Mixture Model for Multiple Instance Regression and Applications in Remote Sensing

Zhuang Wang, Liang Lan, and Slobodan Vucetic

***Abstract*—The Multiple Instance Regression (MIR) problem arises when a data set is a collection of bags, where each bag contains multiple instances sharing the identical real-valued label. The goal is to train a regression model that can accurately predict label of an unlabeled bag. Many remote sensing applications can be studied within this setting. We propose a novel probabilistic framework for MIR that represents bag labels with a mixture model. It is based on an assumption that each bag contains the prime instance which is responsible for the bag label. An expectation-maximization algorithm is proposed to maximize the likelihood of the mixture model. The mixture model MIR framework is quite flexible and several existing MIR algorithms can be described as its special cases. The proposed algorithms were evaluated on synthetic data and remote sensing data for aerosol retrieval and crop yield prediction. The results show that the proposed MIR algorithms achieve higher accuracy than the previous state-of-the-art.**

***Index Terms*— multiple instance regression, multiple instance learning, mixture model, expectation maximization, neural networks, aerosol retrieval, MISR, MODIS, remote sensing**

I. Introduction

In Multiple Instance Learning (MIL), a learner is given a number of labeled bags, each containing many instances of the same type. The goal is to train a model that can accurately predict label of an unseen bag given its instances. The main difference from the traditional supervised learning is that labels are assigned to bags instead of individual instances. The difficulty of the MIL problem depends on the type and variability of instances within each bag.

The most commonly addressed MIL problem is Multiple Instance Classification (MIC), where it is assumed that negative bags contain exclusively negative instances, while positive bags contain at least one positive instance [7]. Interestingly, such a setup covers a substantial number of applications such as drug activity prediction [7], image categorization [19] [3] and retrieval [30] [32], text categorization [1], and stock market prediction [18]. A standard approach to solving MIC is to first discover positive instances as those that are the most different from instances in negative bags and to build a classifier that best discriminates the positive instances from the negative ones. Given an unlabeled bag, the classifier is applied over all its instances and if at least one of them is classified as positive, the whole bag is labeled positive.

In the less-commonly studied Multiple Instance Regression (MIR), bag labels are real-valued. In the original contribution to MIR [22], the assumption was that each bag contains a prime instance that determines its label. A solution was proposed to train a linear predictor for prime instances, but it was not specified how to detect the prime instance and how to use the predictor on unlabeled bags. In [27], the prime instance assumption was replaced with the assumption that bag instances have different relevance and that bag label is a relevance-weighted average of instance-level predictions. Since this assumption results in an NP-hard problem, an approximation was proposed that concurrently determines relevant instances and trains a linear predictor. However, the paper did not describe how to determine relevance of instances of an unlabeled bag and how to predict its label. In [28], it was assumed that instances can be grouped into a number of clusters and that only instances from the "prime" cluster are responsible for bag labels. Upon identifying the prime cluster, each bag is represented by a meta-instance as an average of bag instances weighted by the strength of their assignment to the prime cluster. The MIR problem is then treated as a standard regression problem. This approach provides a clear mechanism for labeling an unlabeled bag, but there is an open question as to how appropriate the prime cluster assumption is for any given application. In [29], a constructive procedure was proposed that postulates that bag label should be the mean or median of instance-level predictions. It assumes that each bag contains high-noise instances whose removal from training improves bag prediction accuracy. The proposed algorithm starts from a training set that contains instances from all bags. It incrementally removes the noisy instances from training data as long as bag-level prediction accuracy continues to increase.

Multiple instance regression naturally arises in several types of remote sensing applications. Let us consider the following two applications, which have been previously studied within the MIR setup, as representatives of remote sensing applications appropriate for MIR. The first application is of the task of predicting of Aerosol Optical Depth (AOD) from remotely-sensed data [29]. An important property of aerosol, that can aid AOD prediction, is that it has small spatial variability over

Manuscript received June 4, 2010; revised June 28, 2011; accepted September 11, 2011.

Zhuang Wang is with Siemens Corporate Research, Princeton, NJ, USA (e-mail: zhuang.wang@siemens.com) and was with the Department of Computer and Information Science, Temple University, PA, USA, while most of this work was performed.

Liang Lan and Slobodan Vucetic are with Department of Computer and Information Science, Temple University (e-mail: lanliang@temple.edu, vucetic@temple.edu).

distances of up to 100 km [10]. On the other hand, sensors aboard satellites gather information in the form of multi-spectral images with a high spatial resolution, where pixels could cover areas as small as $200\times200m^2$. In Figure 2, we show the reflectances at a single spectral band of MISR instrument over a 50x50 km^2 area. Over this area, it is reasonable to assume that AOD is nearly constant. The main reason for the observed reflectance variability is the variability of surface properties within the area. Since the remotely-sensed information is a mixture of surface and atmospheric effects, the surface can be considered as a source of observation noise. With respect to the aerosol retrieval, pixels over darker surfaces can be considered to be of higher quality, while pixels over brighter surfaces are of lower quality. If we treat a multi-spectral satellite image over an area as a bag, the AOD value measured by a highly accurate ground-based instrument (e.g. by an AERONET [9] radiometer) as its label, and a single pixel as an instance, then training an AOD predictor from a set of labeled images is a form of MIR. It should be noted that each pixel is a noisy version of the prime instance, which would be a pixel over the perfectly dark surface.

Aerosol retrieval has several features that make it suitable for MIR: multispectral images labeled by ground-based measurements, pixels with varying levels of noise,, label with low spatial variability. This property is shared by the related applications, such as retrieval of greenhouse gases (water vapor [26], carbon monoxide [8], methane, nitrous oxide, ozone [2]). Beyond retrieval of atmospheric properties, the MIR setup can be appropriate for mapping of land surface temperature [4], soil moisture [11], precipitation [12], ocean salinity [15], and ocean [25] and land [31] biological productivity, that are also characterized by spatially high-variable pixels and low-variable label. MIR can also be useful for applications where remotely-sensed observations are proxies for prediction, such as in numerous studies in environmental monitoring [16], ecology [14], and epidemiology [13].

Another remote sensing application that was previously studied as an MIR problem is the prediction of county-level crop yields [27]. For illustration, in Figure 4(a) we show corn yield in each county of Kansas in 2002, while in Figure 4(b) we show the reflectance at a single spectral band of a MODIS instrument observed a few months before the harvest. As can be seen, each county has a single label value (corn yield reported by the United States Department of Agriculture) and contains multiple pixels with varying reflectance. By considering each county as a bag, its crop yield as the bag label, and pixels as instances, this clearly leads to the MIR setting. In the absence of knowledge about the location of crop fields, it becomes reasonable to calculate a bag label as a weighted average of individual instance predictions. It should be noted that, in addition to the pixels over crop fields, some non-crop pixels (e.g. forests) can also contain valuable information (e.g. leaf area index) for the prediction, while others (e.g. cities, lakes, clouds) can be uninformative. Unlike the aerosol retrieval application, where label is a spatially low-variable property, the bag label in the crop prediction application is an aggregated value over a region. Similar setups could occur in other socio-economic applications of remote sensing.

In this paper, we propose a probabilistic framework for analyzing MIR problems and designing MIR algorithms. Our framework is based on the *prime instance assumption* that a bag label is determined by its prime instance. Under this assumption, we treat the bag label as a random variable described with a mixture model, where the contribution of each instance to labeling is proportional to its probability of being the prime instance. To learn the mixture model, we use the expectation-maximization (EM) algorithm. Given the mixture model, the prediction for an unlabeled bag can be obtained as the weighted average of the instance-level predictions. Within the proposed framework, users have flexibility in modeling the probability that an instance is the prime instance. We study several possible strategies appropriate for remote sensing problems. Moreover, previous MIR algorithms Prime-MIR [22], Pruning-MIR [29], and a baseline algorithm called Instance-MIR can be described as special cases of the proposed framework.

The paper is organized as follows. Section II defines the MIR problem; Section III outlines several previously proposed MIR algorithms; Section IV proposes a mixture model framework for the MIR problem; Section V describes the expectation-maximization MIR algorithm and how the previous MIR algorithms fit this framework; Sections VI and VII compare previous and proposed MIR algorithms on synthetic and remote sensing data.

