Aerosol Optical Depth Prediction from Satellite Observations by Multiple Instance Regression *

Zhuang Wang Vladan Radosavljevic Bo Han Zoran Obradovic
Slobodan Vucetic†

Abstract

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. To help address this challenge, numerous satellite sensors are employed to achieve global-scale monitoring of aerosols. Given the satellite measurements, the common objective is prediction of Aerosol Optical Depth (AOD). An important property of AOD is its low spatial variability on a scale of tens of kilometers. On the other hand, satellite sensors gather information in the form of multi-spectral images with high spatial resolution where pixels could be as small as a few hundred meters. Given an accurate ground-based AOD measurement over a specific location and time, all the pixels in the vicinity can be assumed to have the same AOD. If we treat satellite measurement at a single pixel as an instance, all pixels from the neighborhood can be considered as a bag of instances labeled with the same AOD. Given a number of bags obtained at numerous locations and at different times we can treat the problem of AOD prediction from satellite attributes as Multiple Instance Regression (MIR). An important challenge is that because of rapidly changing surface properties attribute values of pixels from a bag can vary a lot. This study evaluated several MIR approaches on several synthetic data sets and on a data set consisting of 800 labeled bags, each containing hundreds of pixel instances observed over the Continental U.S. by the MISR satellite instrument. The results indicate that the most successful MIR approach consists of an iterative procedure that detects and discards outlying instances and trains a predictor on the remaining ones.

1 Introduction

Recently, a series of low-altitude satellites (e.g. TERRA, AQUA, AURA) have been launched as a part of the Earth Observation System (EOS) that provides steady data streams from multiple instruments. These instruments provide an unprecedented opportunity for long-term global observations of the land surface, the biosphere, solid Earth, the atmosphere, and the oceans. As a basic principle of remote sensing, satellite instruments measure radiances reflected from Earth. The collected data are used to predict the underlying geophysical parameters such as atmospheric temperature profiles, cloud/aerosol properties, or vegetation cover. Achieving accurate estimations is a critical requirement for the success of the ensuing scientific studies.

Among the most challenging climate research problems today is understanding composition, abundance, and variability of aerosols, small airborne particles that reflect and absorb incoming solar radiation, on a global scale. The existing algorithms for aerosol prediction from satellite observations are typically developed and finely tuned manually by teams of domain scientists. Due to the complexity of such predictors, aerosol predictions are computationally costly and the updates of existing predictors are difficult. This gives rise to an increased interest in data-driven predictors trained directly from labeled data. In case of aerosols, the source of high quality labeled data is AERONET, a global network of ground-based radiometers that frequently and accurately measure AOD over several hundred locations throughout the world. The objective of this paper is to illustrate that aerosol prediction can be cast as multiple instance regression and to propose and evaluate several multiple instance regression algorithms.

Over the last decade a large interest has been shown in the problem of multiple instance learning (MIL). In its most general setting a learner is given a number of labeled bags, each containing a number of instances of the related type. The main difference between this scenario and the traditional supervised learning is that the target labels are assigned to the bags instead of the individual instances. The difficulty of the learning problem depends on the type and variability of instances within the bag.

The most commonly addressed multiple instance learning problem is classification, where negative bags

*This work was supported in part by the U.S. National Science Foundation under Grants IIS-0546155 and IIS-0612149.

†Center for Information Science and Technology, Temple University, Philadelphia, PA 19122, USA, {zhuang, vladan, bohan, zoran, slobodan}@ist.temple.edu

contain exclusively negative instances, while positive bags contain at least one positive instance in addition to the arbitrary number of negative instances. Interestingly, this setting covers a substantial number of applications such as drug activity prediction [10], image categorization [6] and retrieval [24], [26], text categorization [1], and stock market prediction [14]. Consequently, numerous algorithms were proposed to solve the classification problem.

