Gaussian Processes Multiple Instance Learning

Minyoung Kim MIKIM@CS.RUTGERS.EDU
Fernando De la Torre FTORRE@CS.CMU.EDU
Robotics Institute, Carnegie Mellon University, PA 15213, USA

Abstract

This paper proposes a multiple instance learning (MIL) algorithm for Gaussian processes (GP). The GP-MIL model inherits two crucial benefits from GP: (i) a principle manner of learning kernel parameters, and (ii) a probabilistic interpretation (e.g., variance in prediction) that is informative for better understanding of the MIL prediction problem. The bag labeling protocol of the MIL problem, namely the existence of a positive instance in a bag, can be effectively represented by a sigmoid likelihood model through the max function over GP latent variables. To circumvent the intractability of exact GP inference and learning incurred by the noncontinuous max function, we suggest two approximations: first, the soft-max approximation; second, the use of witness indicator variables optimized with a deterministic annealing schedule. The effectiveness of GP-MIL against other state-of-the-art MIL approaches is demonstrated on several benchmark MIL datasets.

1. Introduction

In supervised learning the training data consist of pairs of input objects (typically vectors) and the desired outputs, $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. Throughout the paper, we will consider the binary classification setting, where $y_i \in \{+1, -1\}$. In multiple instance learning (MIL) (Dietterich et al., 1997), on the other hand, the assumption that each instance has one label is relaxed in the following manner: (i) we are given B bags of instances $\{\mathbf{X}_b\}_{b=1}^B$ where each bag, $\mathbf{X}_b = \{\mathbf{x}_{b,1}, \dots, \mathbf{x}_{b,n_b}\}$, consists of n_b instances ($\sum_b n_b = n$), and (ii) the labels are provided only at the bag-level such that for each bag b, $Y_b = -1$ if $y_i = -1$ for *all* $i \in b$, and $Y_b = +1$ if $y_i = +1$ for *some* $i \in b$.

The notion of bag together with the labeling protocol often makes the MIL formulation more realistic than the standard classification setting for particular types of applications such as image retrieval (Zhang et al., 2002; Gehler & Chapelle, 2007), and text classification (Andrews et al., 2003). For instance, content-based image retrieval fits the MIL framework well because an image can be represented as a bag comprised of smaller regions/patches (i.e., instances). Given a query for a particular object, one may be interested in deciding only whether the image contains the queried object ($Y_b = +1$) or not ($Y_b = -1$), instead of solving the more difficult (and usually unnecessary) problem of labeling every single patch in the image. Similarly, in text classification, one is more concerned with the concept/topic (i.e., bag label) of an entire paragraph than labeling each of the sentences that comprise the paragraph. Other applications include object detection (Viola et al., 2005), time series classification (Nguyen et al., 2009) and protein identification (Tao et al., 2004).

Traditionally, the MIL problem was tackled by specially tailored algorithms; for example, the hypothesis class of axis-parallel rectangles (Dietterich et al., 1997) and the so-called diverse density to define a measure of proximity between a bag and a positive intersection point (Maron & Lozano-Perez, 1998). Another class of algorithms treats the MIL problem as a standard classification problem at a bag-level via proper development of kernels or distance measures on the bag space (Wang & Zucker, 2000; Gärtner et al., 2002; Tao et al., 2004).

Recently, a different perspective that regards MIL as a missing-label problem has emerged. Unlike the negative instances which are all labeled negatively, the instances in the positive bags are considered as latent variables with the positive bag constraint (i.e., at least one of them is positive, $\sum_i \frac{y_i+1}{2} \geq 1$). In this treatment, a direct approach is to formulate a standard

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

(instance-level) classification problem (e.g., SVM) that can be optimized over the model and the latent variables simultaneously. The *mi-SVM* approach of (Andrews et al., 2003) is derived in this manner.

One drawback of such approaches is that they involve a (mixed) integer programming which is generally difficult to solve. For instance, in (Andrews et al., 2003), certain heuristic optimization methods were employed. Recently, the deterministic annealing (DA) algorithm has been employed (Gehler & Chapelle, 2007), which approximates the original problem to a continuous optimization by introducing binary random variables in conjunction with the temperature-scaled entropy term. The DA algorithm begins with a high temperature to solve a easier convex-like problem, and iteratively reduces the temperature with the warm starts.

Instead of dealing with all the instances in a positive bag individually, a more insightful strategy is to focus on the *most positive* instance, often referred to as the *witness*, which is responsible for determining the label of a positive bag. In the SVM formulation, the *MI-SVM* of (Andrews et al., 2003) directly aims at maximizing the margin of the instance with the most positive confidence w.r.t. the current model $\mathbf{w}$ (i.e., $\max_{i\in b}\langle \mathbf{w}, \mathbf{x}_{b,i}\rangle$), while in the *MICA* algorithm (Mangasarian & Wild, 2008), they indirectly form a witness using convex combination over all instances in a positive bag. The *EM-DD* algorithm of (Zhang et al., 2002) extends the diverse density framework of (Maron & Lozano-Perez, 1998) by incorporating the witnesses. In (Gehler & Chapelle, 2007) the DA algorithms have also been applied to the witness-identifying SVMs, exhibiting superior performance to existing approaches.