II. MIR Problem Setting

In the MIR problem, we are given a set of B labeled bags, $\mathbf{D} = \{(\mathbf{bag}_i, y_i),\ i = 1...B\}$, where $\mathbf{bag}_i = \{\mathbf{x}_{ij},\ j = 1...b_i\}$, $\mathbf{x}_{ij}$ is an attribute vector of the j-th instance from the i-th bag, y_i is the real-valued label of the i-th bag, and b_i is the number of instances in the i-th bag. The objective is to train a regression model $\hat{y}(\mathbf{bag})$ that accurately predicts the label of an unlabeled bag. Accuracy of MIR is defined as the Mean Squared Error (MSE) of bag label predictions,

$$MSE_{bag} = \frac{1}{B}\sum_{i=1}^{B}(y_i - \hat{y}(bag_i))^2 . \quad (1)$$

We assume that the *prime instance* is responsible for each bag label: label y_i of the i-th bag is a function of the prime instance $\mathbf{p}_i$ plus some added noise,

$$y_i = h(\mathbf{p}_i) + \varepsilon , \quad (2)$$

where h is the unknown regression function and ε is the label noise. If the noise is Gaussian, $\varepsilon \sim N(0,\sigma^2)$, the probability of bag label y_i given $\mathbf{p}_i$ can be written as

$$p(y_i \mid \mathbf{p}_i) = N(y_i \mid h(\mathbf{p}_i),\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{(y_i - h(\mathbf{p}_i))^2}{2\sigma^2}\right). \quad (3)$$

In this paper, we will assume that bag instances are noisy or distorted versions of $\mathbf{p}_i$,

TABLE I
SUMMARY OF NOTATION

Notations	Description
$\mathbf{D}$	data set
$\mathbf{bag}_i$	i-th bag
$\mathbf{x}_{ij}$	j-th instance in $\mathbf{bag}_i$
$\mathbf{X}$	set of $\mathbf{x}_{ij}$
y_i	label of $\mathbf{bag}_i$
$\mathbf{Y}$	set of y_i
b_i	# instances in $\mathbf{bag}_i$
B	# bags in $\mathbf{D}$
h, f	functions
$\hat{y}$	MIR predictor
$\boldsymbol{\theta}$	model parameters
$\mathbf{p}_i$	the prime instance in $\mathbf{bag}_i$
z_{ij}	binary indicator for the prime instance
$\mathbf{z}_i$	$[z_{ij}]$, $j = 1,\ldots,b_j$
$\mathbf{Z}$	set of $\mathbf{z}_i$
π_j	prior probability that $\mathbf{x}_{ij}$ is prime
γ_{ij}	posterior probability that $\mathbf{x}_{ij}$ is prime

$$\mathbf{x}_{ij} = \mathbf{p}_i + \delta_{ij}, \tag{4}$$

where δ_{ij} is a deviation of attributes in the j-th instance of the i-th bag from the prime instance.

To help the reader, our notation is summarized in Table 1.

III. Previously Proposed MIR Algorithms

Let us first introduce the benchmark MIR algorithms that we will use to compare with the proposed algorithms. The first two algorithms have been used previously [28,29] as baselines, while the remaining 3 are some of the most prominent recently proposed MIR algorithms.

A. Aggregated-MIR

In this algorithm, the i-th bag is treated as a meta-instance ($\mathbf{x}_i$, y_i) obtained by averaging all its instances, as

$$\mathbf{x}_i = mean(\{\mathbf{x}_{ij}, j = 1...b_i\}). \tag{5}$$

Then, a regression model f is trained using a set of meta-instances $\mathbf{D}_A = \{(\mathbf{x}_i, y_i), j = 1\ldots B\}$. To predict label of an unlabeled bag, the bag is represented as the meta-instance and used as an input to the predictor,

$$\hat{y}(\mathbf{bag}_i) = f(\mathbf{x}_i). \tag{6}$$

If $E[\delta_{ij}] = 0$ in (4) and b_i is large, Aggregated-MIR is appropriate, because averaging creates meta-instances that approach prime instances as the bag size increases. If $E[\delta_{ij}] \neq 0$, Aggregated-MIR produces sub-optimal results.

B. Instance-MIR

An alternative to treating each bag as a single example is to treat each instance as an example. A straightforward application of the instance-as-an-example approach is to represent the j-th instance from the i-th bag as ($\mathbf{x}_{ij}$, y_i), join instances from all bags into a single training data set $\mathbf{D}_I = \{(\mathbf{x}_{ij}, y_i)$, $i = 1...B$, $j = 1...b\}$, and learn a regression model f from the training data. To prevent giving higher importance to large bags, Instance-MIR samples (with repetition) the same number b of instances from each bag to the training data set. In the absence of a principled way to predict bag label, Instance-MIR uses an intuitively reasonable approach: the label of i-th bag is calculated as

$$\hat{y}(\mathbf{bag}_i) = mean(\{f(\mathbf{x}_{ij}), j = 1...b\}). \tag{7}$$

In case when bags contain outlying instances, it can be more appropriate to use the median predictor,

$$\hat{y}(\mathbf{bag}_i) = median(\{f(\mathbf{x}_{ij}), j = 1...b\}). \tag{8}$$

If the distribution of δ_{ij} in (4) is unknown, Instance-MIR can be treated as learning with noisy attributes. A recent study [21] showed that, despite the simplicity, the instance-based algorithms such as Instance-MIR can provide competitive results on many datasets.

C. Cluster-MIR

Aggregated-MIR and Instance-MIR treat all instances equally. Motivated by the crop prediction problem, where instances might have different relevance for the prediction, the assumption made in [28] is that the instances in each bag are drawn from a number of distinct underlying data distributions and that only one distribution is responsible for bag label. The distinct distributions are identified using soft clustering with k components. Then, for the i-th bag, k meta-instances $\mathbf{x}_{ij}$, $j = 1\ldots k$, are created by weighted averaging of bag instances, according to their assignment to each of the k components. By concatenating all meta-instances for the j-th component, a training data set $\mathbf{D}_j = \{(\mathbf{x}_{ij}, y_i)$, $i = 1...B\}$ is constructed and predictor f_j is trained. Among the k predictors, the one with the highest accuracy is selected to provide predictions on unlabeled bags, by weighted averaging of its instance-level predictions. It should be noted that for $k = 1$, Cluster-MIR reduces to Aggregate-MIR.

Cluster-MIR is flexible, as the user has a choice in selecting clustering algorithm, distance metric, number of clusters, and prediction algorithm. The potential drawback is that clustering is performed in an unsupervised manner, without consulting the bag labels. This can result in uninformative clusters and subpar prediction accuracy, even when the underlying assumption is true.

D. Prime-MIR

Prime-MIR [22] is based on the assumption that each bag contains the prime instance. It is the instance-as-an-example approach where only a single instance from each bag is used in training. Prime-MIR is an iterative algorithm that attempts to discover the prime instances and train a linear predictor on them. Given the currently available predictor, the algorithm first selects from each bag the instance which has the lowest prediction error. These instances are called the prime candidates. Then, a new predictor is trained using the prime candidates. The algorithm iterates as long as the prediction error over prime candidates decreases.

In our previous work [29], we made several slight modifications to the originally proposed algorithm to make it more generally applicable: 1) While the original algorithm starts from a randomly generated predictor, we used Instance-MIR algorithm to build the initial predictor. 2) We allowed use of both linear and nonlinear (i.e. neural networks) regression models. 3) The original algorithm does not propose how to use the resulting predictor on an unseen bag. We used the mean and median averaging from (7) and (8).

E. Pruning-MIR

Instance-MIR and Prime-MIR are two extremes of the instance-as-an-example approach. Instance-MIR uses all available instances and suffers when bags contain many noisy instances. Prime-MIR uses a rather sensitive procedure that does not guarantee detection of the prime instance. Moreover, it uses only a small fraction of instances for training, which could prevent accurate training of more complex models.

To address these issues, Pruning-MIR was proposed in our previous work [29]. It starts from the Instance-MIR solution, and in each iteration discards a small fraction of the noisiest instances in each bag, and then trains a new predictor on the remaining instances. By reducing noise, the algorithm is trying to improve quality of training data and increase accuracy. The algorithm runs as long as there is an improvement in prediction accuracy.

The noisiest instances in a bag are defined as those whose predictions are the farthest away from the median prediction over the non-pruned instances. With such definition of noisy examples, Pruning-MIR criterion ensures that the algorithm is less sensitive to the choice of the initial predictor. To predict of label of an unlabeled bag, either (7) or (8) is used.

F. MIR algorithms not considered in evaluation

There are two more prominent MIR algorithms that we did not consider in the evaluation. The first is the algorithm already mentioned in the Introduction that is attempting to determine relevance of instances in training bags [27]. This algorithm was not considered because it does not have a mechanism to determine relevance of instances in unlabeled bags and provide predictions. The second is the algorithm proposed in [5]. It is similar to the Prime-MIR, with the main difference being that the prime instance is the one with the highest prediction. It should be noted that this prime instance assumption is similar to the assumption used in many of the multiple instance classification algorithms mentioned in the Introduction. However, this assumption is not appropriate for remote sensing applications. Let us take the aerosol retrieval as an example: the instance with the highest prediction is likely to be the pixel over the brightest surface within the region, and thus the noisiest instance in the bag. In addition to this issue, the algorithm from [5] can only train a linear predictor and its generalization to nonlinear regression is not trivial.

