A Unifying Framework for Learning Bag Labels from Generalized Multiple-Instance Data

Gary Doran, Andrew Latham, Soumya Ray
Department of Electrical Engineering and Computer Science
Case Western Reserve University
{gary.doran,andrew.latham,sray}@case.edu

Abstract

We study the problem of bag-level classification from generalized multiple-instance (GMI) data. GMI learning is an extension of the popular multiple-instance setting. In GMI data, bags are labeled positive if they contain instances of certain types, and avoid instances of other types. For example, an image of a "sunny beach" should contain sand and sea, but not clouds. We formulate a novel generative process for the GMI setting in which bags are distributions over instances. In this model, we show that a broad class of distribution-distance kernels is sufficient to represent arbitrary GMI concepts. Further, we show that a variety of previously proposed kernel approaches to the standard MI and GMI settings can be unified under the distribution kernel framework. We perform an extensive empirical study which indicates that the family of distribution distance kernels is accurate for a wide variety of real-world MI and GMI tasks as well as efficient when compared to a large set of baselines. Our theoretical and empirical results indicate that distribution-distance kernels can serve as a unifying framework for learning bag labels from GMI (and therefore MI) problems.

1 Introduction

Many real-world problem domains require learning from structured data. For example, consider the content-based image retrieval (CBIR) domain, in which the goal is to retrieve images that contain some object or scene of interest [Maron and Ratan, 1998]. If the image is segmented, then the presence of an object within the image corresponds to the presence of certain classes of segments within the image. For such problems, the multiple-instance (MI) setting offers a richer representation for these structured objects as sets, or "bags," of feature vectors, each of which is called an "instance" [Dietterich *et al.*, 1997]. In CBIR, an image is a bag of segments, each of which can be described with a feature vector. The MI setting can also be applied to other problems with structured objects, such as text categorization, audio classification, or drug discovery.

In the "standard" MI setting, it is assumed that a bag is positive if it contains at least one instance from the positive class of instances. This makes sense for many CBIR problems; for example, an image contains an apple if *some* segment corresponds to an apple. However, prior work has observed that a more complex relationship is required for many real-world problems. Consider a CBIR problem in which the task is to distinguish pictures of deserts, oceans, and beaches [Foulds and Frank, 2010]. Segments in these images are primarily either sand or water. However, in this case, the presence of both sand *and* water is required to distinguish beaches from deserts (only sand) or oceans (only water). Similarly, Wang *et al.* [2004] study the problem of identifying members of the Thioredoxin-fold "superfamily" of proteins. A sequence of amino acids corresponding to a protein (bag) is represented using properties of subsequences (instances) surrounding a central "motif" within the protein sequence. To be a member of this superfamily, the sequence must contain certain subsequences and exclude others. Such concepts can be learned under the generalized MI (GMI) framework [Weidmann *et al.*, 2003; Scott *et al.*, 2005]. Here, we assume that some types of instances are "attractive" and that others are "repulsive." For a bag to be positive, it must contain a certain number of attractive instance types and exclude some number of the repulsive instance types. The generalization allows for richer relationships between bag and instance labels than the standard MI setting, which is a special case of GMI as follows: there is one attractive type, no repulsive types, and a bag is labeled positive if it has at least one instance from the attractive type.

Many existing supervised learning approaches such as decision trees [Blockeel *et al.*, 2005] and support vector machines (SVMs) [Andrews *et al.*, 2003] have been extended to standard MI learning, but none of these approaches are directly applicable to the GMI setting. Furthermore, many of these approaches extend instance-based approaches, and therefore use instance-level hypothesis classes to label bags. An alternative, successfully employed by other prior work [Gärtner *et al.*, 2002; Chen *et al.*, 2006; Foulds, 2008; Zhou *et al.*, 2009; Amores, 2013a], is to explicitly or implicitly construct a feature vector representation for bags and use a standard supervised classifier to solve the bag-labeling task. In fact, these bag-level classifiers often outperform their instance-level counterparts in practice in terms of both accu-

racy and efficiency on the bag-labeling task [Amores, 2013b; Doran and Ray, 2013; Cheplygina *et al.*, 2015]. Given this observation, we are interested in the question: how can we appropriately represent bags so that standard supervised approaches can learn *GMI* concepts?