Although some of these MIL algorithms, especially the SVM-based discriminative methods, are quite effective for a variety of situations, most approaches are nonprobabilistic, thus unable to capture the underlying generative process of the data. In this paper we introduce a novel MIL algorithm using Gaussian processes (GP), which we call it *GPMIL*. Motivated by the fact that a bag label is solely determined by the instance that has the highest confidence toward the positive class, we design the bag class likelihood as the sigmoid function over the maximum GP latent variables on the instances. By marginalizing out the latent variables, we have a nonparametric, nonlinear probabilistic model $P(Y_b|\mathbf{X}_b)$ that fully respects the bag labeling protocol of MIL.

Dealing with a probabilistic bag class model is not completely new. For instance, the Noisy-OR model suggested by (Viola et al., 2005) is a reasonable direction, where the learning is formulated within the functional gradient boosting framework (Friedman, 1999). A similar Noisy-OR modeling has also been proposed recently by (Raykar et al., 2008) where their Bayesian treatment is shown to lead to effective feature selection. In these approaches, however, the bag class model is built from the *instance-level* classification models $P(y_i|\mathbf{x}_i)$, more specifically, $P(Y_b = -1|\mathbf{X}_b) = \prod_{i\in b} P(y_i = -1|\mathbf{x}_i)$ and $P(Y_b = +1|\mathbf{X}_b) = 1 - P(Y_b = -1|\mathbf{X}_b)$, which may incur several drawbacks. First of all, it involves additional modeling effort for the instance-level classifiers, which may be unnecessary, or only indirectly relevant to the bag class decision. Moreover, the Noisy-OR model combines the instance-level classifiers in a product form, treating each instance independently. This ignores the impact of potential interaction among the neighboring instances, which may be crucial for accurate bag class prediction. On the other hand, our GPMIL represents the bag class model directly without using typically unnecessary instance-level classifiers. The interaction among instances is also incorporated through the GP prior which essentially enforces a smoothness regularization along the neighboring structure of the instances.

In addition to the above-mentioned advantages, the most important benefit of the GPMIL, especially which contrasted the SVM-based approaches, is that the kernel hyperparameters can be learned in a principled manner (e.g., empirical Bayes), thus avoiding grid search and being able to exploit a variety of kernel families with complex forms. Unfortunately, one caveat of the GPMIL is the computational issue. To circumvent the intractability in the exact GP inference and learning incurred by the non-continuous max function, we suggest two approximations: the soft-max approximation, and the use of witness indicator variables which can be further optimized by a deterministic annealing schedule. Both approaches often exhibit more accurate prediction than most recent SVM variants.

The paper is organized as follows: In Sec. 2 the GPMIL framework is introduced with the soft-max approximation for inference and learning. The witness variable based approximation for GPMIL is described in Sec. 3. Experimental results on both synthetic data and real-world MIL benchmark datasets are provided in Sec. 4. We conclude the paper in Sec. 5.

2. GP Multiple Instance Learning

This section proposes a novel Gaussian process (GP) model for the MIL problem, which we denote by *GPMIL*. Our approach builds a bag class likelihood model from the GP latent variables, where the likelihood is the sigmoid of the *maximum* latent variables. The

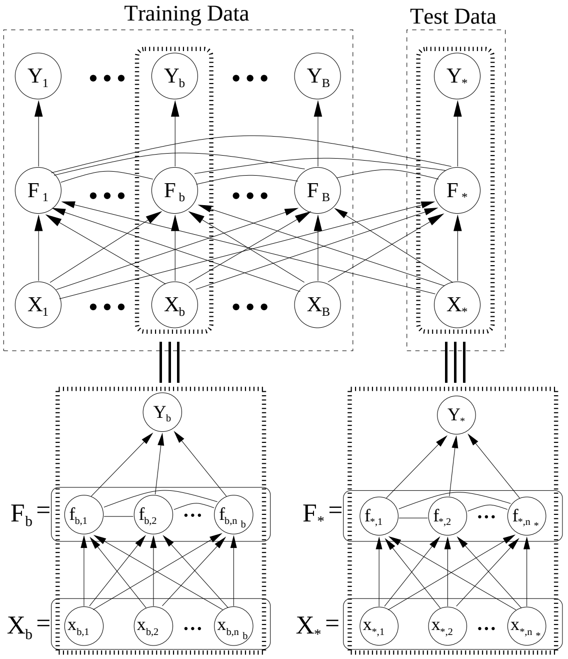

Figure 1. Graphical model for GPMIL.

readers are encouraged to refer to the standard materials, e.g., (Rasmussen & Williams, 2006), for the backgrounds on the Gaussian process and the notations used in the paper.

In the GP framework, we consider the latent variable f for each input $\mathbf{x}$, which can be seen as a function evaluated at (or indexed by) $\mathbf{x}$, i.e., $f(\mathbf{x})$, where the real-valued (nonlinear) function $f(\cdot)$ follows the GP prior. The GP prior is characterized by the covariance function (i.e., kernel) $k(\cdot, \cdot)$ defined on the input space, which implies that:

$$\text{Cov}(f(\mathbf{x}_i), f(\mathbf{x}_j)) = k(\mathbf{x}_i, \mathbf{x}_j), \text{for any } \mathbf{x}_i \text{ and } \mathbf{x}_j. \quad (1)$$