IV. Mixture Model For Multiple Instance Regression

The proposed framework is based on the prime instance assumption that one of the b_i instances in $\mathbf{bag}_i$ is responsible for the bag label y_i. Let us define the b_i-dimensional binary random variable $\mathbf{z}_i = [z_{i1} \ldots z_{ib_i}]$, such that $z_{ij} = 1$ if the j-th instance in the i-th bag is prime and $z_{ij} = 0$, otherwise. Therefore, only one element of $\mathbf{z}_i$ is nonzero, and $\sum_{j=1}^{b_i} z_{ij} = 1$. If $z_{ij} = 1$, the conditional probability $p(y_i \mid \mathbf{bag}_i, z_i)$ is fully determined by instance $\mathbf{x}_{ij}$,

$$p(y_i \mid \mathbf{bag}_i, z_i) = p(y_i \mid \mathbf{x}_{ij}). \tag{9}$$

Let us express $p(y_i \mid \mathbf{bag}_i)$ as the marginal of the joint distribution $p(y_i, z_i \mid \mathbf{bag}_i)$. Using the sum and product rules, we can write

$$\begin{aligned} p(y_i \mid \mathbf{bag}_i) &= \sum_{\mathbf{z}_i} p(y_i, \mathbf{z}_i \mid \mathbf{bag}_i) \\ &= \sum_{\mathbf{z}_i} p(\mathbf{z}_i \mid \mathbf{bag}_i)\, p(y_i \mid \mathbf{z}_i, \mathbf{bag}_i). \end{aligned} \tag{10}$$

By observing that there are b_i possible values of $\mathbf{z}_i$, and using (9), we obtain the following mixture model,

$$\begin{aligned} p(y_i \mid \mathbf{bag}_i) &= \sum_{j=1}^{b_i} p(z_{ij} = 1 \mid \mathbf{bag}_i)\, p(y_i \mid x_{ij}) \\ &= \sum_{j=1}^{b_i} \pi_j(\mathbf{bag}_i)\, p(y_i \mid \mathbf{x}_{ij}), \end{aligned} \tag{11}$$

where we defined $\pi_j(\mathbf{bag}_i) = p(z_{ij} = 1 \mid \mathbf{bag}_i)$ for simplicity of notation. Thus, $\pi_j(\mathbf{bag}_i)$ is the prior probability that the j-th instance is the prime instance of the i-th bag, and $p(y_i \mid \mathbf{x}_{ij})$ is the label probability when the j-th instance is the prime instance. Given the mixture model (11), the label of the i-th bag can be predicted as its expected value,

$$\hat{y}(\mathbf{bag}_i) = E[y_i \mid \mathbf{bag}_i] = \sum_{j=1}^{b_i} \pi_j(\mathbf{bag}_i) E[y_i \mid \mathbf{x}_{ij}]. \tag{12}$$

Therefore, the prediction of a bag label is straightforward and follows directly from the defined mixture model. This differs from other MIR approaches that either do not have a mechanism for label prediction or use heuristics.

The learning problem is to determine $\pi_j(\mathbf{bag}_i)$ and $p(y_i \mid \mathbf{x}_{ij})$ from training data. We assume that both probabilities are parametric functions and express them explicitly as $\pi_j(\mathbf{bag}_i, \boldsymbol{\theta}_g)$ and $p(y_i \mid \mathbf{x}_{ij}, \boldsymbol{\theta}_p)$. The mixture model from (11) can now be rewritten as

$$p(y_i \mid \mathbf{bag}_i, \boldsymbol{\theta}) = \sum_{j=1}^{b_i} \pi_j(\mathbf{bag}_i, \boldsymbol{\theta}_g)\, p(y_i \mid \mathbf{x}_{ij}, \boldsymbol{\theta}_p), \tag{13}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_g, \boldsymbol{\theta}_p)$ are the model parameters.

A convenient way to optimize parameters of a mixture model is to use the Expectation-Maximization (EM) algorithm [6]. Let us denote $\mathbf{X} = (\mathbf{bag}_i, i = 1 \ldots B)$, $\mathbf{Y} = (y_i, i = 1 \ldots B)$, and $\mathbf{Z} = (\mathbf{z}_i, i = 1 \ldots B)$, and observe that we can write

$$p(y_i, \mathbf{z}_i \mid \mathbf{bag}_i, \boldsymbol{\theta}) = \prod_{j=1}^{b_i} (\pi_j(\mathbf{bag}_i, \boldsymbol{\theta}_g)\, p(y_i \mid \mathbf{x}_{ij}, \boldsymbol{\theta}_p))^{z_{ij}} \tag{14}$$

Then, the log-likelihood of the complete data, $\mathbf{D}_{complete} = \{(\mathbf{bag}_i, y_i, \mathbf{z}_i), i = 1...B\}$, can be expressed as

$$\ln p(\mathbf{Y},\mathbf{Z}\,|\,\mathbf{X},\boldsymbol{\theta}) = \ln \prod_{i=1}^{B} p(y_i,\mathbf{z}_i\,|\,\mathbf{bag}_i,\boldsymbol{\theta}) = \sum_{i=1}^{B}\sum_{j=1}^{b_i} z_{ij} \ln(\pi_j(\mathbf{bag}_i,\boldsymbol{\theta}_g)\, p(y_i\,|\,\mathbf{x}_{ij},\boldsymbol{\theta}_p)). \quad (15)$$

EM starts with an initial guess of $\boldsymbol{\theta}$ and then updates it by alternating between an expectation (E) step and a maximization (M) step until convergence. In the **E-step**, the algorithm evaluates the expected value of the log-likelihood (15), with respect to the current estimate of the posterior probability of $\mathbf{Z}$, given $\mathbf{X}$, $\mathbf{Y}$ and θ. By denoting the current parameter estimate as $\boldsymbol{\theta}^{old}$, the expectation can be expressed as

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) = E_{\mathbf{Z}|\mathbf{X},\mathbf{Y},\boldsymbol{\theta}^{old}}[\ln p(\mathbf{Y},\mathbf{Z}\,|\,\mathbf{X},\boldsymbol{\theta})] = \sum_{i=1}^{B}\sum_{j=1}^{b_i} p(z_{ij}=1\,|\,\mathbf{bag}_i, y_i, \boldsymbol{\theta}^{old}) \ln(\pi_j(\mathbf{bag}_i,\boldsymbol{\theta}_g)\, p(y_i\,|\,\mathbf{x}_{ij},\boldsymbol{\theta}_p)). \quad (16)$$

The posterior probability that the *j*-th instance in the *i*-th bag is the prime instance can be calculated as

$$p(z_{ij}=1\,|\,\mathbf{bag}_i, y_i, \boldsymbol{\theta}^{old}) = \frac{p(z_{ij}=1, y_i\,|\,\mathbf{bag}_i,\boldsymbol{\theta}^{old})}{p(y_i\,|\,\mathbf{bag}_i,\boldsymbol{\theta}^{old})} = \frac{\pi_j(\mathbf{bag}_i,\boldsymbol{\theta}_g^{old})\, p(y_i\,|\,\mathbf{x}_{ij},\boldsymbol{\theta}_p^{old})}{\sum_{k=1}^{b_i}\pi_k(\mathbf{bag}_i,\boldsymbol{\theta}_g^{old})\, p(y_i\,|\,\mathbf{x}_{ik},\boldsymbol{\theta}_p^{old})} \quad (17)$$

After defining for simplicity of notation $\gamma_{ij}(\boldsymbol{\theta}^{old}) \equiv p(z_{ij}=1\,|\,\mathbf{bag}_i, y_i, \boldsymbol{\theta}^{old})$, we can express $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$ as

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) = \sum_{i=1}^{B}\sum_{j=1}^{b_i}\gamma_{ij}(\boldsymbol{\theta}^{old})\ln\pi_{ij}(\boldsymbol{\theta}_g) + \sum_{i=1}^{B}\sum_{j=1}^{b_i}\gamma_{ij}(\boldsymbol{\theta}^{old})\ln p(y_i\,|\,\mathbf{x}_{ij},\boldsymbol{\theta}_p), \quad (18)$$

In the **M-step**, the algorithm updates the model parameters $\boldsymbol{\theta}$ to maximize *Q*,

$$\boldsymbol{\theta}^{(new)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(old)}). \quad (19)$$

The resulting EM procedure is summarized in Table 2.

To optimize (19) we have to define the parametric functions $\pi_{ij}(\boldsymbol{\theta}_g)$ and $p(y_i\,|\,\mathbf{x}_{ij},\boldsymbol{\theta}_p)$. This will be discussed in the following section.

V. Training Algorithms

In this section we discuss how to solve the MIR optimization problem (19), depending on how $\pi_j(\mathbf{bag}_i,\boldsymbol{\theta}_g)$ and $p(y_i\,|\,\mathbf{x}_{ij},\boldsymbol{\theta}_p)$ are defined.

A. Label probability

Consistent with the assumption (3), we define $p(y_i\,|\,\mathbf{x}_{ij},\boldsymbol{\theta}_p) = N(y\,|\,f(\mathbf{x}_{ij},\mathbf{w}),\delta^2)$, where $\boldsymbol{\theta}_p = (\mathbf{w},\delta)$, and

TABLE II
The EM Algorithm for Mixture Model MIR

Input: D
Output: *f*
Initialize: $\boldsymbol{\theta}^{old}$
Repeat
  E-step: construct $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$
  M-step: $\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$
  $\boldsymbol{\theta}^{old} = \boldsymbol{\theta}^{new}$
Until convergence

$f(\mathbf{x},\mathbf{w})$ is a regression function with parameters $\mathbf{w}$. Then, the $\ln p(y_i\,|\,\mathbf{x}_{ij},\boldsymbol{\theta}_p)$ term from (18) becomes

$$\ln p(y_i\,|\,\mathbf{x}_{ij},\boldsymbol{\theta}_p) = \frac{(y_i - f(\mathbf{x}_{ij},\mathbf{w}))^2}{\delta^2} + \ln 2\pi\delta^2. \quad (20)$$

B. Prime prior as a deterministic function

In some remote sensing applications, there might be a clear physical interpretation of the prime instance and it can be reasonable to set the values of π_j based purely on domain knowledge. In this case, $\pi_j(\mathbf{bag}_i,\boldsymbol{\theta}_g) = \pi_j(\mathbf{bag}_i)$, and there is no need for the algorithm to learn the parameters $\boldsymbol{\theta}_g$. The only condition in designing the priors is that $\sum_j \pi_{ij} = 1$ and $\pi_{ij} \geq 0$, to ensure that priors can be treated as probabilities. For example, pixels from heavily urban areas are not expected to be correlated with crop yield. Similarly, cloudy pixels and pixels over bright surfaces are considered very noisy from the perspective of aerosol prediction. In both cases we can set their priors to zero and the priors of the remaining pixels to a constant value.

