|\mathbf{A}_i\mathbf{w}\| \leq \xi_i - \mathbf{w}^\top \mathbf{q}_i - b - 1, \quad \forall i\colon y_i = -1 \\ & \mathbf{0} \leq \boldsymbol{\xi}. \end{aligned} \qquad (7)$$

Where we have used $\mathbf{P}_i = \mathbf{A}_i^\top \mathbf{A}_i$. The constraints for positive ellipsoids are not second order cone constraints.

2.2. Ellipsoid estimation

We model the ith bag, $\{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}\}$ with n_i instances, by the empirical mean and covariance of the instances, given by

$$\mathbf{q}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{x}_{ij} \quad \mathbf{P}_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \mathbf{q}_{ij})(\mathbf{x}_{ij} - \mathbf{q}_{ij})^\top. \qquad (8)$$

When the number of instances per bag n_i is larger than the dimensionality of the feature space d, the covariance $\mathbf{P}_i$ is of full rank and strictly positive definite. However, in many datasets $n_i < d$, resulting in a low rank $\mathbf{P}_i$. This is usually the case in the test datasets we consider. The average number of instances in a bag is much lower than d, resulting in a semidefinite covariance $\mathbf{P}_i$.

The covariance matrix gives the shape of the ellipsoid. To find the volume, we derive two types of scaling factors for the covariance matrix under two different distributional assumptions in the following sections.

2.3. Confidence regions

Under the assumption of approximately Gaussian distributed instances per bag we can use the following fact. Recall that for a random variable $\mathbf{x}$ distributed as a p dimensional Gaussian $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the quadratic form $(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mu)$ is distributed as χ^2 with p degrees of freedom. This implies that the ellipsoid

$$(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \leqslant \chi_p^2(\alpha)$$

contains $1 - \alpha$ of the total probability mass.

For an ellipsoid $(\mathbf{q}, \mathbf{P})$ to cover $1 - \alpha$-percent of a p-dimensional multivariate Gaussian distribution with a covariance matrix $\boldsymbol{\Sigma}$ we set $\mathbf{P} = \boldsymbol{\Sigma} \cdot F_{\chi_p^2}^{-1}(1 - \alpha)$, where $F_{\chi_p^2}^{-1}(q)$ is the quantile function (inverse cdf) of χ_p^2 at quantile q. We can use this fact to scale the empirical covariance matrix, estimated from the bag instances. In the next section a more general scaling factor for non Gaussian distributions is derived.

2.4. Probabilistic multiple instance learning

In this section we show that eMIL maximizes the margin between instances from any within bag data distribution with finite mean and covariance with high probability, while enforcing the asymmetry inherent to multiple instance learning. We use similar minimax techniques to robust optimization approaches (Lanckriet et al., 2002; Shivaswamy et al., 2006), that are based on a multivariate Chebyshev's inequality (Bertsimas & Popescu, 2001).

Assuming the instances in a bag are drawn from a probability distribution with mean $\mathbf{q}_i$ and covariance $\boldsymbol{\Sigma}_i$, we want to find a hyperplane that maximises the margin between instances from the two classes. Remember, that for instances in a negative bag we know the label, for instances in a positive bag we only know at least one instance is positive. We formalize this with the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{B} \xi_i \\ \text{s.t.} \quad & \inf_{\mathbf{x}_i \sim (\mathbf{q}_i, \boldsymbol{\Sigma}_i)} \Pr_{\mathbf{x}_i}[y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i] \geq 1 - \alpha_i, \forall y_- \\ & \sup_{\mathbf{x}_i \sim (\mathbf{q}_i, \boldsymbol{\Sigma}_i)} \Pr_{\mathbf{x}_i}[y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \leq 1 - \xi_i] \geq \alpha_i, \qquad \forall y_+ \\ & \mathbf{0} \leq \boldsymbol{\xi}, \end{aligned} \qquad (9)$$