On the contrary, relatively little has been done on the problem of multiple instance regression. It seems that the lack of a motivating real-life application is the major cause for such state of the matter. For example, in the representative multiple instance regression work by Ray and Page [19], the authors evaluate their algorithm only on synthetic linear regression data. One of the contributions of this paper is introduction of a remote sensing problem that nicely fits into the framework of multiple instance regression. The problem deals with remote sensing of aerosols and is explained in more detail in §2. The property of aerosol data is that bag instances are very noisy and that the data generating process is nonlinear.

Another contribution of the paper are several novel multiple instance regression algorithms described in §3 that are suitable for the posed and the related problems. The algorithms are compared to the method of Ray and Page [19], to several baseline methods, and to the operational aerosol predictor developed by the aerosol scientists. In addition to evaluation on the aerosol remote sensing data (§5), we carefully characterized these algorithms on several synthetic data sets (§4). The paper is concluded by an overview of the related work in §6.

2 Remote Sensing of Aerosols as a Multiple Instance Regression Problem

2.1 Satellite Remote Sensing Fundamentals. Remote sensing is defined as the acquisition of information about an object without being in physical contact with it [12]. A typical source of remote sensing data is electromagnetic radiation which is emitted or reflected from the observed object.

Information about our environment could be obtained by imaging the Sun's electromagnetic radiation reflected from the Earth's surface and atmosphere using cameras aboard various satellites. The Sun emits electromagnetic radiation which propagates through the vacuum of space and reaches the Earth's atmosphere. It interacts with the atmosphere, with the Earth's surface, and again with the atmosphere along its path to the imaging sensor.

The radiance observed by the satellite sensors originates from several sources (Figure 1). Path 1 contains solar radiance that was attenuated very little before reaching the surface and reflected back to the atmosphere. Path 2 contains radiance that never reaches the Earth's surface because of scattering in the atmosphere. A certain amount of that radiation is detected by the sensor system. Path 3 contains radiation from the Sun that has undergone some scattering and perhaps some absorption before reaching the surface. Path 4 contains radiation that was reflected by nearby a surface into the sensor system.

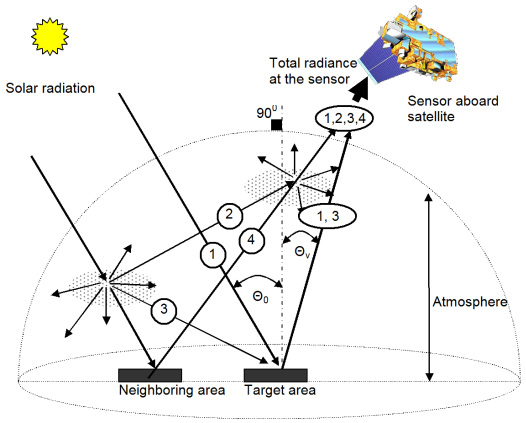

Figure 1: Various paths of radiance received by a satellite remote sensing system

2.2 Remote Sensing of Aerosols. Aerosols are a main factor that affects radiation while it travels through the atmosphere. Aerosols are characterized by Aerosol Optical Depth (AOD), a dimensionless quantity which represents the total attenuation of radiation from the top of the atmosphere down to the surface caused by aerosols. AOD is a very important quantity in building global climate models.

Remote sensing of AOD relies on the concept that solar radiation is modified by aerosols as it travels through the atmosphere. However, the total radiance observed by the satellite sensor is the combination of radiances reflected from both the atmosphere and the surface (Figure 1). To predict AOD using satellite observations, one needs to accurately determine the exact amount of radiance reflected from the atmosphere (Path 2 in Figure 1) as it conveys information directly related to AOD. In this case, radiance reflected from the surface (Paths 1, 3 and 4 in Figure 1) is considered as an unwanted noise. Extracting the Path 2 radiance from the observed radiation is a highly non-linear problem because the mixture of radiances depends on many parameters, such as aerosol and surface types, which

are not known during the prediction process.