We assume that the bag b is comprised of n_b points $\mathbf{X}_b = \{\mathbf{x}_{b,1}, \ldots, \mathbf{x}_{b,n_b}\}$. We then denote the GP latent variables for the bag b by $\mathbf{F}_b = \{f_{b,1}, \ldots, f_{b,n_b}\}$. Similar to the standard GP classification case, we will consider f_i ($\in \mathbf{F}_b$) as a *confidence* score toward the (instance-level) positive class for $\mathbf{x}_i$ ($\in \mathbf{X}_b$). That is, the sign of f_i indicates the (instance-level) class label y_i, and its magnitude implies how confident it is. In MIL, our goal is to devise a bag class likelihood model $P(Y_b|\mathbf{F}_b)$ instead of the instance-level model $P(y_i|f_i)$. Note that the latter is a special case of the former since an instance can be seen as a singleton bag. Once we have the bag class likelihood model, we can then marginalize out all the latent variables $\mathbf{F} = \{\mathbf{F}_b\}_{b=1}^B$ under the Bayesian formalism using the GP prior $P(\mathbf{F}|\mathbf{X})$ given the entire input $\mathbf{X} = \{\mathbf{X}_b\}_{b=1}^B$.

Now, consider the situation where the bag b is labeled as positive ($Y_b = +1$). The chance is determined solely by the single point that is the *most likely positive* (i.e., the largest f). The larger the confidence f, the higher the chance is. The other instances do not contribute to the bag label prediction no matter what their confidence scores are. Hence, we can write the probability of a bag b labeled as positive as:

$$P(Y_b = +1|\mathbf{F}_b) \propto \exp(\max_{i \in b} f_i). \quad (2)$$

Similarly, the odds of the bag b being labeled as negative ($Y_b = -1$) is affected solely by the single point which is the *least likely negative*. As far as that point has a negative confidence f, the label of the bag is negative, and the larger the confidence $-f$, the higher the chance is. This leads to the model:

$$P(Y_b = -1|\mathbf{F}_b) \propto \exp(\min_{i \in b} -f_i). \quad (3)$$

Combining (2) and (3), we have the following bag class likelihood model:

$$P(Y_b|\mathbf{F}_b) = \frac{1}{1 + \exp(-Y_b \max_{i \in b} f_i)}. \quad (4)$$

Note also that (4), in the limiting case where all the bags become singletons (i.e., classical supervised classification), is equivalent to the standard Gaussian process classification model with the sigmoid link[1].

When incorporating the likelihood model (4) into the GP framework, one bottleneck is that we have non-differentiable formulas due to the max function. We approximate it by the soft-max[2]: $\max(z_1, \ldots, z_m) \approx \log \sum_i \exp(z_i)$. This leads to the approximated bag class likelihood model:

$$\begin{aligned} P(Y_b|\mathbf{F}_b) &\approx \frac{1}{1 + \exp(-Y_b \log \sum_{i \in b} e^{f_i})} \\ &= \frac{1}{1 + (\sum_{i \in b} e^{f_i})^{-Y_b}}. \end{aligned} \quad (5)$$

Whereas the soft-max is often a good approximation for the max function, it should be noted that unlike in standard GP classification with the sigmoid link, the negative log-likelihood $-\log P(Y_b|\mathbf{F}_b) = \log(1 + (\sum_{i \in b} e^{f_i})^{-Y_b})$ is not a convex function of $\mathbf{F}_b$ for $Y_b = +1$ (although it is convex for $Y_b = -1$). This corresponds to a non-convex optimization in the approximated GP posterior computation and learning when the Laplace or variational approximation methods are

[1] So, it is also possible to have a probit version of (4), namely $P(Y_b|\mathbf{f}_b) = \Phi(Y_b \max_{i \in b} f_i)$, where $\Phi(\cdot)$ is the cumulative normal function.

[2] It is well known that the soft-max provides relatively tight bounds for the max, $\max_{i=1}^m z_i \leq \log \sum_{i=1}^m \exp(z_i) \leq \max_{i=1}^m z_i + \log m$. Another nice property is that the soft-max is a convex function.

adopted. However, using the (scaled) conjugate gradient search with different starting iterates, one can typically obtain a well-approximated posterior with a meaningful set of hyperparameters.

Before we proceed further to the details of inference and learning, we briefly discuss the benefits of the GPMIL compared to the existing MIL methods. As mentioned earlier, the GPMIL directly models the bag class distribution, without suboptimally introducing instance-level models such as the Noisy-OR model of (Viola et al., 2005). Also, framed in the GP framework, the posterior estimation and the hyperparameter learning can be accomplished by simple gradient search with similar complexity as the standard GP classification, while enables a probabilistic interpretation (e.g., uncertainty in prediction). Moreover, GPMIL have a principled way to learn the kernel hyperparameters under the Bayesian formalism, which is not properly handled by other kernel-based MIL methods.