where $\sup / \inf_{\mathbf{x}_i \sim (\mathbf{q}_i, \boldsymbol{\Sigma}_i)}$ means supremum/infimum over all distributions for $\mathbf{x}_i$ having mean $\mathbf{q}_i$ and covariance $\boldsymbol{\Sigma}_i$. The constraints in Equation (9) differ for instances from positive bags and instances for negative bags. For instances from negative bags we want the smallest (over all distributions with mean $\mathbf{q}_i$ and covariance $\mathbf{\Sigma}_i$) probability of correct classification to be higher than $\alpha_i \in (0, 1]$. For small α_i this gives high worst-case probability of correct classification for all instances in a negative bag. For instances from positive bags the highest (again over all distributions with mean $\mathbf{q}_i$ and covariance $\mathbf{\Sigma}_i$) probability of negative classification needs to be larger than α_i, in other words, there exists a distribution with negative classification higher than α_i. This may seem counter-intuitive at first, but recall that for a positive bag to be classified correctly only one instance needs to be classified correctly, i.e. many of the instances will be classified negative. Here for small α_i this gives low negative classification probability.

We show that considering the worst case distribution with finite mean and covariance results in the same constraints as making an assumption of ellipsoidal bags.

Proposition 2. *The optimization problem in Equation (9) is equivalent to eMIL (Equation (7)).*

To prove Proposition 2 we use the following Lemmas:

Lemma 3.

$$\sup_{\mathbf{x}_i \sim (\mathbf{q}_i, \mathbf{\Sigma}_i)} \Pr_{\mathbf{x}_i}[y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \leq 1 - \xi_i] = \frac{1}{1 + d^2}, \text{with} \tag{10}$$

$$d^2 = \inf_{\mathbf{x}_i | y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \leq 1 - \xi_i} (\mathbf{x}_i - \mathbf{q}_i)^\top \mathbf{\Sigma}_i^{-1} (\mathbf{x}_i - \mathbf{q}_i). \tag{11}$$

Where $\mathbf{x}_i$ is a random vector and the supremum is over all distributions for $\mathbf{x}_i$ with mean $\mathbf{q}_i$ and covariance matrix $\mathbf{\Sigma}_i$.

Proof sketch. This can be shown by using the multivariate Chebyshev inequality from Lanckriet et al. (2002) and setting $\mathcal{S} = \{\mathbf{x}_i | y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \leq 1 - \xi_i\}$. (See Appendix A in the supplementary file for a more detailed proof.) □

Lemma 4. *For $\mathbf{x}_i \sim (\mathbf{q}_i, \mathbf{\Sigma}_i)$:*

$$\begin{aligned} &\inf_{\mathbf{x}_i | y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \leq 1 - \xi_i} (\mathbf{x}_i - \mathbf{q}_i)^\top \mathbf{\Sigma}_i^{-1} (\mathbf{x}_i - \mathbf{q}_i) \\ &= \frac{\max(0, y_i(\langle \mathbf{w}, \mathbf{q}_i \rangle + b) - 1 + \xi_i)^2}{\mathbf{w}^\top \mathbf{\Sigma}_i \mathbf{w}}. \end{aligned} \tag{12}$$

Proof sketch. By considering the distance d to the hyperplane, we obtain the above result. This follows the same logic as the proof in Lanckriet et al. (2002); Shivaswamy et al. (2006), see Appendix A in the supplementary file for a more detailed proof. □

Proof of Proposition 2. Since the objective function in both optimization problems are the same, it is sufficient to show that the constraints are equivalent.

For the probabilistic constraint of negative bags

$$\inf_{\mathbf{x}_i \sim (\mathbf{q}_i, \mathbf{\Sigma}_i)} \Pr_{\mathbf{x}_i}[y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i] \geq \alpha_i, \tag{13}$$

we rewrite it as

$$\sup_{\mathbf{x}_i \sim (\mathbf{q}_i, \mathbf{\Sigma}_i)} \Pr[y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \leq 1 - \xi_i] \leq 1 - \alpha_i. \tag{14}$$

Then we can use Lemma 3 to rewrite the constraints as

$$\begin{aligned} \frac{1}{1 + d^2} &\leq \alpha_i, \forall i \colon y_i = -1 \\ \frac{1}{1 + d^2} &\geq \alpha_i, \forall i \colon y_i = +1, \end{aligned} \tag{15}$$

where d^2 is defined as in Lemma (3).