Kernel methods are a well-studied set of approaches for implicitly constructing feature vector representations of arbitrary objects. Given a set of objects $\mathcal{X}$, a positive-definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a "kernel function" that is implicitly associated with a "feature map" $\phi : \mathcal{X} \to \mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$. Given the uncountably infinite family of potential kernel functions, how can we choose a class of kernel functions that are appropriate for the GMI setting? Some of the first work on MI kernels showed that a family of *set kernels* could be used to represent standard MI concepts [Gärtner *et al.*, 2002]. However, the desirable properties of such kernels hold only under the standard MI assumption in which a bag's label is the logical disjunction of its instances' labels. Subsequent work devised a specialized kernel for the GMI setting [Tao *et al.*, 2004; 2008]. In our work, we propose a new formal generative model that describes GMI concepts in a setting where bags are *distributions* over instances. We show that in this generative model, a family of distribution-distance kernels can be successfully applied to the GMI setting. In particular, we show that a universal kernel based on the maximum mean discrepancy (MMD) distance between distributions is sufficient for representing GMI concepts given some weak assumptions [Christmann and Steinwart, 2010].

Next, we show that some existing bag-level representations of GMI data can be viewed as approximations of certain distribution-distance or distribution-embedding kernels. Prior work [Amores, 2013b] proposed a taxonomy of MI learning algorithms that included categories for "bag-space" approaches that implicitly map bags into a feature space via kernels or distance metrics, and "embedded-space" approaches that first construct feature vectors from bags before applying supervised learning techniques. By showing that some of these embedded-space approaches are actually approximations to bag-space approaches, our work implies that these two separate categories can be unified into a single framework for understanding bag-level representations.

Finally, we empirically evaluate the performance in terms of both accuracy and efficiency of distribution kernels as well as existing bag-level representations across 72 datasets from a variety of real-world domains. We show that distribution-based kernels yield the best performance in terms of these two metrics. Thus, in addition to the theoretical understanding provided by our results, we also provide practical recommendations regarding the use of distribution-based kernels for a wide variety of MI and GMI problems.

2 The GMI Generative Model

In this section, we describe the generative process for GMI data. Although the original work on standard MI learning leaves the precise generative process unspecified [Dietterich *et al.*, 1997], subsequent theoretical analyses either assume the bags are independent and identically distributed (IID) samples from a *single* distribution across bags [Blum and Kalai, 1998], or are drawn from an arbitrary distribution over tuples of instances [Sabato and Tishby, 2012]. Our approach builds on a more recently proposed generative model [Doran and Ray, 2014] that assumes that bags correspond to distinct distributions over instances. This model makes fewer assumptions about the generative process than prior work. It is motivated by domains such as 3-dimensional Quantitative Structure–Activity Relationship (3D-QSAR), where a molecule exists in a dynamic equilibrium whose distribution over shapes is governed by the energy associated with each shape, or text categorization, for which existing successful models such as topic models already treat documents as topic-specific distributions over words or passages [Blei *et al.*, 2003]. We extend this distribution-based model of MI data to the GMI setting.

Figure 1: An example GMI concept of "sunny beach" has attractive types `sand` and `water`, and a repulsive type `cloud`. The desert and ocean images (**top**) only contain one of the two required attractive types. The cloudy beach image (**lower left**) is not a member of the concept because it contains the repulsive type `cloud`. The sunny beach image (**lower right**) satisfies the definition of the concept.

Formally, let $\mathcal{X}$ be the space of instances, and $\mathscr{P}(\cdot)$ denote the space of Borel probability measures over its argument. Then $\mathscr{P}(\mathcal{X})$, the space of probability measures over instances, is the space of bags. Hence, each bag B corresponds to a specific distribution over instances, denoted $\Pr(x \mid B)$. Bags themselves are drawn from some fixed distribution *over bags*, which is given by $D_{\mathcal{B}} \in \mathscr{P}\big(\mathscr{P}(\mathcal{X})\big)$. After being sampled, these bags are labeled by a function $F : \mathscr{P}(\mathcal{X}) \to \{-1, +1\}$. In GMI learning, it is assumed that some types of instances are "attractive" and that others are "repulsive." For a bag to be labeled positive by F, it must contain or "hit" a certain number of attractive instance types and exclude or "miss" some number of the repulsive instance types. Figure 1 shows a GMI concept "sunny beach," for which the attractive types are `sand` and `water`, and a repulsive type is `cloud`. Let $\mathcal{A} = \{A_i\}_{i=1}^{a}$ denote a set of attractive types, and $\mathcal{R} = \{R_i\}_{i=1}^{r}$ be a set of repulsive types, all disjoint closed subsets of $\mathcal{X}$. The set of "other" instances in the support of the instance distribution $D_{\mathcal{X}}$ is defined as $\mathcal{O} \triangleq \operatorname{supp} D_{\mathcal{X}} - \bigcup_{C_i \in \mathcal{A} \cup \mathcal{R}} C_i$, assumed to be a closed set.