2.1. Posterior, Evidence, and Prediction

From the latent-to-output likelihood model (5), our generative GPMIL model can be depicted in a graphical representation as Fig. 1. Following the GP framework, all the latent variables $\mathbf{F} = \{\mathbf{F}_1, \dots, \mathbf{F}_B\} = \{f_{b,i}\}_{b,i}$ are dependent on one another as well as on all the training input points $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_B\} = \{\mathbf{x}_{b,i}\}_{b,i}$, conforming to the following distribution:

$$P(\mathbf{F}|\mathbf{X}) = \mathcal{N}(\mathbf{F}; 0, \mathbf{K}), \tag{6}$$

Similarly, for a new test bag $\mathbf{X}_* = \{\mathbf{x}_{*,1}, \dots, \mathbf{x}_{*,n_*}\}$ together with the corresponding latent variables $\mathbf{F}_* = \{f_{*,1}, \dots, f_{*,n_*}\}$, we have a joint Gaussian prior on the concatenated latent variables, $\{\mathbf{F}_*, \mathbf{F}\}$, from which the predictive distribution on $\mathbf{F}_*$ can be derived as (by conditional Gaussian):

$$\begin{aligned} P(\mathbf{F}_*|\mathbf{X}_*, \mathbf{F}, \mathbf{X}) = \mathcal{N}\Big(&\mathbf{F}_*;\ \mathbf{k}(\mathbf{X}_*)^\top \mathbf{K}^{-1}\mathbf{F}, \\ & k(\mathbf{X}_*, \mathbf{X}_*) - \mathbf{k}(\mathbf{X}_*)^\top \mathbf{K}^{-1}\mathbf{k}(\mathbf{X}_*)\Big), \end{aligned} \tag{7}$$

where $\mathbf{k}(\mathbf{X}_*)$ is the $(n \times n_*)$ train-test kernel matrix whose ij-th element is $k(\mathbf{x}_i, \mathbf{x}_{*,j})$, and $k(\mathbf{X}_*, \mathbf{X}_*)$ is the $(n_* \times n_*)$ test-test kernel matrix whose ij-th element is $k(\mathbf{x}_{*,i}, \mathbf{x}_{*,j})$.

Under the usual i.i.d. assumption, the entire likelihood $P(\mathbf{Y} = [Y_1, \dots, Y_B]|\mathbf{F})$ is the product of the individual bag likelihoods $P(Y_b|\mathbf{F}_b)$ in (5). That is,

$$P(\mathbf{Y}|\mathbf{F}) = \prod_{b=1}^{B} P(Y_b|\mathbf{F}_b) \approx \prod_{b=1}^{B} \frac{1}{1 + (\sum_{i \in b} e^{f_i})^{-Y_b}}. \tag{8}$$

Equipped with (6) and (8), one can compute the posterior distribution $P(\mathbf{F}|\mathbf{Y}, \mathbf{X}) \propto P(\mathbf{F}|\mathbf{X})P(\mathbf{Y}|\mathbf{F})$ and the evidence (or the data likelihood) $P(\mathbf{Y}|\mathbf{X}) = \int_{\mathbf{F}} P(\mathbf{F}|\mathbf{X})P(\mathbf{Y}|\mathbf{F})$, where the GP learning maximizes the evidence w.r.t. the kernel hyperparameters (also known as the empirical Bayes). However, similar to the GP classification cases, the non-Gaussian likelihood term (8) causes intractability in the exact computation, and we resort to some approximation. Here we focus on the Laplace approximation[3].

The Laplace approximation essentially replaces the product $P(\mathbf{Y}|\mathbf{F})P(\mathbf{F}|\mathbf{X})$ by a Gaussian with the mean equal to the mode of the product, and the covariance equal to the inverse Hessian of the product evaluated at the mode. For this purpose, we rewrite

$$\begin{aligned} P(\mathbf{Y}|\mathbf{F})P(\mathbf{F}|\mathbf{X}) &= \exp(-S(\mathbf{F})) \cdot |\mathbf{K}|^{-1/2} \cdot (2\pi)^{-n/2}, \\ \text{where} \ \ S(\mathbf{F}) &= \textstyle\sum_{b=1}^{B} l(Y_b, \mathbf{F}_b) + \frac{1}{2}\mathbf{F}^\top \mathbf{K}^{-1}\mathbf{F}, \\ l = -\log P(Y_b|\mathbf{F}_b) &\approx \log\Big(1 + (\textstyle\sum_{i \in b} e^{f_i})^{-Y_b}\Big). \end{aligned} \tag{9}$$

We first find the minimum of $S(\mathbf{F})$, namely

$$\widehat{\mathbf{F}} = \arg\min_{\mathbf{F}} S(\mathbf{F}), \tag{10}$$

where the optimum is denoted by $\widehat{\mathbf{F}}$. Solving (10) can be done by gradient search as usual. Unlike the standard GP classification, however, notice that $S(\mathbf{F})$ is a non-convex function of $\mathbf{F}$ since the Hessian of $S(\mathbf{F})$, $\mathbf{H} + \mathbf{K}^{-1}$, is generally not positive definite, where $\mathbf{H}$ is the block diagonal matrix whose b-th block has the ij-th entry $[\mathbf{H}_b]_{ij} = \frac{\partial^2 l(Y_b, \mathbf{F}_b)}{\partial f_i \partial f_j}$ for $i, j \in b$. Although this may hinder obtaining the global minimum easily, $S(\mathbf{F})$ is bounded below by 0 (from (9)), and the (scaled) conjugate or Newton-type gradient search with different initial iterates can yield a reliable solution.