Next we use Lemma 4 and rearrange the terms to finally get the constraints

$$\begin{aligned} -\langle \mathbf{w}, \mathbf{q}_i \rangle - b &\geq 1 - \xi_i + \kappa(\alpha_i)\sqrt{\mathbf{w}^\top \mathbf{\Sigma}_i \mathbf{w}}, \ \forall y_- \\ \langle \mathbf{w}, \mathbf{q}_i \rangle + b &\geq 1 - \xi_i - \kappa(\alpha_i)\sqrt{\mathbf{w}^\top \mathbf{\Sigma}_i \mathbf{w}}, \ \forall y_+, \end{aligned} \tag{16}$$

where $\kappa(\alpha_i) = \sqrt{\frac{1-\alpha_i}{\alpha_i}}$

Now, if $\mathbf{A}_i$ in Equation (7) is set to $\mathbf{A}_i = \kappa(\alpha_i)\mathbf{\Sigma}_i^{1/2}$ the equivalence can be seen. □

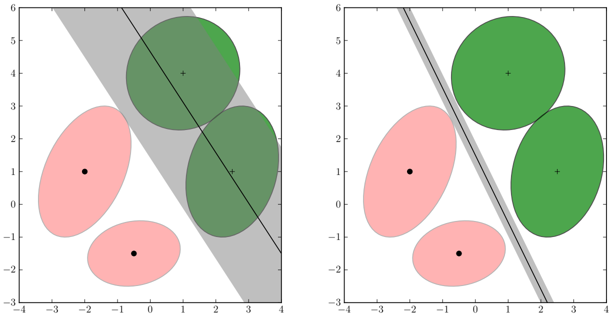
This shows that by considering the worst case distribution (not necessarily Gaussian) with finite mean and covariance scales the ellipsoid by a factor $\kappa(\alpha_i)^2$.

2.5. Relation to robust classification

In robust classification, the goal is to find a classifier, that is robust to random perturbations in the feature space (Ben-Tal et al., 2009). If we assume an ellipsoidal uncertainty set $\mathcal{U}_i = \{\mathbf{u}_i : \mathbf{u}_i^\top \mathbf{P}_i \mathbf{u}_i \leq 1\}_{i=1}^n$, and seek to find a maximum-margin classifier, we get the formulation of a robust SVM (Sra et al., 2011):

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C \cdot \sum_{i=1}^{N} \xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^T \mathbf{q}_i + b) \geq 1 - \xi_i + \|\mathbf{P}_i^{1/2}\mathbf{w}\| \\ & 0 \leq \xi_i \quad \forall \xi_i. \end{aligned} \tag{17}$$

The difference to eMIL (Equation (7)) is the fact that $\|\mathbf{P}_i^{1/2}\mathbf{w}\|$ is not multiplied with the label y_i. This leads to a hyperplane that separates ellipsoids, whereas in eMIL the positive ellipsoids can overlap the hyperplane. See Figure 2(b) for an illustration.



(a) Maximum margin hyperplane for eMIL.

(b) Maximum margin hyperplane for robust SVM.

Figure 2. Two negative (red) and two positive (green) ellipsoids with separating hyperplane and margin, notice the different hyperplane and margin for eMIL (a) and robust SVM (b).

2.6. Other MIL approaches

The MIL setting is very natural in many applications such as text classification (Andrews et al., 2002), image retrieval (Gehler & Chapelle, 2007) and object detection (Viola et al., 2006). For example, content based image retrieval represents an image as a bag containing image patches (examples $\mathbf{x}_{ij}$) and for a particular query, one is interested in returning images (bags $\{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}\}$) that contain the object, instead of solving the more complex problem of labeling every patch in the image.

Unfortunately, due to the weak labeling, it is unclear during training time of MIL methods how to allocate the positive label. Since any number of examples in a positive bag may be positive, one would naively have to look at all possible labelings that includes at least one positive label. For learning a classifier, the bag label is traditionally inferred as the max over the classification of all instances in the respective bag: $y_i = \operatorname{sgn} \max_{j=1}^{n_i}(\langle \mathbf{w}, \mathbf{x}_{ij}\rangle + b)$ (Andrews et al., 2002). This results in a non-convex optimization problem for finding a maximum-margin hyperplane for classification due to the negative max function not being convex. It also results in potentially expensive computations which involve optimising a combinatorial problem. Recently, there has been several proposals of making some assumptions about the structure of the bag such as using Markov random fields (Warrell & Torr, 2011) and low dimensional manifolds (Babenko et al., 2011).