Note that this is equivalent to the assumption that there is a continuous function that defines the boundary with separation between each of the types of instances. Formally, we say that a bag B hits a type $C_i \in \mathcal{A} \cup \mathcal{R}$ if $\int_{C_i} \mathrm{d}\,\mathrm{P}(x \mid B) \geq \pi_i$ for threshold π_i and misses C_i if $\int_{C_i} \mathrm{d}\,\mathrm{P}(x \mid B) = 0$. That is, there must be probability π_i of sampling an instance from type C_i within a bag to hit the type, or zero probability to miss the type. Given this notion, we can extend the distribution-based generative process to the GMI setting:

Definition 1 (GMI-GEN)**.** *Let $\Pi_A = \{\pi_i\}_{i=1}^{a}$ and $\Pi_R = \{\pi_i\}_{i=1}^{r}$ be sets of positive type-specific threshold parameters for the attractive and repulsive types. Then* $\text{GMI-GEN}(\mathcal{A}, \mathcal{R}, \Pi_A, \Pi_R, \alpha, \rho)$ *is the set of $(D_{\mathcal{B}}, F)$ s.t.:*

1. *The support of the instance distribution defined by $D_{\mathcal{X}} \triangleq \int B \,\mathrm{d}\,\mathrm{Pr}_{D_{\mathcal{B}}}(B)$ is a compact set.*
2. *For each bag B in the support of $D_{\mathcal{B}}$, and each type in $\mathcal{A} \cup \mathcal{R}$, B either hits or misses the type.*
3. *A bag is positive if and only if it hits at least α of the types in $\mathcal{A}$ and misses at least ρ of the types in $\mathcal{R}$.*

Condition 3 allows for flexibility in using the attractive and repulsive types to define classes. To illustrate, consider a variant of our "sunny beach" example called "nice beach" with an extra repulsive type `trash`. A beach is nice if it does not have *both* `trash` and `cloud`, so selecting $\rho = 1$ with these two types can be used to express this condition.

An assumption of the standard MI setting is that there also exists some instance-labeling function $f : \mathcal{X} \to \{-1, +1\}$, which has the relationship with F that F should label a bag $B \in \mathscr{P}(\mathcal{X})$ negative if the labels of the instances sampled from B are *almost surely* negative, and positive if there is some nonzero probability π of sampling a positive instance within the bag. Note that given an instance-labeling function f and $\mathcal{X}_+ = \{x \in \mathcal{X} : f(x) = +1\}$, the $(D_{\mathcal{B}}, F)$ in $\text{GMI-GEN}(\{\mathcal{X}_+\}, \varnothing, \{\pi\}, \varnothing, 1, 0)$ correspond to the standard MI concepts. Hence, GMI-GEN is indeed a generalization of the standard MI setting. Having described possible generative processes for GMI data, we can begin to discuss kernel methods for learning from data generated in the manner described in Definition 1.

3 Learning GMI-GEN Concepts with Kernels

We first show that certain classes of distribution kernels are sufficient to represent concepts in GMI-GEN. Then, we show that existing bag kernels and feature space embeddings can be viewed under one unifying framework provided by our generative process.

3.1 Distribution Kernels for GMI Data

Given our generative model where bags are distributions, we are interested in the application of kernels constructed for *learning from distributions*, which have been explored by prior work [Muandet *et al.*, 2012]. One particularly successful kernel-based representation for distributions is the "mean embedding." Given an instance space $\mathcal{X}$ and instance kernel k with feature map ϕ, the mean embedding of a distribution $\mathcal{P} \in \mathscr{P}(\mathcal{X})$ is given by $\mu(\mathcal{P}) \triangleq \mathrm{E}_{x \sim \mathcal{P}}[\phi(x)]$. The associated "mean embedding kernel" K defined on distributions is given by $K(\mathcal{P}, \mathcal{Q}) \triangleq \mathrm{E}_{(x,x') \sim \mathcal{P} \times \mathcal{Q}}[k(x, x')]$ for $\mathcal{P}, \mathcal{Q} \in \mathscr{P}(\mathcal{X})$.

The mean embedding has some desirable theoretical properties. First, given a sample $X \sim \mathcal{P}^n$, the *empirical* mean embedding $\widehat{\mu}(X) \triangleq \frac{1}{n} \sum_{x_i \in X} \phi(x_i)$ converges quickly to the underlying embedding $\mu(\mathcal{P})$ as sample size n increases [Smola *et al.*, 2007]. Furthermore, whenever the instance kernel k is *characteristic*, the mean embedding μ is an injective mapping of distributions into the kernel feature space, or reproducing kernel Hilbert space (RKHS), $\mathcal{H}$ [Sriperumbudur *et al.*, 2010]. This means that distinct distributions will have unique feature representations. A kernel is characteristic if it is *universal*, meaning that its RKHS $\mathcal{H}$, interpreted as a space of functions over $\mathcal{X}$, is uniformly dense in $C(\mathcal{X})$, the space of bounded, continuous functions over $\mathcal{X}$ [Micchelli *et al.*, 2006]. The commonly used radial basis function (RBF) kernel, $k(x, x') = e^{-\gamma \|x - x'\|^2}$, is universal.