2.3 Aerosol Prediction as Multiple Instance Regression. An important property of aerosols that can aid AOD prediction is that it has small spatial variability and is nearly unchanged over distances of up to 100 km [13]. 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 $200 \times 200m^2$. If we consider multispectral observations over a single pixel as an instance, neighboring pixels can be considered as instances from a single bag and labeled with a single AOD value. The attribute values (i.e. multispectral observations) over neighboring pixels can change widely and are a function of surface properties that can change rapidly over relatively small distances. Such a scenario clearly corresponds to *multiple instance regression* where the bag instances are noisy versions of the prime instance.

In the aerosol remote sensing domain, the prime instance would correspond to pixels over dark surfaces that are perfectly absorbing the incoming radiation. In this case, only the atmospheric component of the solar radiation (Path 2 from Figure 1) would be observed by the satellite sensor and the aerosol optical depth would be a deterministic function of the observed reflectance. Because the Earth's surface is not perfectly absorbing and is highly variable over space and time, it is not realistic to expect that bags in the aerosol prediction problem contain prime instances. This property violates the assumption of the existence of a prime instance posed in the work of Ray and Page [19].

3 Multiple Instance Regression (MIR) Approaches

3.1 Task Definition. We define the task of MIR as follows. We are given a labeled set with B bags, $D = \{bag_j, 1...B\}$, where $bag_j = \{(x_{ij}, y_j), i = 1...b_j\}$, x_{ij} is an attribute vector of the i-th instance from the j-th bag, y_i is the real-valued target value of the j-th bag, and b_j is the number of instances in the j-th bag. We assume that the instances in a bag are noisy or distorted versions of the prime instance

$$x_{ij} = p_j + \delta_{ij} \tag{3.1}$$

where p_j is the noise-free, or prime, instance and δ_{ij} is attribute noise that follows some unknown distribution. The target of bag_j is a function of the prime instance x_j with some added noise

$$y_j = g(p_j) + \epsilon_j \tag{3.2}$$

where g is the regression function and ϵ_j is the target noise with zero mean. The goal is to train a regression model that can accurately predict the target of an unseen bag given a set of its instances.

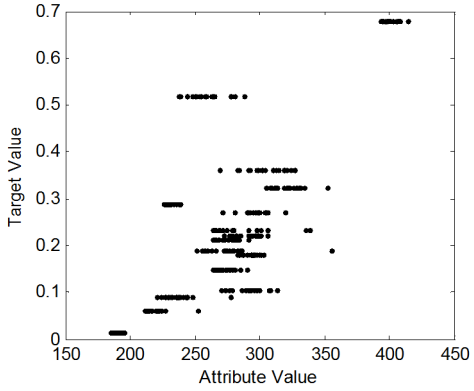

Figure 2: Example plot of 20 bags from the aerosol remote sensing data. 20 instances are shown for each bag.

To illustrate the MIR data, in Figure 2 we plot instances from 20 bags from the aerosol remote sensing data (explained in detail in §5). The figure shows the relationship between pixel reflectance at $440nm$ observed by the MISR satellite instrument and the AOD value measured by a ground-based sun photometer. Each bag shown appears as a horizontal line. It is indicative that there is large variation in the reflectance values within each of the bags. Additionally, several bags appear to have outlying AOD values that are likely caused by significant cloud presence during the measurement. The relationship between reflectance and AOD appears to be near linear.

3.2 Algorithms. In the following we describe five algorithms based on one of the two MIR approaches - the first that aggregates attribute information from all bag instances and represents each bag as a single training example; and the second that is treating each bag instance as a separate training example.

3.2.1 Aggregated-MIR. In this algorithm, bag_j is treated as a meta-instance (x_j, y_j), where x_j is obtained by averaging over all of its instances as $x_j = mean(\{x_{ij}, i = 1...b_j\})$. Then, a regression model f is trained using a set of meta-instances $\{(x_j, y_j), j = 1...B\}$. To predict the label of an unseen bag, the bag is first represented as the meta-instance (x, y) by the attribute averaging and the prediction is given as $f(x)$.

Aggregated-MIR is optimal when the attribute noise δ_{ij} from (1) has zero mean, $E[\delta_{ij}] = 0$, in all bags, $j = 1...B$, and when the bags are large. This is

because attribute averaging creates meta-instances that approach prime instances as the bag size increases.