In Section 4, we compare our method to the following algorithms for MIL: Two traditional approaches to solve the MIL problem, the earliest one being the method of axis-parallel rectangles (APR) (Dietterich et al., 1997), that was specifically designed for the MUSK1 and MUSK2 datasets and an extension of the diverse-density algorithm (Maron & Lozano-Pérez, 1998), EMDD, using Expectation-Maximization to find a positive witness (Zhang et al., 2002). We also compare to two extensions of a support vector machine mi-SVM (maximizing instance margin) and MI-SVM (maximizing bag margin), that lead to mixed-integer programs (Andrews et al., 2002); deterministic annealing methods to solve mi-SVM (AL-SVM) and MI-SVM (AW-SVM) (Gehler & Chapelle, 2007); a convex semi definite programming to the maximum instance margin problem (SDP) (Guo, 2009); MICA, an algorithm that uses convex combinations of positive bag instances (Mangasarian & Wild, 2008); and a recent approach developing a Gaussian process by building bag likelihood models from the GP latent variables (GPMIL) (Kim & la Torre, 2010).

Algorithm 1 eMIL: Sequential SOCP

```
Initialise $(\mathbf{w}_0, b_0)$ according to Equation (18)
while $\ell(\mathbf{w}_k, b_k) - \ell(\mathbf{w}_{k+1}, b_{k+1}) > \epsilon$ do
    Find the optimal solution $(\mathbf{w}_{k+1}, b_{k+1})$ of Equation (22), given $(\mathbf{w}_k, b_k)$.
end while
```

3. Solving eMIL

3.1. Difference of Convex Functions

We propose two approaches to optimize the resulting non-convex optimization problem (6). We derive a concave convex procedure (CCCP) in the following subsection, and a quasi-Newton approach (L-BFGS) in Section 3.3. While CCCP gives consistently lower optimal values on all the datasets that we tried, the gradient based method is usually much faster, especially in very high dimensional problems. However, the lower objective value typically also does not translate into significant improvements on test accuracy. For both approaches, we initialize by setting

$$\mathbf{w}_0, b_0 = \arg\min_{\mathbf{w},b} \sum_{i=1}^{B} \max(0, 1 - y_i(\mathbf{w}^\top \mathbf{q}_i + b)) + \frac{\lambda}{2}\|\mathbf{w}\|^2, \tag{18}$$

which is the maximum-margin hyperplane that separates the means of the bags. This can be seen as a first order approximation of the within-bag distribution.

3.2. Solving eMIL with CCCP

We can express the objective function (6) as a difference of convex functions and use CCCP to solve it, by solving a series of convex programs. See Yuille &

Rangarajan (2003) for the introduction of the CCCP, Sriperumbudur & Lanckriet (2009) for its convergence proof, and Le Thi & Pham Dinh (2005) for an overview on difference of convex functions algorithm. The solution of eMIL with CCCP is shown in Algorithm 1.

The decomposition of (6) into the difference of two convex functions $g(\mathbf{w}, b)$, $h(\mathbf{w}, b)$ is as follows:

$$\begin{aligned}\min_{\mathbf{w},b} & \sum_{y_-} \max\left(0, 1 + f(\mathbf{P}_i; \mathbf{q}_i)\right) \\ & + \sum_{y_+} \max\left(0, -1 + f(\mathbf{P}_i; \mathbf{q}_i)\right) + \frac{\lambda}{2}\|\mathbf{w}\|^2 && (19) \\ & - \sum_{y_+} \left(-1 + f(\mathbf{P}_i; \mathbf{q}_i)\right), && (20)\end{aligned}$$

where the first three lines (Equation (19)) correspond to $g(\mathbf{w}, b)$ and the last line (Equation (20)) to $-h(\mathbf{w}, b)$. Given the decomposition, CCCP proceeds by linearizing the concave part $-h(\mathbf{x}, b)$ at $\mathbf{w}_k, b_k$, solving the resulting convex optimization problem Equation (21), obtaining the optimal value $\mathbf{w}_{k+1}, b_{k+1}$ and repeating until convergence. The linearisation of $-h(\mathbf{x}, b)$ (Equation (20)) at $\mathbf{w}_k, b_k$ is given by: $-\langle \mathbf{w}, \partial h(\mathbf{w}_k, b_k)\rangle$. By taking the sum over the positive examples out of the inner product we arrive at:

$$\begin{aligned}\min_{\mathbf{w},b} & \sum_{y_-} \max\left(0, 1 + f(\mathbf{P}_i; \mathbf{q}_i)\right) \\ & + \sum_{y_+} \max\left(0, -1 + f(\mathbf{P}_i; \mathbf{q}_i)\right) + \frac{\lambda}{2}\|\mathbf{w}\|^2 && (21) \\ & - \sum_{y_+} \left(\left\langle \mathbf{w}, \frac{\mathbf{P}_i \mathbf{w}_k}{\sqrt{\mathbf{w}_k^\top \mathbf{P}_i \mathbf{w}_k}} + \mathbf{q}_i \right\rangle + b\right).\end{aligned}$$