As a consequence, the first term in (18) can be treated as a constant during the EM procedure and (19) can be simplified to

$$\boldsymbol{\theta}^{new} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{B}\sum_{j=1}^{b_i}\gamma_{ij}(\boldsymbol{\theta}^{old})\left(\frac{(y_i - f(\mathbf{x}_{ij},\mathbf{w}))^2}{\delta^2} + \ln 2\pi\delta^2\right), \quad (21)$$

where $\boldsymbol{\theta} = \boldsymbol{\theta}_p = (\mathbf{w},\delta)$.

To optimize $\mathbf{w}$ in (21), we can treat δ as constant. The resulting problem is equivalent to minimizing the weighted squared error of *f* at the instance level,

$$E_f = \sum_{i=1}^{B}\sum_{j=1}^{b_i}\gamma_{ij}(\boldsymbol{\theta}^{(old)})(y_i - f(\mathbf{x}_{ij},\mathbf{w}))^2. \quad (22)$$

Let us consider the setup where *f* is a feedforward neural network with weights $\mathbf{w}$. We use neural networks because they are powerful and easy to train. The basic method for neural network training is gradient descent, where weights $\mathbf{w}$ are updated in the negative direction of the gradient of the cost function as $\mathbf{w} \leftarrow \mathbf{w} - \eta \cdot \partial E_f / \partial \mathbf{w}$, where η is the learning rate. Table 3 gives the details of calculation of the gradient $\partial E_f / \partial \mathbf{w}$. We note that the gradient contains partial derivatives $\partial f(\mathbf{x}_{ij},\mathbf{w}) / \partial \mathbf{w}$, which can be calculated efficiently using the backpropagation procedure.

Given the neural network, the remaining step is to optimize

TABLE III
THE NEURAL NETWORK TRAINING PROCEDURE FOR OPTIMIZING f

The cost function:

$$E_f = \sum_{i=1}^{B}\sum_{j=1}^{b_i}\gamma_{ij}(\boldsymbol{\theta}^{old})(y_i - f(\mathbf{x}_{ij},\mathbf{w}))^2 ;$$

Weight update rule:

$\mathbf{w} \leftarrow \mathbf{w} - \eta \cdot \partial E_f / \partial \mathbf{w}$, where

$$\frac{\partial E_f}{\partial \mathbf{w}} = -2\sum_{i=1}^{B}\sum_{j=1}^{b_i}\gamma_{ij}(\boldsymbol{\theta}^{old})(y_i - f(\mathbf{x}_{ij},\mathbf{w}))\frac{\partial f(\mathbf{x}_{ij},\mathbf{w})}{\partial \mathbf{w}}.$$

TABLE IV
THE EM ALGORITHM FOR MIXTURE MODEL MIR

The cost function:

$$E_g = -\sum_{i=1}^{B}\sum_{j=1}^{b_i}(\gamma_{ij}(\boldsymbol{\theta}^{old})\ln \pi_{ij}(\boldsymbol{\theta}_g)) ;$$

Weight update rule:

$\boldsymbol{\theta}_g \leftarrow \boldsymbol{\theta}_g - \eta \cdot \partial E_g / \partial \boldsymbol{\theta}_g$, where

$$\frac{\partial E_g}{\partial \boldsymbol{\theta}_g} = -\sum_{i=1}^{B}\sum_{j=1}^{b_i}(\gamma_{ij}(\boldsymbol{\theta}^{old}) - \pi_{ij}(\boldsymbol{\theta}_g))\frac{\partial g(\mathbf{x}_{ij},\boldsymbol{\theta}_g)}{\partial \boldsymbol{\theta}_g}$$

δ. By setting the derivative of (21) with respect to δ to zero, the optimal δ is obtained in a closed-form as

$$\delta^2 = \frac{\sum_{i=1}^{B}\sum_{j=1}^{b_i}\gamma_{ij}(\boldsymbol{\theta}^{(old)})(y_i - f(\mathbf{x}_{ij},\mathbf{w}))^2}{\sum_{i=1}^{B}\sum_{j=1}^{b_i}\gamma_{ij}(\boldsymbol{\theta}^{(old)})}. \quad (23)$$

We call the resulting EM algorithm EM$^{\mathrm{D}}$-MIR, where D stands for Deterministic prior.

C. Prime prior as a function of prediction deviation

The deterministic prior introduced in Section V.B allows for consulting the existing prediction models. Given this possibility, as a special case of the deterministic prior, we define

$$\pi_j(\mathbf{bag}_i) = \frac{1}{C_i} N(f(\mathbf{x}_{ij},\mathbf{w}^{old}) \mid \mu_i, \nu_i^2), \quad (24)$$

where C_i is the normalization constant that ensures that all priors in the bag sum to 1 (i.e. $\sum_j \pi_{ij} = 1$). Parameter μ_i is defined as the median of all predictions in $\mathbf{bag}_i$, $\mu_i = median(\{f(\mathbf{x}_{ij}, \mathbf{w}^{(old)})\})$ and ν_i as the multiple of Median Absolute Deviation (MAD)

$$\nu_i = 1.48 \cdot MAD = 1.48 \cdot median(\{|\, f(\mathbf{x}_{ij},\mathbf{w}^{old}) - \mu_i \,|\}). \quad (25)$$

MAD is a robust measure of the variability of a univariate sample. The introduced prior (24) is related to the idea used in Pruning-MIR from Section III.D. It gives high probability to instances whose prediction is close to the median prediction.

We call the resulting EM algorithm EM$^{\mathrm{PD}}$-MIR, where PD stands for Predictive Deviation.

D. Prime prior as a parametric function

In this strategy, the prime prior is defined as a parametric function $\pi_j(\mathbf{bag}_i, \boldsymbol{\theta}_g)$. To ensure that $\sum_{j=1}^{b_i}\pi_j = 1, \pi_j \geq 0$, we define $\pi_{ij}(\mathbf{bag}_i, \boldsymbol{\theta}_g)$ as the soft-max function,

$$\pi_j(\mathbf{bag}_i, \boldsymbol{\theta}_g) = \frac{\exp(g(\mathbf{x}_{ij},\mathbf{w}_g))}{\sum_{k=1}^{b_i}\exp(g(\mathbf{x}_{ik},\mathbf{w}_g))}, \quad (26)$$

where g is a parametric function. Specifically, we will assume that g is a feedforward neural network with weights $\mathbf{w}_g$. Thus, $\boldsymbol{\theta}_g = \mathbf{w}_g$, and the parameters of the mixture model from (11) are $\boldsymbol{\theta} = (\mathbf{w}, \delta, \mathbf{w}_g)$.

We should observe that weights $\mathbf{w}_g$ influence (19) only through its first term, while weights $\mathbf{w}$ and scalar δ influence (19) only through its second term. Therefore, the M-step reduces to separate optimization problems. The optimization of $\mathbf{w}$ and δ was explained in Section V.B. In the following, we discuss learning of $\mathbf{w}_g$. After plugging (26) into the first term of (19), the optimization problem reduces to the minimization of the cross entropy between the prime prior and posterior probabilities. This implies that g is optimized such that the prime prior approximates the prime posterior. The procedure for training of neural network g is summarized in Table 4.

We call the resulting EM algorithm EM$^{\mathrm{G}}$-MIR, where superscript G is a reminder that the prior is parameterized.

E. MIR predictor

Following (11), and given the learned parameters $\boldsymbol{\theta} = (\mathbf{w}, \delta, w_g)$, the label of the i-th bag is predicted as

$$\hat{y}(\mathbf{bag}_i) = \sum_{j=1}^{b_i}\pi_j(\mathbf{bag}_i, \boldsymbol{\theta}_g) f(\mathbf{x}_{ij}, \mathbf{w}), \quad (27)$$

where π_j can be a predetermined deterministic function, or calculated using (24) or (26).

F. Reductions to other Approaches

Three previously proposed algorithms, Prime-MIR [22], Pruning-MIR [29], and Instance-MIR, can be interpreted as the special cases of the proposed mixture model MIR framework.