Prior work describes the class of functions over distributions that can be represented using the mean embedding kernel [Muandet *et al.*, 2012]. In particular, the RKHS of the mean embedding kernel with some universal instance kernel is dense in the set:

$$\mathcal{F} = \left\{ \mathcal{P} \mapsto \int_{\mathcal{X}} g \,\mathrm{d}\mathcal{P} : \mathcal{P} \in \mathscr{P}(\mathcal{X}),\, g \in C(\mathcal{X}) \right\}. \qquad (1)$$

These are the functions we get by taking the expected value of a fixed but arbitrary continuous function with respect to probability distributions. However, the function class $\mathcal{F}$ is a strict subset of $C(\mathscr{P}(\mathcal{X}))$, the set of all bounded, continuous functions over the set of probability distributions, with respect to the weak topology on $\mathscr{P}(\mathcal{X})$ [Muandet *et al.*, 2012]. Thus, the mean embedding defined in terms of a universal kernel with respect to $C(\mathcal{X})$ is not itself universal with respect to $C(\mathscr{P}(\mathcal{X}))$.

MMD. However, as shown by Christmann and Steinwart [2010], it *is* possible to construct a universal kernel with respect to $C(\mathscr{P}(\mathcal{X}))$ using an additional level of embedding. That is, using the RBF kernel defined with respect to the mean embeddings of two distributions $\mathcal{P}$ and $\mathcal{Q}$, $k(\mathcal{P}, \mathcal{Q}) = e^{-\gamma \|\mu(\mathcal{P}) - \mu(\mathcal{Q})\|_{\mathcal{H}}^2}$, is universal with respect to $C(\mathscr{P}(\mathcal{X}))$ when μ is injective and $\mathcal{X}$ is compact. Note that this kernel is equivalent in form to the RBF kernel but treats $\mathcal{H}$ rather than $\mathcal{X}$ as the input space. Since the quantity $\|\mu(\mathcal{P}) - \mu(\mathcal{Q})\|_{\mathcal{H}}$ is known as the MMD [Smola *et al.*, 2007], we refer to this as the MMD kernel. This iterated embedding is also called the "level-2" embedding by prior work [Muandet *et al.*, 2012].

Are MMD distribution kernels expressive enough to represent bag-level GMI concepts from Definition 1? Our first result answers this question affirmatively.

Proposition 1. *Let $(D_{\mathcal{B}}, F)$ be an element of* $\text{GMI-GEN}(\mathcal{A}, \mathcal{R}, \Pi_A, \Pi_R, \alpha, \rho)$*. Then the universal MMD kernel can arbitrarily approximate in the uniform norm a function that separates bags according to F.*

Proof. Since a universal kernel can arbitrarily approximate continuous functions, it suffices to show that a GMI concept F is separable with a continuous function.

Taking for granted that $\mathcal{X}$ is a normal space,[1] by Urysohn's Lemma, there exists a continuous function $h_i : \mathcal{X} \to [0, 1]$ for each type C_i such that $x \in C_i \implies h_i(x) = 1$ and $x \notin C_i \wedge x \in \operatorname{supp} D_{\mathcal{X}} \implies h_i(x) = 0$. This follows from the assumption that all sets of instance types and their complements within $\operatorname{supp} D_{\mathcal{X}}$ are disjoint closed sets.

Given the existence of the h_i, the function $H_i(B) = \min\left\{1, \frac{1}{\pi_i}\int_{\mathcal{X}} h_i \,\mathrm{d}\,\mathrm{P}(x \mid B)\right\}$ is 1 if B hits type C_i and 0 if B misses C_i. Furthermore, each H_i is a continuous function over bags since it is a composition of continuous functions.

Then, for a GMI concept F as described in Definition 1, the following concept separates F: $G(B) = 2\min\left\{\sum_{i=1}^{a} H_i(B) - \alpha, \sum_{i=a+1}^{a+r}(1 - H_i(B)) - \rho\right\} + 1$. The minimum is at least 1 if the number of hits and misses are both above their respective thresholds α and ρ. Thus, $G(B) \geq 1$ for positive bags according to F, and $G(B) \leq -1$ for negative bags.