By introducing slack variables $\boldsymbol{\xi}$, using $\mathbf{P}_i = \mathbf{A}_i^\top \mathbf{A}_i$ and finally converting the remaining objective function into second order cone constraint, we can now rewrite Equation (21) to get the equivalent (in terms of optimal solution $(\mathbf{w}, b)$) constrained optimisation problem Equation (22), which is a second order cone program (SOCP).

$$\begin{array}{lll} \min\limits_{\theta,\mathbf{w},b,\boldsymbol{\xi}} \theta & & \\ \text{s.t.} \quad \left\| \begin{pmatrix} \gamma \\ \sqrt{\frac{\lambda}{2}}\mathbf{w} \end{pmatrix} \right\| & \leq -\gamma + 1 & \\ \|\mathbf{A}_i \mathbf{w}\| + \mathbf{w}^\top \mathbf{q}_i + b & \leq \xi_i - 1, & \forall i\colon y_i = -1 \\ \|\mathbf{A}_i \mathbf{w}\| + \mathbf{w}^\top \mathbf{q}_i + b & \leq \xi_i + 1, & \forall i\colon y_i = +1 \\ \mathbf{0} & \leq \boldsymbol{\xi}. & \end{array} \qquad (22)$$

Where γ is just a placeholder for

$$\frac{1}{2}\left[1 - \sum_{y_+}\left(\left\langle \mathbf{w}, \frac{\mathbf{P}_i \mathbf{w}_k}{\sqrt{\mathbf{w}_k^\top \mathbf{P}_i \mathbf{w}_k}} + \mathbf{q}_i \right\rangle + b\right) + \sum_{i=1}^{B} \xi_i - \theta\right].$$

See Appendix B in the supplementary file for details.

3.3. Solving eMIL with BFGS

Another way of solving eMIL is to find a local minimum with a gradient based method. We use the quasi-Newton method L-BFGS (Byrd et al., 1995). To get a gradient of eMIL we use a smoothed version of the hinge-loss similar to (Chapelle, 2007; Wang et al., 2008), which has the following form $\ell_\delta(\mathbf{P}_i; \mathbf{q}_i, y_i, \mathbf{w}, b) =$

$$\begin{cases} \frac{(1 - y_i \cdot f(\mathbf{P}_i; \mathbf{q}_i))^2}{2\delta} & \text{if } 1 - \delta < y_i \cdot f(\mathbf{P}_i; \mathbf{q}_i) \leq 1 \\ 1 - y_i \cdot f(\mathbf{P}_i; \mathbf{q}_i) - \frac{\delta}{2} & \text{if } y_i \cdot f(\mathbf{P}_i; \mathbf{q}_i) \leq 1 - \delta \\ 0 & \text{if } y_i \cdot f(\mathbf{P}_i; \mathbf{q}_i) > 1, \end{cases}$$

where we choose an appropriately small δ. The gradient is shown in Appendix D in the supplementary file.

4. Benchmark datasets

We compare the performance of eMIL on the following datasets: The MUSK1 and MUSK2 datasets described in (Dietterich et al., 1997), three image annotation datasets (Elephant, Fox, Tiger) introduced in (Andrews et al., 2002) and 7 splits of the TREC9 dataset (TST1, TST2, TST3, TST4, TST7, TST9 and TST10) also described in (Andrews et al., 2002). The TREC9 datasets are extremely high-dimensional and sparse, having 66000 to 67000 features of which only a maximum of 30 are non-zero per instance. On average the MUSK1 and MUSK2 datasets contain approximately 6 and 60 instances per bag respectively. The average bag sizes for the image annotation and TREC9 data are 7 and 8 respectively. We minimize the regularized empirical risk function, Equation (6) using L-BFGS (Section 3.3).[1] To avoid numerical problems, when the ellipsoids are low rank we add a tiny positive constant to the diagonal of $\mathbf{P}_i$. This preserves the shape of the ellipsoid and makes $\mathbf{P}_i$ positive definite. We also experimented with different scaling factors (see Section 2.3 and Section 2.4), but could not generally improve test accuracy compared to simply using the estimated covariance matrix. By setting $\kappa(\alpha_i) = 1$ we implicitly use $\alpha_i = 0.5$ for all bags.