Special case 1 (Prime-MIR): Let us assume that $\pi_j = 1/b_i$ for all instances in $\mathbf{bag}_i$ and that the δ parameter in (20) is predetermined and very small. By denoting *prime*(i) as the index of the instance predicted with the smallest error, it follows that $p(y_i \mid \mathbf{x}_{i,prime(i)}, \boldsymbol{\theta}_p) >> p(y_i \mid \mathbf{x}_{ik}, \boldsymbol{\theta}_p)$, for all $k \neq prime(i)$. From there it follows that $\gamma_{i,prime(i)} \approx 1$ and $\gamma_{ik} \approx 0$ for all $k \neq prime(i)$. In this case, the cost function (22) can be approximated with

$$E_f = \sum_{i=1}^{B}(y_i - f(\mathbf{x}_{i,prime(i)}, \mathbf{w}))^2. \quad (28)$$

Note that the instance with index *prime*(i) is exactly the prime candidate introduced in the Prime-MIR algorithm. Learning the weight $\mathbf{w}$ consists of repeatedly minimizing (28) and recalculating the prime candidates, which is equivalent to the Prime-MIR algorithm. Prediction of the bag label using (27) is simply the average of all predictions in a bag and it is consistent with the Prime-MIR prediction heuristic (7) explained in

Section III.C.

Special case 2 (Pruning-MIR): Let us assume that the δ parameter in (20) is predetermined and very large, and define $\pi_j = 0$ for the noisiest bag instances and $\pi_j = const$ for the remaining ones. Following Pruning-MIR described in Section III.D, the noisiest instances are those whose prediction is the farthest from the median prediction in the bag. It is worth pointing out here that the prior defined in (24) and (26) is consistent with this description and can be treated as a particular way of defining the noisiest instances.

Since δ is very large, the $p(y_i \mid \mathbf{x}_{ij}, \boldsymbol{\theta}_p)$ values of all bag instances are very similar and $\gamma_{ij} \approx \pi_j$. In this case, the cost function (22) can be approximated with

$$E_f = \sum_{i=1}^{B} \sum_{j,\pi_j \neq 0} (y_i - f(\mathbf{x}_{ij}, \mathbf{w}))^2 , \quad (29)$$

and the resulting procedure consists of minimizing (29) and removing the noisiest instances from each bag. This is exactly the procedure followed by Pruning-MIR. Following (27), the bag label is the average prediction over the low-noise instances, which is a slight departure from the original Pruning-MIR mean/median predictor.

Special case 3 (Instance-MIR): Let us assume that the δ parameter in (20) is predetermined and very large and define $\pi_j = 1/b_i$. Since δ is very large, the $p(y_i \mid \mathbf{x}_{ij}, \boldsymbol{\theta}_p)$ values of all bag instances are very similar and $\gamma_{ij} \approx \pi_j$. In this case, the cost function (22) can be approximated with

$$E_f = \sum_{i=1}^{B} \sum_{j=1}^{b_i} (y_i - f(\mathbf{x}_{ij}, \mathbf{w}))^2 , \quad (30)$$

and the resulting procedure consists of minimizing (30). It is interesting to observe that, in this case, the EM procedure terminates after a single iteration. This resulting algorithm is exactly the baseline Instance-MIR algorithm. Following (27), the bag label is predicted as the average over the bag instances.

Discussion of special cases. The interpretation of Prime-MIR and Pruning-MIR within the proposed framework reveals that they require very restrictive assumptions about the δ parameter – it is fixed as either a very large or a very small number. These models also use deterministic prime priors, which further reduces their representation power. Therefore, the proposed EM framework for MIR is a significant extension of these two previously proposed approaches. As for the Instance-MIR, although it can be described as a special case of the EM framework, its appeal lies in its simplicity and its usefulness for benchmarking. Finally, it should be observed that Aggregate-MIR and Cluster-MIR cannot be described within the proposed framework.

VI. Experiments on Synthetic Data

In this section, we present evaluation of the proposed algorithms on several synthetic datasets. We also evaluated baseline methods, Instance-MIR and Aggregated-MIR, and the previously proposed methods, Cluster-MIR, Prime-MIR [22] and Pruning-MIR [29], which are all described in Section III. We used the Root Mean Squared Error (RMSE) of bag label predictions defined in (1) to assess the performance.

A. Synthetic Data Sets

We constructed synthetic data sets following the data generating process described in (2) and (4). To facilitate the interpretation of the results, we used one-dimensional attribute vectors and regression function $h(x)$ for generating bag labels. We studied linear regression, $h(x) = x$, and nonlinear regression, $h(x) = x^2$. The synthetic data sets differed in the way we generated label noise ε_i in (2) and attribute noise δ_{ij} in (4).

MIR-Gaussian(B, b, σ, s). For each bag, $\mathbf{bag}_i$, $i = 1 \ldots B$, we generated the prime instance as a random number between 0 and 1. The bag label y_i was generated by adding the Gaussian noise ε_i with mean 0 and variance σ^2. Then, we generated b (=5, 100) instances for each bag as noisy versions of the prime instance using (4), where δ_{ij} was Gaussian noise with mean zero and variance s^2. MIR-Gaussian generator is idealistic and all MIR algorithms described before should achieve good accuracy.

MIR-Outlier1(B, b, σ, s). Real-life remote sensing data are likely to introduce more complex attribute and target noise than the Gaussian noise used in MIR-Gaussian generator. For example, in aerosol prediction problem, bags over highly variable terrain will contain a large fraction of outlying instances, while bags over bright terrain will have instances with biased noise distribution.

To simulate these properties, MIR-Outlier1 generates bags with different fractions of outliers. Specifically, q_i% of instances in the i-th bag are generated using MIR-Gaussian generator, where q_i is a random number between 50 and 100, and the remaining instances are generated as outliers. The attribute in the j-th outlier instance of the i-th bag is generated as $x_{ij} = p_i + \delta_{ij} + \nu_i$, where δ_{ij} is the Gaussian noise with variance $25s^2$ and ν is an offset generated as a random number between –0.25 and 0.25.

MIR-Outlier2(B, b, σ, s). In addition to outlying instances generated by MIR-Outlier1, real-life data are characterized by outlying target values. Starting from MIR-Outlier1 generator, MIR-Outlier2 generates outlying targets in 20% of the randomly selected bags as $y_j = h(p_j) + \varepsilon_j$, where ε_j is Gaussian additive noise with mean zero and variance $25\sigma^2$.

B. Experimental Design

Feedforward neural networks with one hidden layer and 5 hidden nodes were used as the regression model f. For the prime prior function g in EM^G-MIR, we used feedforward neural networks with one input and one hidden node. The resilient backpropagation algorithm [24] was iterated 200 epochs for training of f and 50 epochs for g. For Pruning-MIR, 5 iterations of instance removal were used, as in [29]. For Instance-MIR, Prime-MIR, and Pruning-MIR, the median predictor (8) was used in testing. The proposed EM-MIR algorithms were iterated until convergence (the increase of the objective (18) is

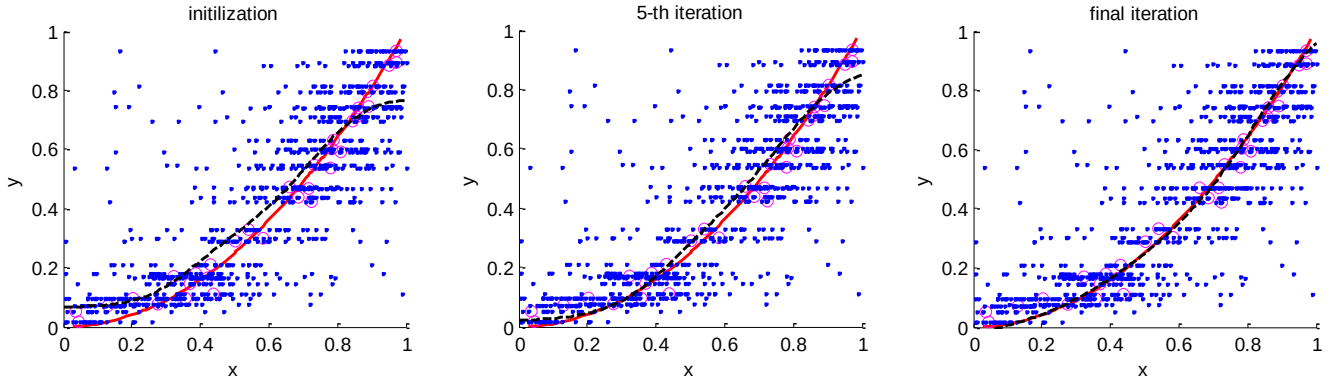


Fig. 1. Improvement of regression curves of EM^{G2}-MIR on Outlier1 data for $h(x)=x^2$. (Red solid curve is the true function, black slash curve is the regression function, blue dots are the bag instances and magenta circles are the prime instances. Each horizontal line of dots corresponds to a bag.)

less than 0.1). For Cluster-MIR[1], we experimented with different numbers of clusters, k = {5, 15, 30} but only reported the result with the lowest RMSE. We also evaluated the case when both attribute **x** and the prediction deviation are used as inputs to g and we denote this algorithm EM^{G2}-MIR.

C. Results

The summary of experimental results on the synthetic data for h (**x**) = x and b=100 is shown in Table 5.a, for h (**x**) = x^2 and b=100 in Table 5.b, and for h (**x**) = x^2 and b=5 in Table 5.c. RMSE are calculated on the test data. Each result is an average of 10 runs of the algorithm. Values in bold indicate the best results for each data set.