Finally, note that G is a composition of the continuous $\min$ function with continuous functions over bags, so it is in fact continuous and can be approximated by an element of the RKHS of the MMD kernel. □

A corollary of this result is that the mean embedding kernel suffices to represent standard MI concepts. Recall that standard MI concepts are of the form $\text{GMI-GEN}(\{\mathcal{X}_+\}, \varnothing, \{\pi\}, \varnothing, 1, 0)$ for some $\pi > 0$ and $\mathcal{X}_+ = \{x \in \mathcal{X} : f(x) = +1\}$, where f is an instance-labeling function. Then we have:

Corollary 1. *Let $(D_{\mathcal{B}}, F)$ be a standard MI concept, an element of* $\text{GMI-GEN}(\{\mathcal{X}_+\}, \varnothing, \{\pi\}, \varnothing, 1, 0)$. *Then the mean embedding kernel with a universal instance kernel can separate bags with respect to F.*

Proof. By Proposition 1, there is a single function H_+ such that a bag separating function is given by:

$$\begin{aligned} G(B) &= 2\,(H_+(B) - 1) + 1 = 2H_+(B) - 1 \\ &= 2\min\left\{1, \tfrac{1}{\pi}\int h_+ \,\mathrm{d}\,\mathrm{P}(B)\right\} - 1 \\ &= \min\left\{1, \tfrac{2}{\pi}\int h_+ \,\mathrm{d}\,\mathrm{P}(B) - 1\right\}. \end{aligned}$$

Hence, it follows that the function $G'(B) = \frac{2}{\pi}\int h_+ \,\mathrm{d}\,\mathrm{P}(B) - 1$ also separates bags.

Now, given that $G'(B)$ is a linear rescaling of the function $\int h_+ \,\mathrm{d}\,\mathrm{P}(B)$, where h_+ is a continuous function, it can be approximated by an element of the RKHS of the mean embedding kernel. This fact follows from the results of [Muandet *et al.*, 2012] as shown in Equation 1. □

The results above show that the mean embedding and MMD kernels can learn MI and GMI concepts. Note that these results do not require prior knowledge of the π parameters, only that they exist. These results assume that the entire bag distribution B_i is known during training. In practice, only samples of instances are observed for each bag. However, since empirical estimates of mean embeddings converge quickly to the underlying embeddings as the sample size within each bag increases [Sriperumbudur *et al.*, 2010], we argue (and empirically show) that these approaches will also work well in practice.

[1] A topological space is normal if for any disjoint closed sets A and B in the space, there are disjoint open sets U and V containing A and B, respectively [Folland, 1999].

3.2 Relationship of Prior Work to the MMD

In prior work, some other bag-level kernels have been defined for solving GMI classification problems. In this section, we discuss these approaches and their relationship to the mean embedding and MMD kernels. In particular we show that these approaches are either special cases, or approximations of the MMD, or learn hypotheses that are a proper subset of those representable by the MMD. Thus, although these approaches were placed in two distinct categories by prior work [Amores, 2013b], we argue that they can be unified under our proposed generative model.

Box-Counting Kernel. The box-counting kernel is motivated by the assumption that "attractive" and "repulsive" types of points as described in Definition 1 are contained within axis-parallel boxes in the feature space [Tao *et al.*, 2004; 2008]. This is a stronger assumption than is made in Proposition 1, which allows these sets of types to be arbitrary closed sets. Hence, the hypothesis space of the MMD kernel subsumes that of the box-counting kernel. The box-counting kernel constructs a Boolean feature corresponding to *every* axis-parallel box in a discretized version of the feature space. A feature has value `1` if the corresponding box contains a point in the bag and value `0` otherwise.

Because it is intractable to explicitly enumerate all such features, Tao *et al.* [2004] use a "box-counting" kernel by observing that the inner product between two Boolean feature vectors above will be equal to the number of boxes that contain points from both corresponding bags. However, the box-counting problem is #P-complete, so an approximation is used to make even the kernel computation tractable [Tao *et al.*, 2004]. The approximation scheme finds a value within a factor of ϵ of the true count with probability $1 - \delta$, in $\mathsf{poly}(m_u, k, \frac{1}{\epsilon}, \frac{1}{\delta})$ time, where m is the bag size and k is the dimensionality of the input feature space. In contrast to this, the MMD kernel can be efficiently computed exactly.