To be able to compare the performance of our method with previous methods we follow (Andrews et al.,

[1] eMIL is available at `http://bioweb.me/emil`

Table 1. Classification accuracy on the MUSK datasets (top block), image annotation (middle block), and TREC9 data. See Section 2.6 for description of previous approaches. Performance of those were obtained from the respective papers.

Dataset	eMIL	APR	EMDD	MI-SVM	mi-SVM	MICA	AL-SVM	AW-SVM	SDP	GPMIL
Musk 1	84.5	**92.4**	84.8	77.9	87.4	84.4	85.7	85.7	69.5	89.5
Musk 2	86.0	89.2	84.9	84.3	83.6	**90.5**	86.2	83.8	61.3	87.3
Tiger	**88.8**	-	72.1	84.0	78.4	82.0	78.5	83.0	73.6	87.4
Elephant	**84.0**	-	78.3	81.4	82.2	82.5	79.5	82.0	74.8	83.8
Fox	58.3	-	56.1	57.8	58.2	62.0	63.5	63.5	56.8	**65.8**
TST1	**95.9**	-	85.8	93.9	93.6	-	-	-	92.7	94.4
TST2	79.2	-	84.0	84.5	78.2	-	-	-	75.1	**85.3**
TST3	**86.8**	-	69.0	82.2	87.0	-	-	-	74.3	86.1
TST4	84.0	-	80.5	82.4	82.8	-	-	-	77.7	**85.3**
TST7	80.4	-	75.4	78.0	**81.3**	-	-	-	72.5	80.3
TST9	69.0	-	65.5	60.2	67.5	-	-	-	59.9	**70.8**
TST10	**83.4**	-	78.5	79.5	79.6	-	-	-	74.4	80.4

2002) and employ the following procedure on all of the datasets: We use 10-fold cross-validation and search coarsely for an optimal regularization parameter λ. This procedure is repeated 10 times on random permutations of the data and the results are averaged.

4.1. Feature space corresponding to kernels

For the MUSK-datasets we use a Gaussian kernel with $\gamma = 10^{-6}$. Since we optimise eMIL in the primal, we use kernel PCA to project the infinite dimensional feature vector to a lower-dimensional subspace. For MUSK2 we additionally restrict the number of basis vectors to 2500 to save memory.

To be able to optimise in the primal, we explicitly compute a finite dimensional representation of the features corresponding to the kernel. Following Zien et al. (2007), we use kernel PCA (Schölkopf & Smola, 2002) to find a d dimensional representation of the data from the kernel $k(\mathbf{x}_i, \mathbf{x}_j)$. Since the representer theorem ensures that the optimal solution $\mathbf{w}$ lies in a finite dimensional subspace, we first find a basis for this subspace and then represent the instances in terms of this basis.

The basis needs to satisfy two criteria: (1) each basis vector has to be expressed in terms of the feature maps, and (2) the basis vectors should be orthonormal. Hence for a kernel matrix $\mathbf{K}$, we need to find a set of coefficients in a matrix $\mathbf{A}$ such that $\mathbf{A}^\top \mathbf{K} \mathbf{A} = \mathbf{I}$. One way to do so is to compute the eigenvalue decomposition of $\mathbf{K} = \mathbf{V} \Lambda \mathbf{V}^\top$ and set $\mathbf{A} = \mathbf{V} \Lambda^{-\frac{1}{2}}$.

4.2. Results

On the musk datasets (Table 1, *top*) eMIL shows comparable performance to the methods motivated by finding a witness instance for positive bags (MI-SVM, MICA, AL-SVM) and the instance level maximum margin methods (mi-SVM, AW-SVM). The good accuracy of APR on the musk datasets can be explained by the fact that the hypothesis class of axis-parallel rectangles was specifically developed for this particular dataset. For the image annotation datasets (Table 1, *middle*) eMIL has the best accuracy on the tiger and elephant dataset and beats the MIL-SVM methods on the Fox dataset. Our method achieves highest accuracy for some of the TREC9 datasets (Table 1, *bottom*), and comparable accuracy to the best method (GPMIL or mi-SVM) on all the TREC9 datasets, apart from TST2.