For MIR-Gaussian(100, b, 0.05, 0.1), all employed MIR methods achieved similar accuracies, as expected by the design of the data set. However, the proposed EM-MIR algorithms were slightly more accurate than the alternatives on the nonlinear regression data. The success of Aggregated-MIR on linear regression data can be attributed to the fact that instance attribute was created by adding Gaussian noise with mean zero to the prime instance (as discussed in III.A).

For MIR-Outlier1(100, b, 0.05, 0.1), the performance of Instance-MIR is markedly inferior to other algorithms due to a large level of attribute noise and the known effect of regression function attenuation when learning on data with noisy attributes. EM^{PD}-MIR and EM^{G2}-MIR were the most successful approaches reflecting their robustness to the outlying instances. EM^{G2}-MIR was more accurate than EM^{G}-MIR, which shows the importance of prediction deviation for discriminating noisy instances in this data set. Similar results were also observed on MIR-Outlier2 (100, b, 0.05, 0.1) data set, but it should be noted that the advantage of EM-MIR algorithms over competitors was even more pronounced. For small bag size of b=5, accuracy of all MIR algorithms deteriorated, indicating their sensitivity to bag size. The relative performances of MIR algorithms mostly remained similar. However, it should be noted that performances of Aggregated, Cluster, and EM^{G} suffered the most, while Prime was the most resilient to the reduction in bag size.

In Figure 1, the details of the evolution of the EM^{G2}-MIR algorithm are illustrated. We can see that the regression curve was refined during the EM iterations and eventually closely matched the regression function.

TABLE V.A
RMSE(× 100) ON 3 SYNTHETIC DATA SETS, BAG SIZE = 100, H(X)=X.

Algorithms	Datasets		
	MIR-Gaussian	MIR-Outlier1	MIR-Outlier2
Aggregated	**5.0±0.0**	7.0±0.0	7.3±0.0
Instance	6.5±0.0	14.4±0.0	14.6±0.0
Prime	6.3±0.5	14.1±1.9	11.4±0.2
Pruning	5.1±0.0	**5.3±0.0**	6.4±0.0
Cluster	7.1±0.2	6.5±0.1	7.9±0.5
EM^{PD}	5.4±0.0	5.4±0.0	**6.3±0.0**
EM^{G}	5.9±0.0	7.2±0.1	7.8±0.5
EM^{G2}	5.7±0.3	5.8±0.5	6.9±0.5

TABLE V.B
RMSE(× 100) ON 3 SYNTHETIC DATA SETS, BAG SIZE = 100, H(X)=X^2.

Algorithms	Datasets		
	MIR-Gaussian	MIR-Outlier1	MIR-Outlier2
Aggregated	5.7±0.1	7.5±0.1	9.8±0.3
Instance	6.1±0.1	10.8±0.3	10.8±0.1
Prime	7.2±1.7	8.1±1.8	9.9±1.7
Pruning	6.0±0.1	7.1±0.4	8.9±0.2
Cluster	5.7±0.3	7.0±0.1	9.2±0.3
EM^{PD}	**5.6±0.0**	**5.9±0.1**	**7.1±0.1**
EM^{G}	5.8±0.0	8.6±0.2	9.4±0.2
EM^{G2}	**5.6±0.1**	6.8±0.9	7.5±0.6

TABLE V.C
RMSE(× 100) ON 3 SYNTHETIC DATA SETS, BAG SIZE = 5, H(X)=X^2.

Algorithms	Datasets		
	MIR-Gaussian	MIR-Outlier1	MIR-Outlier2
Aggregated	7.1±0.2	13.3±0.2	14.0±0.6
Instance	6.9±0.1	11.4±0.2	13.0±0.3
Prime	6.9±0.3	9.7±0.1	10.4±0.5
Pruning	6.7±0.1	9.5±0.1	10.6±0.3
Cluster	7.3±0.2	12.4±0.1	15.1±1.2
EM^{PD}	**6.6±0.1**	**9.3±0.2**	10.2±0.2
EM^{G}	6.7±0.0	13.1±1.7	12.1±1.0
EM^{G2}	**6.6±0.1**	10.4±1.4	**9.9±0.4**

[1] We used feedforward neural networks (with the same architecture as in the other algorithms) as the prediction algorithm to replace with SVR (the one used in [36]]) because we observed neural networks achieved lower RMSE than SVR on our datasets.

VII. Experiments on Remote Sensing Data

We start this section by describing 5 remote sensing data sets and then summarize how different MIR algorithms performed on them.

A. Remote Sensing Data Sets

Two applications studied in this paper are Aerosol Optical Depth (AOD) retrieval and crop yield prediction from remotely sensed data. The summary of the 5 data sets is in Table 6.

TABLE VI
SUMMARY OF REMOTE SENSING DATA SETS

Data sets	#bags	#instances in bags	Dimensions
AOD-MISR1	800	100	16
AOD-MISR2	800	100	16
AOD-MODIS	1364	100	12
CROP-WHEAT	388	100	40
CROP-CORN	368	100	40

Aerosol Retrieval (Figure 2)**.** Aerosols are small airborne particles that both reflect and absorb incoming solar radiation and whose effect on the Earth's radiation budget is one of the biggest challenges of current climate research. AOD retrieval from satellite measurements is an important remote sensing task. We used collocated ground-based and satellite data from two instruments, MISR and MODIS.

AOD-MISR1 data set is a collection of 800 bags collected at 35 AERONET [9] ground sites (Figure 3) within the continental U.S. between 2001 and 2004 from the MISR satellite. Each bag consisted of 100 instances, representing randomly selected pixels within 20-kilometer radius around the AERONET site. The instance attributes were 12 reflectances from the three middle MISR cameras as well as 4 solar and view zenith angles. The bag target value was the AOD measured by the AERONET instrument within 30 minutes of the satellite overpass.

AOD-MISR2 data set has the same properties as AOD-MISR1. The only difference is that the 100 bag instances were sampled only from the non-cloudy pixels. The two data sets were used to better characterize different MIR algorithms, because cloudy pixels are known to be noisy and lead to reduced retrieval quality.

AOD-MODIS data set was constructed using the MODIS satellite instrument. AOD-MODIS consists of 1,364 bags collected at 45 AERONET sites within the continental U.S. between 2002 and 2004. Each bag consists of 100 instances. The instance attributes were 7 MODIS reflectances and 5 solar and view zenith angles and the bag label was the corresponding AERONET AOD measurement.

To support the MIR research, AOD-MISR1, AOD-MISR2, AOD-MODIS, and the synthetic data, can be found at `http://www.dabi.temple.edu/~vucetic/MIR.html`.

Crop Yield Prediction (Figure 4). The goal of crop yield prediction is to estimate accurately crop yield using the remote sensing observations over a specific region. The WHEAT and CORN data sets [27] used in this study consist of more than 350 labeled bags collected between 2001 and 2004. Each bag represents one of 100 counties within Kansas, USA. Bag labels are average wheat and corn yields within the county based on the USDA records. Each bag consists of instances representing 100 randomly selected MISR pixels within each county. Each instance is a 92-dimensional vector comprising of 46 daily observations during the growing season in two spectral bands. The observations during the first 15 and last 11 days were noisy and the corresponding attributes were removed from the data set. Instances in the resulting data set contained 40 attributes.

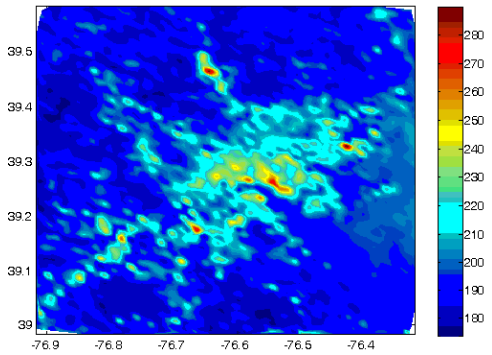


Fig. 2. AOD prediction: 50 x 50 km^2 MISR reflectance

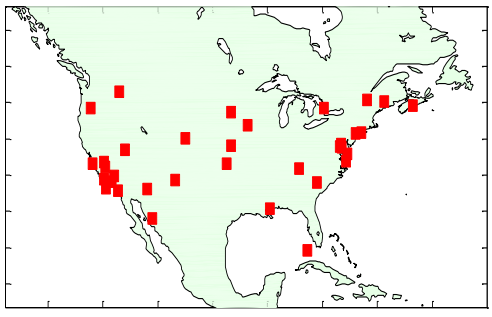
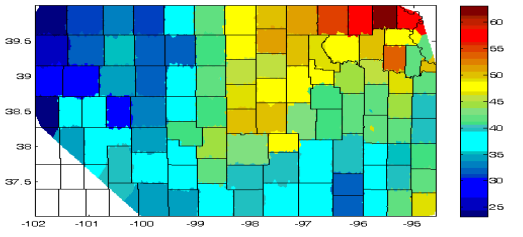
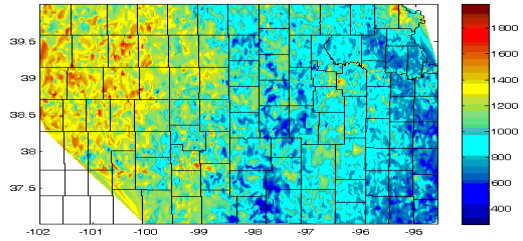


Fig. 3. Locations of 35 ground-based stations for aerosol measurement



(a) Kansas corn yield over counties



(b) Kansas reflectance over counties

Fig. 4. Crop yield prediction

B. Experimental Design

We evaluated various MIR algorithms on the 5 described data sets. As the regression model f in all algorithms, we used feedforward neural networks with 1 hidden layer and 10 hidden nodes. As function g in EM^{G}-MIR and EM^{G2}-MIR, we used feedforward neural networks with one hidden node. The resilient backpropagation algorithm was iterated 200 epochs for training of f and 50 epochs for g. For Pruning-MIR, we used five iterations of instance removal. For Instance-MIR, Prime-MIR and Pruning-MIR, the median predictor (8) was used in the testing stage. The EM-MIR algorithms were iterated until convergence (the increase of the objective (18) is less than 0.1). For Cluster-MIR, we still experimented with different k ({5, 15, 30}) and only reported the best result.