We then approximate $S(\mathbf{F})$ by a quadratic function using its Hessian evaluated at $\widehat{\mathbf{F}}$, namely $\mathbf{H}(\widehat{\mathbf{F}}) + \mathbf{K}^{-1}$. Yet, in order to enforce a convex quadratic form, we need to address the case that $\mathbf{H} + \mathbf{K}^{-1}$ is not positive definite, which although very rare, could happen as gradient search only discovers a point close (not exactly the same) to local minima. We approximate it to the closest positive definite matrix by projecting it onto the PSD cone. More specifically, we let $\mathbf{Q} \approx \mathbf{H} + \mathbf{K}^{-1}$, with $\mathbf{Q} = \sum_i \max(\lambda_i, \epsilon)\mathbf{v}_i\mathbf{v}_i^\top$, where λ and $\mathbf{v}$ are the eigenvalues/vectors of $\mathbf{H} + \mathbf{K}^{-1}$, and ϵ is a small positive constant. In this way $\mathbf{Q}$ is a positive

[3] Although it is feasible, here we do not take the variational approximation into consideration for simplicity. Unlike the standard GP classification, it is difficult to perform, for instance, the Expectation Propagation (EP) approximation since moment matching, the core step in EP that minimizes the KL divergence between the marginal posteriors, requires integration over the likelihood function in (5), which requires further elaboration.

definite matrix closest to the Hessian with precision ϵ. Letting $\widehat{\mathbf{Q}}$ be $\mathbf{Q}$ evaluated at $\widehat{\mathbf{F}}$, we approximate $S(\mathbf{F})$ by the following quadratic function (i.e., using the Taylor expansion)

$$S(\mathbf{F}) \approx S(\widehat{\mathbf{F}}) + \frac{1}{2}(\mathbf{F} - \widehat{\mathbf{F}})^\top \widehat{\mathbf{Q}}(\mathbf{F} - \widehat{\mathbf{F}}), \qquad (11)$$

which leads to Gaussian approximation for $P(\mathbf{F}|\mathbf{Y}, \mathbf{X})$

$$P(\mathbf{F}|\mathbf{Y}, \mathbf{X}) \approx \mathcal{N}(\mathbf{F}; \widehat{\mathbf{F}}, \widehat{\mathbf{Q}}^{-1}). \qquad (12)$$

The data likelihood (i.e., evidence) immediately follows from the similar approximation,

$$P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) \approx \exp(-S(\widehat{\mathbf{F}}))|\widehat{\mathbf{Q}}|^{-1/2}|\mathbf{K}|^{-1/2}. \qquad (13)$$

We then maximize (13) w.r.t. the kernel parameters $\boldsymbol{\theta}$ by gradient search. The overall learning algorithm is depicted in Algorithm 1.

Algorithm 1 GPMIL Learning

```
**Input:** Initial guess $\boldsymbol{\theta}$, the tolerance parameter $\tau$.
**Output:** Learned hyperparameters $\boldsymbol{\theta}$.
(a) Find $\widehat{\mathbf{F}}$ from (10) for current $\boldsymbol{\theta}$.
(b) Compute $\widehat{\mathbf{Q}}$ using the PSD cone projection.
(c) Maximize (13) w.r.t. $\boldsymbol{\theta}$.
**if** $||\boldsymbol{\theta} - \boldsymbol{\theta}^{old}|| > \tau$ **then**
  Go to (a).
**else**
  Return $\boldsymbol{\theta}$.
**end if**
```

Given a new test bag $\mathbf{X}_* = \{\mathbf{x}_{*,1}, \ldots, \mathbf{x}_{*,n_*}\}$, it is easy to derive the predictive distribution for the corresponding latent variables $\mathbf{F}_* = \{f_{*,1}, \ldots, f_{*,n_*}\}$. Using the Gaussian approximated posterior (12) together with the conditional Gaussian prior (7), we have:

$$P(\mathbf{F}_*|\mathbf{X}_*, \mathbf{Y}, \mathbf{X}) = \mathcal{N}\Big(\mathbf{F}_*;\ \mathbf{k}(\mathbf{X}_*)^\top \mathbf{K}^{-1}\widehat{\mathbf{F}},$$
$$k(\mathbf{X}_*, \mathbf{X}_*) + \mathbf{k}(\mathbf{X}_*)^\top(\mathbf{K}^{-1}\widehat{\mathbf{Q}}^{-1}\mathbf{K}^{-1} - \mathbf{K}^{-1})\mathbf{k}(\mathbf{X}_*)\Big).$$

Finally, the predictive distribution for the test bag class label Y_* can be obtained by marginalizing out $\mathbf{F}_*$, namely

$$P(Y_*|\mathbf{X}_*, \mathbf{Y}, \mathbf{X}) = \int_{\mathbf{F}_*} P(\mathbf{F}_*|\mathbf{X}_*, \mathbf{Y}, \mathbf{X})P(Y_*|\mathbf{F}_*). \qquad (14)$$

The integration in (14) generally needs further approximation. If one is only interested in the mean prediction (i.e., the predicted class label), it is possible to approximate $P(\mathbf{F}_*|\mathbf{X}_*, \mathbf{Y}, \mathbf{X})$ by a delta function at its mean (mode), $\boldsymbol{\mu} := \mathbf{k}(\mathbf{X}_*)^\top \mathbf{K}^{-1}\widehat{\mathbf{F}}$, which yields the test prediction:

$$\text{Class}(Y_*) \approx \text{sign}\bigg(\frac{1}{1 + (\sum_{i \in *} e^{\mu_i})^{-1}} - 0.5\bigg). \qquad (15)$$

3. GPMIL using Witness Variables

Although the approach in Sec. 2 is reasonable, one drawback is that the target function we approximate (i.e., $S(\mathbf{F})$) is not in general a convex function (due to the non-convexity of $-\log P(Y_b|\mathbf{F}_b)$), where we perform the PSD projection step to find the closest convex function in the Laplace approximation. This section addresses this issue in a different way by introducing the so-called *witness latent variables* which indicate the most probably positive instances in the bags.