In the case when the attribute noise does not have zero mean or when the noise distribution is characterized by heavy tails (that creates a significant number of outlying instances) Aggregated-MIR would produce sub-optimal results. In this case, alternative approaches that treat every instance as a training example might be advantageous.

3.2.2 Instance-MIR. A straightforward application of the instance-as-a-training-example approach is to represent the i-th instance from the j-th bag as (x_{ij}, y_j), join instances from all bags in a single training data set $D = \{(x_{ij}, y_j), j = 1...B, i = 1...b_j\}$, and learn a regression model from the training data. To prevent giving higher importance to large bags, Instance-MIR samples (with repetition) the same number N of instances from each bag to the training data set $D = \{(x_{ij}, y_j), j = 1...B, i = 1...N\}$.

A recent study [18] showed that, despite its simplicity, the Instance-MIR algorithm can provide competitive results compared to other multiple instance learning algorithms on many datasets.

3.2.3 Prediction of an unlabeled bag. An important issue when designing an MIR predictor is how to predict the label of an unseen bag. It is reasonable to assume that in the absence of some prior knowledge all bag instances should be given an equal chance to contribute to the final prediction. Following this assumption, a straightforward approach is to apply the resulting predictor on all instances from a bag and calculate the bag label as the average prediction. By denoting as the resulting predictor, prediction for the j-th bag is calculated as

$$\hat{y} = mean\left(\{f\left(x_{ij}\right), i = 1...b_j\}\right).$$

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

$$\hat{y} = md\left(\{f\left(x_{ij}\right), i = 1...b_j\}\right).$$

To illustrate the difference between the mean and median averaging, in Figure 3 we show the histogram of predictions of instances in a single bag in an actual AOD prediction experiment. Clearly, there are a few outlying instances, having unusually high predicted AOD and prediction averaging would positively bias the bag prediction. Nevertheless, we evaluate both mean and median averaging in §4 and 5.

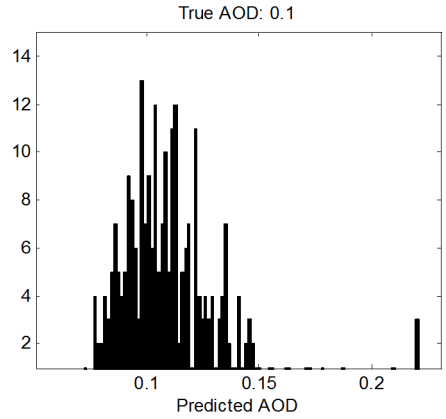

Figure 3: A histogram of the predicted AODs of all instances of a bag in MISR data. The true AOD is 0.10.

3.2.4 Prime-MIR [19]. This algorithm belongs to the instance-as-an-example approach where only a single instance from each bag is selected to the training set. It is based on an assumption that there is a prime instance in each bag, which is a representative of the bag. The remaining bag instances are assumed to be noisy versions of the prime generated by the additive noise. The regression function is assumed to be linear.

The algorithm attempts to discover the prime instances and train a predictor on them. Prime-MIR is an Expectation-Maximization (EM) like procedure. Given the currently available predictor, in the E step the algorithm selects from each bag the instance which has the lowest prediction error. These instances are called the prime candidates. In the M step, a new prediction model is trained by using the prime candidate instances. The algorithm iterates as long as the prediction error over prime instance candidates continues to decrease.

We have made several modifications to the originally proposed algorithm. (1) While the original algorithm starts from a randomly generated predictor, we use Instance-MIR algorithm to build the initial predictor. (2) We explore 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 use the mean and median averaging approaches described in §3.2.3.

3.2.5 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 attribute noise is high. Prime-MIR uses a rather sensitive procedure that does not guarantee detection of the least noisy instance. Furthermore,

as explained in §2.3, the prime instance assumption is unrealistic because bag instances in the remote sensing data are distorted versions of the prime instances. Finally, Prime-MIR uses only a small fraction of instances for training, which could prevent accurate training of more complex models.