Accuracy of each MIR algorithm was evaluated using 5-cross-validation (5-CV) where the bags were randomly split into 5 subsets; one subset was reserved for testing and the others for training of an MIR predictor; the procedure was repeated 5 times, each with different subset reserved for testing. The 5-CV was repeated 10 times and the average RMSE is reported in Table 7. Values in bold indicate the best results for each data set.

We used two more algorithms in addition to the already described ones. A baseline MEAN predictor was used for benchmarking of other algorithms. In MEAN, the average target value from training bags was used to predict label of a test bag. In aerosol retrieval, it has been known that the darkest and brightest pixels are highly likely to be noisy. This knowledge has been used in the operational MOIDS AOD prediction algorithms that remove the darkest and brightest pixels prior to prediction. Following this, we used the deterministic prior that assigns $\pi_{ij} = 0$ to 20% of the darkest and 50% of the brightest instances[2] and $\pi_{ij} = const$ to the remaining instances. We call the resulting EM algorithm EM^{D}-MIR in Table 7.

C. Results

From Table 7, it could be seen that all MIR algorithms achieved significant improvement over MEAN on all 5 data sets. EM^{PD}-MIR and EM^{G2}-MIR were the most accurate on the two crop data sets. EM^{D}-MIR, EM^{G}-MIR and Pruning-MIR had lower, but still competitive, accuracies. Interestingly, Aggregated-MIR was highly competitive, which indicates that attribute averaging was a fortunate choice for the crop data sets. On the other hand, accuracy of Instance-MIR was quite low, which indicates that instance attributes were quite noisy. Prime-MIR was less accurate showing that the aggressive search for prime candidates was not appropriate on the crop data. Cluster-MIR was the least accurate MIR approach which shows the clustering assumption was not held on the remote sensing data.

EM^{G2}-MIR was the most accurate algorithm on AOD data sets, and it was followed by the other EM^{PD}-MIR algorithms. Unlike its performance on crop data, Aggregate-MIR was the least accurate MIR method on AOD data. Instance-MIR fared only slightly better. This indicates that AOD instances are both noisy and biased: 1) land cover is a source of attribute noise and it tends to be quite variable, and 2) the near-prime instances correspond to dark surface pixels (causing the mean of the attribute noise in (4) to be positive). Accuracy of Prime-MIR was overall better than on the crop data, but it still significantly lagged the best MIR algorithms. Pruning-MIR was quite competitive.

Comparing results on AOD-MISR1 and AOD-MISR2 data sets, the accuracy of all algorithms was higher on the non-cloudy AOD-MISR2. Importantly, it could be seen that EM^{G2}-MIR and EM^{PD}-MIR, in addition to being the most accurate on both data sets, appeared very robust to the existence of cloudy pixels, unlike other MIR algorithms. We point to the impressive result on AOD-MISR1 data, where RMSE of EM^{G2}-MIR was more than 25% lower than RMSE of baseline

TABLE VII
RESULTS ON 5 REMOTE SENSING DATA SETS. FOR AOD DATA, ACCURACY IS RMSE × 100; FOR CROP DATA, ACCURACY IS RMSE

Algorithms	Data sets				
	CROP-WHEAT	CROP-CORN	AOD-MISR1	AOD-MISR2	AOD-MODIS
MEAN	11.3±0.5	39.0±0.4	18.6±0.8	18.6±0.8	19.3±0.7
Aggregated	5.4±0.1	28.4±0.3	10.4±0.1	8.1±0.1	11.8±0.1
Instance	6.3±0.2	30.0±0.3	10.0±0.1	8.2±0.1	11.2±0.1
Prime	6.6±0.4	31.4±0.5	9.5±0.3	8.2±0.1	11.1±0.3
Pruning	5.7±0.2	27.7±0.2	8.3±0.1	8.1±0.3	10.6±0.1
Cluster	7.5±0.3	33.3±2.1	10.5±0.6	9.8±1.0	11.8±0.6
EM^{D}	5.4±0.1	28.7±0.6	8.3±0.6	7.4±0.4	10.8±0.4
EM^{PD}	5.1±0.1	**26.8±0.2**	7.8±0.1	7.7±0.1	9.8±0.1
EM^{G}	5.4±0.1	27.4±0.7	8.5±0.1	7.6±0.3	10.0±0.2
EM^{G2}	**4.9±0.1**	27.5±0.2	**7.5±0.1**	**7.3±0.1**	**9.5±0.1**

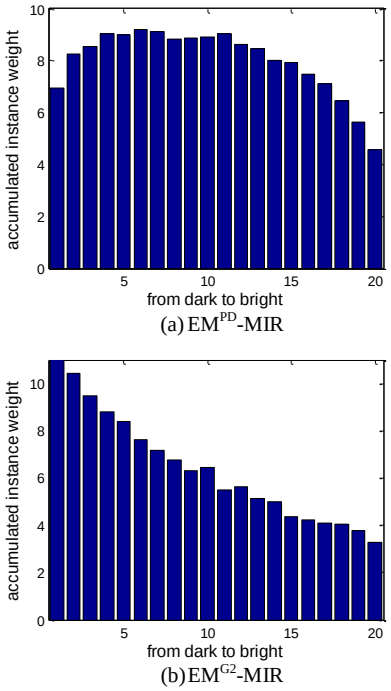


Fig. 5. Accumulated prime prior probability, sorted by darkness in prediction stage on AOD-MISR2 data

[2] For two AOD-MISR data sets, brightness is defined based on the reflectance from the nadir camera at wavelength 466nm. For AOD-MODIS data set, we used the reflectance at wavelength 446nm. For two crop data sets, ideally prior should be set according to land cover information. However, this information is unavailable for this data set. Therefore, we still used the same prior as for the aerosol data sets. Specifically, we used the median value of the time series at red wavelength to calculate brightness.

Instance-MIR. Comparison between AOD-MISR1 and AOD-MISR2 results indicate that data cleaning based on the domain knowledge can be very useful and could blur the difference between various MIR algorithms.

While results from Table 7 show that domain knowledge can be helpful in improving accuracy of MIR algorithms, we explored to what extent it can be revealed by the proposed EM framework. In Figure 5.(a) we show the histogram of π_{ij} values of bag instances calculated by EM^{PD}-MIR. The x-axis is sorted from the darkest to brightest instance. We can observe that the weight given to the brightest and darkest instances is low. This outcome is consistent with the pixel masking procedure commonly applied in aerosol retrieval. For example, in MODIS operational algorithm, 10% of the darkest and 60% of the brightest pixels are masked prior to retrieval, as it has been observed that this results in higher retrieval accuracy [23]. The justification for removal of the darkest pixels is that they might represent contamination by cloud or topographic shadows. Figure 5.(b) shows the results by using EM^{G2}-MIR. We can see the parametric function g in EM^{G2}-MIR learned that darker instances are more important than brighter instances. Considering that EM^{G2}-MIR was the most accurate algorithm, it could be concluded that removal of the darkest pixels might not be necessary on the AOD-MISR2 data. This result adds to the ongoing discussion about the need to mask the darkest pixels before aerosol retrieval [20].

VIII. Conclusions

The Multiple Instance Regression (MIR) setting is applicable when instances can be assigned to multiple bags, each with a single real-valued label. Interestingly, MIR is very relevant to a number of remote sensing problems. As illustrated in this paper, this includes retrieval of atmospheric parameters and spatially aggregated quantities from high-resolution satellite images.

The basic assumption in the proposed Expectation -Maximization (EM) algorithm for MIR is that the bag label is a noisy function of the ideal, or prime, instance. Given this assumption, we used a mixture model that determines bag label probability as a weighted sum of the label probabilities from individual instances. The learning objective was to determine the prior function and the prediction functions and it was solved by the EM algorithm.

The proposed MIR framework is very flexible – it allows encode the domain knowledge about the problem through the prior function, but it could also be applied well when relatively little is known about the problem. It is evident that the success of the proposed MIR approach in any particular application will depend on the modeling skills and domain knowledge about the studied problem. The proposed framework also subsumes several previously proposed MIR algorithms as the special cases.