5. Defective wheel classification

In this section we apply eMIL to a real world MIL problem: Detecting defective wheels of freight trains from multiple dynamic vertical wheel force measurements. Late or undetected wheel defects on railway vehicles result in increased infrastructure maintenance due to damage of the railway infrastructure, like track systems or civil engineering works, and reduced availability of the vehicle pool, maintenance compounds and infrastructure. Most importantly, wheel defects are the major source of noise and vibration emissions of rail traffic. This makes the automatic, reliable and timely detection of wheel defects an essential part of any railway infrastructure safety monitoring system.

For detecting wheel defects, we are given eight measurements per wheel, obtained by eight sensors installed on the tracks as the train runs over the measurement site in full operational speed. A defect only

Table 2. Classification accuracy on the wheel defect dataset. Rank gives the average rank. If two methods are equal, they both get the same rank. See text for details.

Acc.	eMIL	ALSVM	AWSVM	ALPSVM	SVM
Avg.	0.70	0.64	0.68	0.63	0.67
Std. dev.	0.05	0.08	0.06	0.07	0.07
Rank	1.5	3.5	2.	3.5	2.5

impacts a measurement if it hits the part of the track where the sensor is installed directly. This results in eight measurements per wheel, with usually one measurement affected by a defective wheel. By considering all measurements from different sensors for the same wheel as a bag a natural setting for MIL is obtained. The labeled data was obtained by running a test train with a known configuration of wheel defects, resulting in 100 positive and negative bags.

5.1. Experimental protocol

Each measurement (instance) consists of a time series of the vertical wheel force. We use the Global Alignement (GA) kernel for time series, described in Cuturi et al. (2007) and Cuturi (2011). The GA kernel can be seen as a generalization of dynamic time warping (DTW) with a soft-max over all the alignments. To optimise eMIL in the primal we again project the features corresponding to this kernel to a lower dimensional subspace (see Section 4.1).

We compare our algorithm (eMIL) to the three deterministic annealing methods described in Gehler & Chapelle (2007) in Table 2. We chose these methods because they solve the mi-SVM and MI-SVM formulations of MIL and our method could be seen as a generalization of the maximum bag margin method MI-SVM. Furthermore, an implementation was readily available. ALP-SVM is a balancing extension of AL-SVM, it needs to know an estimate of the fraction of positive points in a positive bag $p^\star$ a priori. For ALP-SVM, we provide extra information by setting $p^\star$ to 1/8 because we expect one sensor on average to see a defect. In addition, we compare a baseline method using a standard Support Vector Machine (denoted “SVM”). To convert from bag labels to instance labels, we set all instance labels to the bag label.

The reported accuracy is averaged over 10 random permutations of the following two stage evaluation scheme: Half of the data is split of for model selection and half for evaluation. On the first half of the data the optimal parameter for regularization is searched over $\lambda \in 10^{[-2,\dots,-5]}$. This is done with estimating test error with 5-fold cross validation for all values of λ and the λ with lowest test error is kept. This λ is then used to train the classifier on the full first half of the dataset and test error is computed on the second evaluation half of the dataset. If multiple parameter values give the same test accuracy, the one closest to the average is kept for training.

5.2. Results on the wheel data

From Table 2 we see that eMIL has the highest average accuracy on the test set. However, due to the large variation between the different splits of the data, the standard deviation is large. We also compared the ranks of the methods for each split, with the method with highest accuracy obtaining rank 1, and the lowest rank 5. We see in Table 2 that eMIL has average rank 1.5, which is the best among all considered methods. Interestingly, the naive SVM approach performs well.

6. Discussion

Motivated by the real world application of detecting wheel defects from multiple dynamic force measurements, we derive an ellipsoidal algorithm to solve MIL, resulting in a classifier that optimises a class conditional distance between an ellipsoid and a hyperplane. We show that representing bags as ellipsoids amounts to finding a robust solution. Using only the assumption that the instances are samples from a distribution with finite mean and covariance, we derive an appropriate scaling factor for eMIL. We propose two approaches to solve the optimization problem: a CCCP approach which results in a sequential SOCP, and a quasi-Newton method based on L-BFGS.