For each bag b, we introduce the witness indicator random variables $\mathbf{P}_b = [p_{b,1}, \ldots, p_{b,n_b}]^\top$, where $p_{b,i}$ represents the probability that $\mathbf{x}_{b,i}$ is considered as a *witness* of the bag b. We call an instance a *witness* if it contributes to the likelihood $P(Y_b|\mathbf{F}_b)$. Note that $\sum_i p_{b,i} = 1$, and $p_{b,i} \geq 0$ for all $i \in b$. In the MIL formalism, as $P(Y_b|\mathbf{F}_b)$ is solely dependent on the most likely positive instance, it is ideal to put all the probability mass to a single instance as:

$$p_{b,i} = \begin{cases} 1 & \text{if } i = \arg\max_j f_{b,j} \\ 0 & \text{otherwise} \end{cases} \qquad (16)$$

Alternatively, it is also possible to define a soft witness assignment[4] using a sigmoid function:

$$p_{b,i} = \frac{\exp(\lambda f_{b,i})}{\sum_{j \in b} \exp(\lambda f_{b,j})}, \qquad (17)$$

where λ is the parameter that controls the smoothness of the assignment.

Once $\mathbf{P}_b$ is given, we then define the likelihood as a sigmoid of the weighted sum of f_i's with weights p_i's:

$$P(Y_b|\mathbf{F}_b, \mathbf{P}_b) = \frac{1}{1 + \exp(-Y_b \sum_i p_{b,i} f_{b,i})}. \qquad (18)$$

The aim here is to replace the *max* or the *soft-max* function in the original derivation by the *expectation*, $\sum_i p_{b,i} f_{b,i}$, a linear function of $\mathbf{F}_b$ given the witness assignment $\mathbf{P}_b$. Notice that given $\mathbf{P}_b$, the negative log-likelihood of (18) is a convex function of $\mathbf{F}_b$.

In the full Bayesian treatment, one marginalizes out $\mathbf{P}_b$, namely $P(Y_b|\mathbf{F}_b) = \int_{\mathbf{P}_b} P(Y_b|\mathbf{F}_b, \mathbf{P}_b)P(\mathbf{P}_b|\mathbf{F}_b)$, where $P(\mathbf{P}_b|\mathbf{F}_b)$ is a Dirac's delta function with the point support given as (16) or (17). However, this simply leads to the very non-convexity raised by the original version of our GPMIL. Rather we pursue the

[4] This has a close relation to (Gehler & Chapelle, 2007)'s deterministic annealing approach to SVM. Similar to (Gehler & Chapelle, 2007), one can also consider a scheduled annealing, where the inverse of the smoothness parameter λ in (17) serves as the annealing temperature. See Sec. 3.1 for further details.

coordinate-wise convex optimization by separating the process of approximating $P(Y_b|\mathbf{F}_b)$ into two individual steps: (i) find the witness indicator $\mathbf{P}_b$ from $\mathbf{F}_b$ using (16) or (17), and (ii) (while fixing $\mathbf{P}_b$) represent the likelihood as the sigmoid of the weighted sum (18), and perform posterior approximation. We alternate these two steps until convergence. Note that in this setting the Laplace approximation becomes quite similar to that of the standard GP classification, having the additional alternating optimization as an inner loop.

3.1. Deterministic Annealing

When we adopt the soft witness assignment in the above formulation, it is easy to see that (17) is very similar to the probability assignment in the deterministic annealing (i.e., Eq. (11) of (Gehler & Chapelle, 2007)) while the smoothness parameter λ now acts as the inverse temperature in the annealing schedule. Motivated by this, we can have an annealed version of posterior approximation. More specifically, it initially begins with a small λ (large temperature) corresponding to a uniform-like $\mathbf{P}_b$, and repeats the following: perform a posterior approximation starting from the optimum $\mathbf{F}_b$ in the previous stage to get a new $\mathbf{F}_b$, then increase λ to reduce the entropy of $\mathbf{P}_b$.

4. Empirical Results

We conducted experimental evaluation in both synthetic data and real-world benchmark datasets including the traditional MUSK datasets (Dietterich et al., 1997), image annotation and text classification datasets. We ran two different approximation schemes for our GPMIL, which are denoted by: (a) SOFT-MAX = the soft-max approximation with the PSD projection described in Sec. 2, and (b) WDA = the approximation using the witness indicator variables with the deterministic annealing optimization discussed in Sec. 3. In the SOFT-MAX, the GP inference/learning optimization is done by (scaled) conjugate gradient search with different starting iterates. In the WDA, we started with a large temperature (e.g., $\lambda = 1e{-}1$), and decreased it in log-scale (e.g., $\lambda \leftarrow 10 \cdot \lambda$) until there is no significant change in the quantities to be estimated. For both methods, we first estimated the kernel hyperparameters by empirical Bayes (i.e., maximizing the evidence likelihood) and used the learned hyperparameters for the test prediction.