YARDS. Another set of approaches for both MI and GMI learning construct a feature vector representation for each bag. The "yet another radial distance-based similarity measure" (YARDS) approach constructs a representation for bags as follows: First, each instance in a dataset is represented using a feature vector of length $|X|$, where X is the set of all instances in the dataset, with each feature an RBF kernel between that instance and one of the $x_i \in X$. This mapping, $x \mapsto \big[k(x, x_1), \ldots, k(x, x_{|X|})\big]$, where k is the RBF kernel, is called the empirical kernel map [Schölkopf and Smola, 2002], which we denote $\widehat{\phi}(x)$. YARDS then proceeds to represent each bag as the average of these empirical kernel mappings, as in $\widehat{\widehat{\mu}}(B_i) = \frac{1}{|B_i|}\sum_{x_i \in B_i} \widehat{\phi}(x_i)$. We call

this the *doubly* empirical mean embedding, since it is empirical in terms of both the kernel feature map as well as the estimate of the underlying mean embedding. This embedding is equivalent to the implicit kernel feature map with the empirical mean embedding up to a linear rescaling of the features [Schölkopf and Smola, 2002]. Finally, YARDS uses another RBF kernel to embed $\widehat{\widehat{\mu}}$ into a feature space, making it a doubly empirical version of the MMD kernel. Thus, when YARDS is used with a standard SVM and an RBF kernel, it can learn the same concepts as the MMD kernel. However, some practical differences between these approaches are observed in the experiments, likely due to feature rescaling.

NSK. The normalized set kernel (NSK) is an early kernel method proposed for the standard MI setting [Gärtner *et al.*, 2002]. In its most basic form, the set kernel between two bags is formed by a sum of pairwise instance kernel values between all pairs of instances across the two bags. Early work showed empirically that the kernel performs better when it is normalized, for example, by dividing by the bag sizes. In fact, the NSK with "averaging normalization" is equivalent to the empirical mean embedding kernel [Smola *et al.*, 2007]; that is, $k_{\mathrm{NSK}}(X, X') = \frac{1}{|X||X'|}\sum_{x\in X}\sum_{x'\in X'} k(x, x') = \langle\widehat{\mu}(X), \widehat{\mu}(X')\rangle$. The NSK is *complete* for standard MI classification, meaning that when the RKHS of k contains a function that separates instances of different classes, the RKHS of a corresponding k_{NSK} contains a function that separates bags of different classes [Gärtner *et al.*, 2002; Doran and Ray, 2013]. Proposition 1 shows this also holds for the standard MI setting within GMI-GEN.

mi-Graph. The MI-Graph and mi-Graph [Zhou *et al.*, 2009] approaches first construct graphs for each bag by connecting two instances in a bag with an edge if they are within a distance of τ of each other. The parameter τ is chosen heuristically as the average distance between instances in a bag. The corresponding edge is weighted with a normalized reciprocal of the distance between the instances.

Like the NSK, the MI-Graph kernel is a sum of pairwise kernel values between instances *and edges* across two bags and their corresponding graphs based on an instance kernel and a kernel defined on edges. However, since the number of edges in a bag graph grows roughly as the square of the bag size, computing all pairwise edge kernel values is quartic in terms of the bag size. The mi-Graph kernel is a computationally more efficient version of MI-Graph that is equivalent to a weighted version of the mean embedding kernel. Under the view of bags as distributions, mi-Graph can be viewed as performing the mean embedding on a *biased* sample, or a sample drawn from a modified version of the bag's distribution.

EMD. The earth-mover's distance (EMD), also known as the Wasserstein metric, is a popular distance metric commonly used within the CBIR domain [Rubner *et al.*, 2000]. The EMD is a proper distance metric between distributions, and its name comes from an intuitive description of how it operates. If one views one distribution as a pile of dirt, and the other distribution as a hole in the ground, then the EMD is a measure of the minimum amount of work, in terms of mass of dirt times Euclidean distance across the ground traveled, that it takes to fill the hole with the pile. The EMD kernel is formally defined via $k_{\mathrm{EMD}}(B_i, B_j) = e^{-\gamma\,\mathrm{EMD}(B_i,B_j)}$, which is similar to that of the MMD kernel. We hypothesize that because the EMD produces a similar representation of bag distributions to that of the MMD kernel, it can achieve similar performance on MI and GMI tasks.

4 Empirical Evaluation

In this section, we evaluate our hypothesis that the universal MMD kernel can efficiently learn accurate GMI concepts. We evaluate this hypothesis by comparing the MMD kernel to several baselines, described below.

4.1 Methodology

To evaluate our hypothesis, we use 52 existing datasets from 3D-QSAR [Dietterich *et al.*, 1997], CBIR [Andrews *et al.*, 2003; Maron and Ratan, 1998; Rahmani *et al.*, 2005], text categorization [Andrews *et al.*, 2003; Settles *et al.*, 2008], audio classification [Briggs *et al.*, 2012], and TRX protein sequence classification [Wang *et al.*, 2004]. Of these, only the TRX dataset is known to require a GMI concept; thus, we augment our results with 20 semi-synthetic GMI datasets derived from the multi-label natural scenes datasets [Zhou and Zhang, 2006]. For each of the 5 instance types the bags in these datasets can contain (desert, mountains, sea, sunset, or trees), we form 20 datasets in which one of these is the attractive class and another is the repulsive class (e.g., images of mountains with no trees). Some of these datasets should be more difficult for standard MI methods; however, we cannot verify that all of them are *strictly* GMI without instance labels. We use 10-fold stratified cross-validation to evaluate algorithm performance in terms of accuracy, with 5-fold inner cross-validation and random parameter search [Bergstra and Bengio, 2012] used to select parameters. Details of the parameter ranges used for each kernel, full numerical tables of results, and additional figures are available in supplementary material online[2].