To address these issues, we propose an EM-like procedure that in each E step discards a small fraction of "the noisiest instances". In the M step, a new predictor is trained on the remaining instances. In this manner, the algorithm is gradually removing noisy instances while using the remaining instances for training.

The algorithm runs as long as improvement in prediction accuracy is observed. Taking into the consideration that prediction of an unseen bag is obtained by mean or median prediction averaging (§3.2.3), we define the accuracy as the Mean Squared Error (MSE) of bag label predictions,

$$(3.3) MSE = \sum_{j}^{B} \left(y_j - md\left(\left\{f\left(x_{ij}\right), i = 1...N\right\}\right)\right)^2,$$

where *mean* can be used instead of *median* if desired.

The remaining question is how we define "the noisiest instances." In the following we describe two alternatives that were experimentally evaluated in Sections 4 and 5.

GlobalPruning-MIR. In this algorithm, starting from the resampled global training data set $D = \{(x_{ij}, y_j), j = 1...B, i = 1...N\}$, a regression model is trained and $r\%$ of the instances with the highest prediction error are removed. The model is retrained on the pruned data set and the procedure is repeated until convergence.

The intuition behind this algorithm is that level and properties of attribute noise in different bags could vary significantly. For example, in our aerosol application, some bags are over bright surfaces (deserts) with inherently large attribute noise; some are over highly variable terrain (mountains) with large variance in attribute values within a bag; and some are over dark surfaces (dense forests) that introduce small attribute noise. Additionally, some bags might have large target noise that can occur during cloudy days. GlobalPruning-MIR achieves uneven bag pruning for this type of data.

BalancedPruning-MIR. There are two potential issues with GlobalPruning-MIR. (1) Due to different levels of pruning it inherently weights clean bags higher than the noisy ones. In the extreme cases, it can result in the complete removal of noisy bags. This property is undesirable because it can lead to loss of very useful information. (2) Similarly to Prime-MIR, the pruning of instances with the highest error is self reinforcing - the current predictor will remove the most difficult instances in the E step and, in doing so, it will ensure that the newly trained predictor is not significantly different from the preceding one. This will impede the chances of developing a substantially improved predictor as compared to the initially trained one.

To address the first issue, each E step discards $r\%$ of "the noisiest instances" from each bag. To address the second issue, the noisiest instances in a bag are defined as those whose predictions are the farthest away from the median prediction over the non-pruned bag instances. This subtle change in the pruning criterion ensures that the algorithm is less sensitive to the choice of the initial predictor.

As a summary, in Algorithm 1 we give the pseudocode of the BalancedPruning-MIR algorithm.

Algorithm 1 BalancedPruning-MIR

Input: $D = \{bag_j, j = 1...B\}$, where $bag_j = \{(x_{ij}, y_j), i = 1...N\}$.
Output: A regression model $f(x)$

(Initial M-step)
Train $f_{new}(x)$ over D
Calculate $NewMSE$ of $f_{new}(x)$ using (3.3)
repeat
  $OldMSE = NewMSE$
  $f_{old}(x) = f_{new}(x)$
  (E-step)
  for every bag_j **do**
    for every instance i in bag_j **do**
      $Score_{ij} = \left(f\left(x_{ij}\right) - md\left(bag_j\right)\right)^2$
    end for
    Update bag_j by removing r% of its instances with the highest Score
  end for
  (M-Step)
  Train $f_{new}(x)$ over the updated bags
  Calculate $NewMSE$ of $f_{new}(x)$ using (3.3)
until $NewMSE < OldMSE$ or no instances can be removed the regression model $f(x) = f_{old}(x)$

4 Experiments on Synthetic Data

To characterize the behavior of the various MIR algorithms we performed a series of experiments on synthetic data.