4.1. Synthetic Data

This section tested the effectiveness of GPMIL in accurately estimating the kernel hyperparameters from data. We constructed the synthetic 1D dataset generated by a GP prior with random formation of the bags. More specifically, we first sampled the input data points $\mathbf{x}$ uniformly from the real line $[-30, 30]$. We generated 1000 samples and sampled the latent variables f from the GP prior distribution with the covariance matrix set equal to the (1000×1000) kernel matrix from the input samples. The kernel had a particular form, the RBF kernel $k(\mathbf{x}, \mathbf{x}') = \exp(-||\mathbf{x} - \mathbf{x}'||^2/2\sigma^2)$, where the hyperparameter is set to $\sigma = 3.0$. Note that we assumed that the RBF kernel form is known to the algorithm, and the goal is to estimate σ as accurately as possible. The actual instance-level class output y is determined as $y = \texttt{sign}(f)$. Fig. 2 depicts the instance-level input and output samples (i.e., f (and y) vs. $\mathbf{x}$).

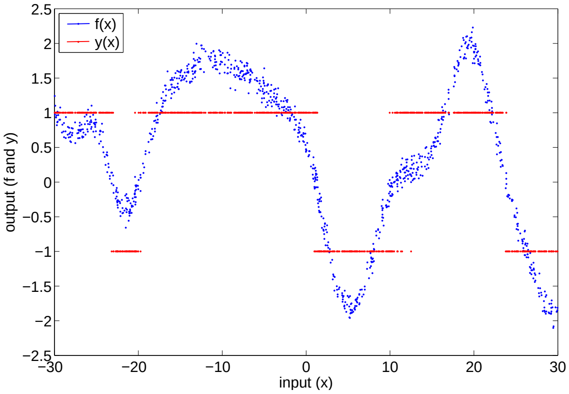

Figure 2. Visualization of the synthetic 1D dataset. It depicts the instance-level input and output samples, where the bag formation is done by randomly grouping the instances. See text for details.

To form the bags, we did the following procedure. For each bag b, we randomly assigned its bag label Y_b uniformly from $\{+1, -1\}$. The number of instances n_b was also chosen uniformly at random from $\{1, \ldots, 10\}$. When $Y_b = -1$, we randomly selected n_b instances from the negative instances of the 1000-sample pool. On the other hand, when $Y_b = +1$, we flipped the 10-side fair coin to decide the positive instance portion $pp \in \{0.1, 0.2, \ldots, 1.0\}$, with which the bag is constructed from $\lceil pp \times n_b \rceil$ instances selected randomly from the positive instances and the rest (also randomly) from the negative instance pool. We generated 100 bags and randomly repeated this process 20 times.

We then performed the GPMIL hyperparameter learning starting from the initial guess $\sigma = 1.0$. We computed the average σ estimated for 20 trials. The results were: $\mathbf{3.2038} \pm \mathbf{0.2700}$ for the SOFT-MAX approach, and $\mathbf{3.0513} \pm \mathbf{0.2149}$ for the WDA approach, which are very close to the true $\sigma = 3.0$. This demonstrates a unique benefit of our GPMIL algorithm, namely that we can estimate the kernel parameters precisely in a principled manner, which is hard to achieve with most existing MIL approaches that rely on heuristic grid search on the hyperparameter space.

4.2. Benchmark Datasets

4.2.1. The MUSK Datasets

The MUSK datasets (Dietterich et al., 1997) have widely served as the benchmark dataset for the MIL algorithms. The datasets contain the description of molecules using multiple low-energy conformations. The feature vector $\mathbf{x}$ is 166-dimensional. There are two different types of bag formation denoted by MUSK1 and MUSK2, where the MUSK1 has approximately $n_b = 6$ conformations (instances) per bag, while the MUSK2 takes $n_b = 60$ instances per bag on average. For comparison with existing MIL algorithms, we followed an experimental setting similar to that of (Andrews et al., 2003; Gehler & Chapelle, 2007), where we conducted 10-fold cross validation. This is further repeated 5 times with different (random) partitions, and the average errors are reported. We used the RBF kernel. The test errors are shown in Table 1. Our GPMIL with `SOFT-MAX` and `WDA` are depicted with and without parentheses, respectively.

In the table, our approaches are also compared with: EMDD=(Zhang et al., 2002), MICA=(Mangasarian & Wild, 2008), and MI-SVM/mi-SVM=(Andrews et al., 2003) as described in the introduction. In addition, we compared GPMIL with the recent approach of (Gehler & Chapelle, 2007) that extend several different SVM variants by deterministic annealing optimization. They include: AL-SVM = Extension of mi-SVM, AW-SVM = Extension of witnesses-identifying SVMs such as MI-SVM and MICA, and ALP-SVM = AL-SVM with the additional constraint on the expected number of positive instances per bag.

Our GPMIL algorithms, for both approximation strategies `WDA` and `SOFT-MAX`, exhibit superior classification performance to existing approaches for the two MUSK datasets. One exception is the MICA where the reported error is the smallest on the MUSK2 dataset. This can be mainly due to the use of L1-regularizer in the MICA that yields a sparse solution suitable for the large-scale MUSK2 dataset. As is also alluded in (Gehler & Chapelle, 2007), it may not be directly comparable with the other methods.