We use the method described by [Demšar, 2006] to statistically compare the kernel approaches. We use the Friedman test to reject the null hypothesis that the algorithms perform equally at an $\alpha = 0.001$ significance level, and an $\alpha = 0.05$ significance level for the Nemenyi test and resulting critical difference diagram shown in Figure 2.

4.2 Results: Accuracy

The results in Figure 2 are generally consistent with the theoretical discussion presented above. The NSK and mi-Graph approaches produce very similar representations of the data, and also perform very similarly with no significant difference across datasets. The MMD kernel outperforms YARDS, which offers only an approximation of the same representation. Although the MMD kernel offers greater representational power than the NSK, the performance of these approaches is statistically equivalent across the MI and GMI datasets. On the other hand, restricting the analysis to the semi-synthetic GMI datasets, using the Wilcoxon signed-rank test to perform a pairwise comparison, the MMD does significantly outperform the NSK as expected. Interestingly, the

[2] http://engr.case.edu/ray_soumya/

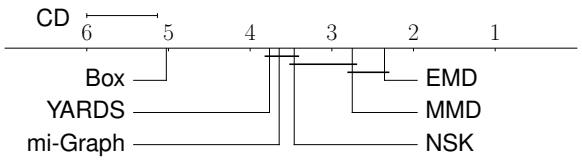

Figure 2: Ranks (lower is better) of the bag kernel approaches on the bag-labeling task. The critical difference diagram shows the average rank of each technique across datasets, with techniques being statistically different at an $\alpha = 0.05$ significance level if the ranks differ by more than the critical difference (CD) indicated above the axis. Thick horizontal lines indicate statistically indistinguishable groups.

Table 1: Complexity of computing bag-level kernel entries, where m denotes bag size, $|X|$ is the number of instances in the dataset, ϵ is an approximation factor for the box-counting kernel, and $1 - \delta$ is the probability of ϵ-approximation.

Technique	Complexity	Technique	Complexity
Box Counting	$O\left(m^2 \frac{1}{\epsilon^2} \log \frac{1}{\delta}\right)$	EMD	$O\left(m^3\right)$
YARDS	$O\left(m\,\lvert X\rvert\right)$	MMD	$O\left(m^2\right)$
mi-Graph	$O\left(m^2\right)$	NSK	$O\left(m^2\right)$

box-counting kernel, designed for the GMI setting, is typically outperformed by other kernels, even on GMI problems.

One somewhat surprising result is that the EMD kernel is slightly better than the MMD kernel, though not significantly so. We conjecture that the explanation for this result is that both the MMD and EMD kernels are actually members of the same family of distribution-distance kernels. Given any distance metric d, a kernel can be constructed using the generalized RBF kernel $k_d(x_i, x_j) = e^{-\gamma d(x_i, x_j)^p}$ [Schölkopf and Smola, 2002]. Both the MMD and EMD kernels are of this generalized form based on distance metrics defined on the space of distributions (or samples from distributions). In fact, both distance metrics induce the same weak topology on the space of distributions over instances, $\mathscr{P}(\mathcal{X})$, when the instance space is separable [Sriperumbudur *et al.*, 2010]. As discussed above, the MMD kernel is universal in its ability to represent continuous functions over distributions. Although it intuitively seems that the EMD kernel should have similar representational abilities, it is still an open question whether the EMD kernel is similarly universal over the space of distributions. However, given the empirical performance of the EMD kernel, we conjecture that this kernel is at least proficient at representing GMI concepts.

4.3 Results: Efficiency

Within the set of bag kernel classifiers, each approach requires differing amounts of time to construct the kernel matrix and train a classifier. The trade-off between training time and accuracy can be informative for selecting a bag kernel to apply in practice. Table 1 shows the time complexity of each algorithm. While the EMD occasionally outperforms the MMD in terms of accuracy, it is computationally more intensive. In practice, the NSK is the fastest algorithm given our implementations. The NSK, MMD, and EMD kernels lie on the Pareto frontier of algorithms ranked by both accuracy and running time (see supplementary materials).

5 Conclusion

In this paper, we have analyzed the GMI and MI settings in a model where bags are distributions over instances. Our theoretical results show that a class of distribution-based kernels are sufficient to represent GMI concepts. Furthermore, we show that many other existing approaches can be viewed as approximations or special cases of this approach. Finally, our empirical results indicate that the most accurate and efficient approaches are either distribution-distance or distribution-embedding kernels. Accordingly, we recommend the use of these distribution-based kernels for a wide variety of MI and GMI problem domains in practice.