Our algorithm results in state of the art performance on benchmark MIL datasets, demonstrating the effectiveness of the method. For classifying defective wheels with multiple instance time series data eMIL consistently outperforms AL-SVM, AW-SVM and ALP-SVM, which are recent improvements to SVM type MIL approaches. We are currently working with our collaborators in the rail industry to test this approach in the safety monitoring system.

Acknowledgements

We would like to thank Stefan Koller from SBB for help with the freight train data, Brian McWilliams for helpful discussions and one anonymous reviewer for detailed comments. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

Andrews, Stuart, Tsochantaridis, Ioannis, and Hofmann, Thomas. Support vector machines for multiple-instance learning. In *NIPS*, 2002.

Babenko, Boris, Verma, Nakul, Dollár, Piotr, and Belongie, Serge. Multiple instance learning with manifold bags. In *ICML*, pp. 81–88, 2011.

Ben-Tal, Aharon, Ghaoui, Laurent El, and Nemirovski, Arkadi. *Robust Optimization*. Princeton University Press, 2009.

Bertsimas, Dimitris and Popescu, Ioana. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15: 780–804, 2001.

Byrd, R.H., Lu, P., Nocedal, J., and Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5): 1190–1208, 1995.

Chapelle, Olivier. Training a support vector machine in the primal. *Neural Computation*, 19(5):1155–1178, 2007.

Cuturi, M., Vert, J.-P., Birkenes, O., and Matsui, T. A kernel for time series based on global alignments. In *ICASSP*, volume 2, 2007.

Cuturi, Marco. Fast global alignment kernels. In *ICML 2011*, 2011.

Dietterich, Thomas G., Lathrop, Richard H., and Lozano-Perez, Toms. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31 – 71, 1997.

Gehler, Peter V. and Chapelle, Olivier. Deterministic annealing for multiple-instance learning. In *AISTATS*, 2007.

Guo, Yuhong. Max-margin multiple-instance learning via semidefinite programming. In *Advances in Machine Learning*, volume 5828 of *Lecture Notes in Computer Science*, pp. 98–108. Springer, 2009.

Kim, M. and la Torre, F. De. Gaussian processes multiple-instance learning. In *ICML*, 2010.

Lanckriet, Gert R. G., Ghaoui, Laurent El, Bhattacharyya, Chiranjib, and Jordan, Michael I. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3:555–582, 2002.

Le Thi, Hoai An and Pham Dinh, Tao. The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research*, 133:23–46, 2005. ISSN 0254-5330.

Mangasarian, OL and Wild, E.W. Multiple instance classification via successive linear programming. *Journal of Optimization Theory and Applications*, 137(3):555–568, 2008.

Maron, O. and Lozano-Pérez, T. A framework for multiple-instance learning. In *NIPS*, 1998.

Schölkopf, Bernhard and Smola, Alexander J. *Learning with Kernels*. MIT Press, 2002.

Shivaswamy, Pannagadatta K., Bhattacharyya, Chiranjib, and Smola, Alexander J. Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, 7:1283–1314, 2006.

Sra, S., Nowozin, S., and Wright, S.J. *Optimization for Machine Learning*. Mit Press, 2011.

Sriperumbudur, Bharath K. and Lanckriet, Gert R. G. On the convergence of the concave-convex procedure. In *NIPS*, pp. 1759–1767, 2009.

Viola, P., Platt, J., and Zhang, C. Multiple instance boosting for object detection. In *NIPS*, 2006.

Wang, Li, Zhu, Ji, and Zou, Hui. Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*, 24(3):412–419, 2008.

Warrell, J. and Torr, P. Multiple-instance learning with structured bag models. *Energy Minimazation Methods in Computer Vision and Pattern Recognition*, pp. 369–384, 2011.

Yuille, A.L. and Rangarajan, A. The concave-convex procedure. *Neural Comput.*, 15(4):915–936, 2003.

Zhang, Qi, Goldman, Sally A., Yu, Wei, and Fritts, Jason E. Content-based image retrieval using multiple-instance learning. In *ICML*, 2002.

Zien, A., De Bona, F., and Ong, C.S. Training and approximation of a primal multiclass support vector machine. In *ASMDA*, 2007.