4.2.2. Image Annotation

This section reports experiments on image annotation datasets devised by (Andrews et al., 2003) using the COREL image database. Each image is treated as a bag comprised of the segments (instances) that are represented as feature vectors of color, text, and shape descriptors. Three datasets were formed for the object categories: tiger, elephant, and fox, regarding images containing the object as positive, and the rest as negative. There were 100/100 positive/negative bags, each of which contains $2 \sim 13$ instances. Similar to (Andrews et al., 2003; Gehler & Chapelle, 2007), we conducted 10-fold cross validation. This is further repeated 5 times with different (random) partitions. We employed the RBF kernel. Table 1 shows the test errors. The proposed GPMIL algorithms achieved significantly higher accuracy than the best competing approaches most of the time. Comparing the two approximation methods for GPMIL, `WDA` often outperformed `SOFT-MAX`, implying that the approximation based on witness variables followed by a proper deterministic annealing schedule can be more effective than the softmax approximation with the spectral convexification.

4.2.3. Text Classification

We also demonstrated the effectiveness of the GPMIL algorithm on the text categorization task. We used the MIL datasets provided by (Andrews et al., 2003) obtained from the well-known TREC9 database. The original data is composed of 54000 MEDLINE documents annotated with 4903 subject terms, each defining a binary concept. Each document (bag) is decomposed into passages (instances) of overlapping windows of 50 or fewer words. Similar to the settings in (Andrews et al., 2003), a smaller subset is used, which is comprised of 7 concepts (binary classification problems), each of which has roughly the same number (about 1600) of positive/negative instances from 200/200 positive/negative bags.

In Table 2 we report the average test errors of the GPMIL with the `WDA` approach, together with those of competing models from (Andrews et al., 2003). For MI-SVM and mi-SVM, only the linear SVM errors are shown since the linear kernel outperforms polynomial/RBF kernels most of the time. In the GPMIL we also employed the linear kernel. We see that for a large portion of the problem sets, our GPMIL exhibits a significant improvement over the methods provided in the original paper (EM-DD, mi-SVM, and MI-SVM).

5. Conclusion

This paper proposes GPMIL, a new model that allows incorporating bag class likelihood models into the GP framework, yielding nonparametric probabilistic models that can capture the underlying generative process of MIL. Using GPMIL, the kernel hyperparameters can

Table 1. Test errors on MUSK and Image Annotation Datasets. For GPMIL, we report the errors of `WDA` (without parentheses) and `SOFT-MAX` (with parentheses). In AW-SVM and AL-SVM, for the two annealing schedules suggested by (Gehler & Chapelle, 2007), we only show the ones with smaller errors. Boldfaced numbers indicate the best results.

Dataset	GPMIL	EMDD	MI-SVM	MICA	AW-SVM	mi-SVM	AL-SVM	ALP-SVM
MUSK1	**10.53** (11.48)	15.2	22.1	15.6	14.3	12.6	14.3	13.7
MUSK2	12.75 (12.13)	15.1	15.7	**9.5**	16.2	16.4	13.8	13.8
TIGER	**12.63** (12.86)	27.9	16.0	18.0	17.0	21.6	21.5	14.0
ELEPHANT	**16.20** (17.13)	21.7	18.6	17.5	18.0	17.8	20.5	16.5
FOX	34.25 (36.80)	43.9	42.2	38.0	36.5	41.8	36.5	**34.0**

Table 2. Test errors on text classification. Boldfaced numbers indicate the best results.

Dataset	GPMIL	EMDD	MI-SVM	mi-SVM
TST1	**5.57**	14.2	6.1	6.4
TST2	**14.67**	16.0	15.5	21.8
TST3	13.88	31.0	17.8	**13.0**
TST4	**14.71**	19.5	17.6	17.2
TST7	19.69	24.6	22.0	**18.7**
TST9	**29.20**	34.5	39.8	32.5
TST10	**19.58**	21.5	20.5	20.4

be learned in a principled manner, thus avoiding grid search and being able to exploit a variety of kernel families with complex forms. To address the intractability of exact GP inference and learning, we have suggested several approximation schemes including softmax with a PSD projection and the witness latent variables that can be optimized by deterministic annealing. For many benchmark MIL datasets, we have demonstrated that the proposed methods can yield superior prediction performance than existing state-of-the-art approaches. In our future work, we will consider different approximation algorithms to further improve both accuracy and computational efficiency.

References

Andrews, Stuart, Tsochantaridis, Ioannis, and Hofmann, Thomas. Support vector machines for multiple-instance learning, 2003. NIPS.

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

Friedman, J. Greedy function approximation: a gradient boosting machine, 1999. Technical Report, Dept. of Statistics, Stanford University.

Gärtner, Thomas, Flach, Peter A., Kowalczyk, Adam, and Smola, Alex J. Multi-instance kernels, 2002. ICML.

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

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

Maron, Oded and Lozano-Perez, Tomas. A framework for multiple-instance learning, 1998. NIPS.

Nguyen, N. H., Torresani, L., de la Torre, F., and Rother, C. Weakly supervised discriminative localization and classification: a joint learning process, 2009. ICCV.

Rasmussen, Carl Edward and Williams, Christopher K. I. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Raykar, Vikas C., Krishnapuram, Balaji, Bi, Jinbo, Dundar, Murat, and Rao, R. Bharat. Bayesian multiple instance learning: Automatic feature selection and inductive transfer, 2008. ICML.

Tao, Qingping, Scott, Stephen, Vinodchandran, N. V., and Osugi, Thomas Takeo. SVM-based generalized multiple-instance learning via approximate box counting, 2004. ICML.

Viola, Paul A., Platt, John C., and Zhang, Cha. Multiple instance boosting for object detection, 2005. NIPS.

Wang, Jun and Zucker, Jean-Daniel. Solving the multiple-instance problem: A lazy learning approach, 2000. ICML.

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