Acknowledgements

G. Doran was supported by GAANN grant P200A090265 from the US Department of Education and NSF grant CNS-1035602. S. Ray was partially supported by CWRU award OSA110264. This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University.

References

[Amores, 2013a] J. Amores. MILDE: Multiple instance learning by discriminative embedding. *Knowledge and Information Systems*, 42(2):381–407, 2013.

[Amores, 2013b] J. Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013.

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

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

[Blei *et al.*, 2003] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Blockeel *et al.*, 2005] H. Blockeel, D. Page, and A. Srinivasan. Multi-instance tree learning. In *Proceedings of the International Conference on Machine Learning*, pages 57–64, 2005.

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

[Briggs *et al.*, 2012] F. Briggs, X. Fern, and R. Raich. Rank-loss support instance machines for MIML instance annotation. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 534–542, 2012.

[Chen *et al.*, 2006] Y. Chen, J. Bi, and J. Wang. MILES: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1931–1947, 2006.

[Cheplygina *et al.*, 2015] V. Cheplygina, D. Tax, and M. Loog. Multiple instance learning with bag dissimilarities. *Pattern Recognition*, 48(1):264–275, 2015.

[Christmann and Steinwart, 2010] A. Christmann and I. Steinwart. Universal kernels on non-standard input spaces. pages 406–414, 2010.

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

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

[Doran and Ray, 2013] G. Doran and S. Ray. A theoretical and empirical analysis of support vector machine methods for multiple-instance classification. *Machine Learning Journal*, 2013.

[Doran and Ray, 2014] G. Doran and S. Ray. Learning instance concepts from multiple-instance data with bags as distributions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1802–1808, 2014.

[Folland, 1999] G. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 1999.

[Foulds and Frank, 2010] J. Foulds and E. Frank. A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, 25(01):1–25, 2010.

[Foulds, 2008] J. Foulds. *Learning instance weights in multi-instance learning*. PhD thesis, Univ. of Waikato, 2008.

[Gärtner *et al.*, 2002] T. Gärtner, P. Flach, A. Kowalczyk, and A. Smola. Multi-instance kernels. In *Proceedings of the International Conference on Machine Learning*, pages 179–186, 2002.

[Maron and Ratan, 1998] O. Maron and A. Ratan. Multiple-instance learning for natural scene classification. In *Proceedings of the International Conference on Machine Learning*, pages 341–349, 1998.

[Micchelli *et al.*, 2006] C. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

[Muandet *et al.*, 2012] K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Schölkopf. Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems*, pages 10–18, 2012.

[Rahmani *et al.*, 2005] R. Rahmani, S. Goldman, H. Zhang, J. Krettek, and J. Fritts. Localized content based image retrieval. In *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 227–236, 2005.

[Rubner *et al.*, 2000] Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

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

[Schölkopf and Smola, 2002] B. Schölkopf and A. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

[Scott *et al.*, 2005] S. Scott, J. Zhang, and J. Brown. On generalized multiple-instance learning. *International Journal of Computational Intelligence and Applications*, 5(1):21–35, 2005.

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

[Smola *et al.*, 2007] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

[Sriperumbudur *et al.*, 2010] B. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 99:1517–1561, 2010.

[Tao *et al.*, 2004] Q. Tao, S. Scott, N. Vinodchandran, and T. Osugi. SVM-based generalized multiple-instance learning via approximate box counting. In *Proceedings of the International Conference on Machine Learning*, 2004.

[Tao *et al.*, 2008] Q. Tao, S. Scott, N. Vinodchandran, T. Osugi, and B. Mueller. Kernels for generalized multiple-instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2084–2098, 2008.

[Wang *et al.*, 2004] C. Wang, S. Scott, J. Zhang, Q. Tao, D. Fomenko, and V. Gladyshev. A study in modeling low-conservation protein superfamilies. Technical report, Department of Computer Science, University of Nebraska, 2004.

[Weidmann *et al.*, 2003] N. Weidmann, E. Frank, and B. Pfahringer. A two-level learning method for generalized multi-instance problems. In *Machine Learning: ECML 2003*, pages 468–479. 2003.

[Zhou and Zhang, 2006] Z. Zhou and M. Zhang. Multi-instance multi-label learning with application to scene classification. pages 1609–1616, 2006.

[Zhou *et al.*, 2009] Z. Zhou, Y. Sun, and Y. Li. Multi-instance learning by treating instances as non-IID samples. In *Proceedings of the International Conference on Machine Learning*, pages 1249–1256, 2009.