The experimental results show that EM-MIR is superior to the previously proposed MIR algorithms, as well as to the baseline algorithms that treat MIR as the standard supervised learning problem. This indicates that EM-MIR could be useful when solving many remote sensing and related problems in other disciplines. We also observed that the proposed and existing MIR algorithms are sensitive to instance and label noise, and that their accuracy deteriorates with the complexity of the regression problem and it increases as the size of the bags increases. These properties should be kept in mind when applying MIR algorithms in practice.

Acknowledgment

Authors thank Alex Yates (Temple Univ.) for proofreading the paper. This work was supported in part by the U.S. National Science Foundation under Grants IIS-0546155 and IIS-0612149.

References

[1] S. Andrews, I. Tsochantaridis and T. Hofmann, “Support vector machines for multiple-instance learning,” in Advances in Neural Information Processing Systems, vol. 15, pp. 561–568, 2003.

[2] D. Balis, M. Kroon, M. E. Koukouli, E. J. Brinksma, G. Labow, J. P. Veefkind and R. D. McPeters, “Validation of ozone monitoring instrument total ozone column measurements using Brewer and Dobson spectrophotometer ground-based observations,” in Journal of Geophysical Research, vol. 112, D24S46, 2007.

[3] Y. Chen and J. Z. Wang, “Image categorization by learning and reasoning with regions,” in J. Machine Learning Research, vol. 5, pp. 913–939, 2004.

[4] C. Coll, V. Caselles, J. M. Galve, E. Valor, R. Niclòs, J.M. Sánchez and R. Rivas, “Ground measurements for the validation of land surface temperatures derived from AATSR and and MODIS data,” in Remote Sensing of Environment, vol. 97(3), 2005.

[5] J. Davis, V. S. Costa, S. Ray and D. Page, "An integrated approach to feature invention and model construction for drug activity prediction," in Proc. of the International Conference on Machine Learning, pp.217-224, 2007.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood estimation from incomplete data via the EM algorithm,” in Journal of the Royal Statistical Society, Series B 39, 1977.

[7] T. Dietterich, R. Lathrop and T. Lozano-Pérez, “Solving the multiple-instance problem with axis-parallel rectangles,” in Artificial Intelligence, vol. 89, pp. 31–71, 1997.

[8] L. K. Emmons, M. N. Deeter, J. C. Gille, D. P. Edwards, J. -L. Attie’, et al, “Validation of measurements of pollution in the Ttroposphere (MOPITT) CO retrievals with aircraft in situ profiles,” in Journal of Geophysical Research, vol. 110, D03309, 2004.

[9] B. N. Holben, T. F. Eck, I. Slutsker, T. Tanre, J. P. Buis, A. Setzer, E. Vermote, J. A. Reagan, Y. J. Kaufman, T. Nakajima, F. Lavenu, I. Jankowiak and A. Smirnov, “AERONET: a federated instrument network and data archive for aerosol characterization,” in Remote Sensing of Environment, vol. 37, pp. 2403–2412, 1998.

[10] C. Ichoku, D. A. Chu, S. Mattoo, Y. Kaufman, L. Remer, D. Tanré, I. Slutsker and B. N. Holben, “A spatio-temporal approach for global validation and analysis of MODIS aerosol products,” in Geophysical Research Letters, vol. 29, pp. 1–4, 2002.

[11] T.J. Jackson, R. Bindlish, M. Cosh, “Validation of AMSR-E Soil Moisture Products Using in Situ Observations,” in Journal of the Remote Sensing Society of Japan, vol. 29(1), pp. 263-270, 2009.

[12] I. Jobard, F. Chopin, J. C. Berges and R. Roca, “An intercomparison of 10-day satellite precipitation products during West African monsoon,” in International Journal of Remote Sensing, vol. 32 (9), pp. 2353 – 2376, 2011.

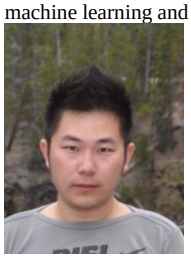
[13] S. Kalluri, P. Gilruth, D. Rogers andM. Szczur, “Surveillance of arthropod vector-borne infectious diseases using remote sensing techniques: a review,” in PLoS Pathog. 3: 116, 2007.

[14] J.T. Kerr andM. Ostrovsky, “From space to species: ecological applications for remote sensing,” in Trends Ecol. Evol. 18, pp. 299–305, 2003.

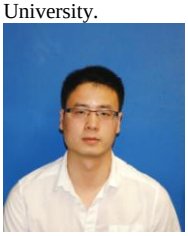
[15] G. S. E. Lagerloef, Y. Chao and F. R. Colomb, “Aquarius/SAC-D Ocean Salinity Mission Science Overview,” in IEEE International Geoscience and Remote Sensing Symposium, pp. 1675 – 1677, 2006.

[16] Y. Liu, C.J. Paciorek and P. Koutrakis, "Estimating regional spatial and temporal variability of PM2.5 concentrations using satellite data, meteorology, and land use information," in Environmental Health Perspectives, vol. 117(6), pp. 886-892, 2009.
[17] O. Maron, Learning from Ambiguity, Dept. of Electrical and Computer Science, Massachusetts Inst. of Technology, Cambridge, 1998.
[18] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in Advances in Neural Information Processing Systems 10, pp. 570–576, 2003.
[19] O. Maron and A. L. Ratan, "Multiple-instance learning for natural scene classification," in Proc. Int. Conf. on Machine Learning, pp. 341–349, 1998.
[20] M. M. Oo, M. Jerg, E. Hernandez, A. Picon, B.M. Gross, F. Moshary andS.A. Ahmed, "Improved MODIS Aerosol Retrieval Using Modified VIS/SWIR Surface Albedo Ratio Over Urban Scenes, " in IEEE Transactions on Geoscience and Remote Sensing, vol. 48(3), pp. 983-1000, 2010.
[21] S. Ray and M. Craven, "Supervised versus multiple instance learning: an empirical comparison," in Proc. Int. Conf. on Machine Learning, pp. 697–704, 2005.
[22] S. Ray and D. Page, "Multiple-instance regression," in Proc. Int. Conf. on Machine Learning, pp. 425–432, 2001.
[23] L. Remer, A. E. Wald and Y.J. Kaufman, " Angular and Seasonal Variation of Spectral Surface Reflectance Ratios: Implications for the Remote Sensing of Aerosol Over Land, " in IEEE Transactions on Geoscience and Remote Sensing, vol. 39(2), 2001.
[24] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: the rprop algorithm," in Proc. IEEE Int. Conf. on Neural Networks, 1993.
[25] G.H. Tilstone, I. M. Angel-Benavides, Y. Pradhan, J. D. Shutler, S. Groom and S. Sathyendranath, "An assessment of chlorophyll-a algorithms available for SeaWiFS in coastal and open areas of the Bay of Bengal and Arabian Sea," in Remote Sensing of Environment, vol.115 (9), pp. 2277-2291, 2011
[26] D.C. Tobin , H.E. Revercomb, R.O. Knuteson, B.M. Lesht, L.L. Strow, et al, "Atmospheric radiation measurement site atmospheric state best estimates for Atmospheric Infrared Sounder temperature and water vapor retrieval validation," in Journal of Geophysical Research, vol. 111, D09S14, 2006.
[27] K. L. Wagstaff and T. Lane, "Salience assignment for multiple-instance regression," in Proc. Int. Conf. on Machine Learning Workshop on Constrained Optimization and Structured Output Spaces, 2007.
[28] K. L. Wagstaff, T. Lane and A. Roper, "Multiple-instance regression with structured data," in Proc. Int. Workshop on Mining Complex Data, 2008.
[29] Z. Wang, V. Radosavljevic, B. Han, Z. Obradovic and S. Vucetic, "Aerosol optical depth prediction from satellite observations by multiple instance regression," in Proc. SIAM Conf. on Data Mining, 2008.
[30] C. Yang and T. Pérez, "Image database retrieval with multiple-instance learning techniques," in Proc. IEEE Int. Conf. Data Engineering, pp. 233–243, 2000.
[31] W. Yang, B. Tan, D. Huang, M. Rautiainen, N. V. Shabanov, Y. Wang, J. L. Privette, et al, "MODIS leaf area index products: From validation to algorithm improvement," in IEEE Transactions on Geoscience and Remote Sensing, vol. 44(7), pp. 1885-1898, 2006.
[32] Q. Zhang, S. A. Goldman, W. Yu and J. Fritts, "Content-based image retrieval using multiple-instance learning," in Proc. Int. Conf. on Machine Learning, pp. 682–689, 2002.

Zhuang Wang received the B. A. degree in Electronic Commerce from Wuhan University, China, in 2006 and Ph.D. degree in Computer and Information Sciences from Temple University, USA in 2010. He is currently a Research Scientist at Siemens Corporate Research, NJ. His research interests are in machine learning and data mining.



Liang Lan received the B.S. degree in Bioinformatics from Huazhong University of Science and Technology, China, in 2007 and M.S. degree in Computer and Information Science from Temple University, USA in 2009. He is currently a Ph.D candidate in Computer and Information Science at Temple University.



Slobodan Vucetic received the B.S. and M.S. degrees in Electrical Engineering from the University of Novi Sad, Serbia, in 1994 and 1997, and Ph.D. degree in Electrical Engineering from Washington State University in 2001. He is currently an Associate Professor in Department of Computer and Information Sciences at Temple University. His research interests are data mining and machine learning.