4.1 Synthetic data generation. We constructed three types of MIR data generators that all follow the data generating process described in equations (1) and (2). We used a single real-valued attribute

because it allowed us to better characterize the behavior of various MIR algorithms (high dimensional aerosol remote sensing data was used in experiments described in §5). We used two types of regression functions g: linear, $g(x) = x$, and nonlinear, $g(x) = x^2$.

MIR-Gaussian(B, σ, s)**.** For each bag, $bag_j, j = 1...B$, we generated the prime instance, x_j, as a random number between 0 and 1. The bag label y_j was generated as $y_j = g(x_j) + \epsilon_j$, where ϵ_j is Gaussian additive noise with mean zero and variance σ^2. Then, we generated 100 instances within the bag as noisy versions of the prime instance as $x_{ij} = x_j + \delta_{ij}$, where δ_{ij} is Gaussian additive noise with mean zero and variance s^2. Instances of data sets generated by MIR-Gaussian are shown in Figures 6 and 9. MIR-Gaussian generator is idealistic and is suitable for the Aggregated-MIR algorithm. The remaining 4 algorithms described in §3.2 should also achieve good accuracy on such data.

MIR-Outlier1(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, 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-Outlier generates bags with different fractions of outliers. Specifically, $q_j\%$ of instances in j-th bag are generated using MIR-Gaussian generator, where q_j is a random number between 50 and 100, and the remaining $N(100 - q_j)$ instances are generated as outliers. The attribute in i-th outlier instance of j-th bag is generated as $x_{ij} = x_j + \delta_{ij} + \alpha_j$, where δ_{ij} is the Gaussian noise with variance $25s^2$ and α is an offset generated as a random number between -0.25 and 0.25. Instances of data sets generated by MIR-Outlier1 are shown in Figures 7 and 10.

MIR-Outlier2(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 $y_j = g(x_j) + \epsilon_j$, where ϵ_j is Gaussian additive noise with mean zero and variance $25\sigma^2$. Instances of data sets generated by MIR-Outlier1 are shown in Figures 8 and 11.

4.2 Experimental design. We compared Aggregated, Instance, Prime, GlobalPruning, and Balanced-Pruning MIR algorithms on 3 types of data sets explained in 4.1. The pruning parameter used in GlobalPruning and BalancedPru-ning in all experiments shown in Sections 4 and 5 was set to 5%. The choice was made because lower values result in slow convergence, while large r values result in aggressive pruning.

In the first set of experiments, we explored MIR algorithms when regression function g is linear, $g(x) = x$. Parameters in all three data generators were set to $B = 1100$, $\sigma = 0.05$, $s = 0.1$. One hundred of the generated bags were used for training, while the remaining 1,000 were used for testing. The regression model used on this data was linear regression trained using ordinary least squares algorithm. For each choice of the (MIR data generator, MIR algorithm) pair we run 20 experiments. We report the mean as one standard deviation of the Root Mean Squared Error (RMSE) calculated as the root of MSE from equation (3). The RMSE accuracy is reported for both mean and median averaging described in §3.2.2.

In the second set of experiments, we explored MIR algorithms when regression function g is nonlinear, $g(x) = x^2$. Feedforward neural networks (NN) with one hidden layer, 10 hidden sigmoid neurons, and one linear output neuron were used as the regression model. Two hundred epochs of the resilient backpropagation algorithm were used for NN training. The experimental design was identical to the first set of experiments.

Experiments on MIR-Gaussian Data. In Tables 1 and 2, we show the results on MIR-Gaussian$(100, 0.05, 0.1)$ data. Figures 6 and 9 allow us to visually compare the performances. In the linear regression case (Table 1), Aggregated and Balanced-Pruning achieved slightly but significantly better results than the other three algorithms. Good performance of Aggregated was expected on this data set, while performance of BalancedPruning highlights the strength of the instance pruning strategy. The number of iterations before convergence of both pruning algorithms is relatively small which is due to the fact that there are not many outlying instances. As expected, no significant difference was observed between the mean and median averaging.

Algorithms	RMSE×100		Iterations
	Mean	Median	
Instance	6.1±0.6	6.1±0.6	No
Aggregated	**5.2±0.4**		No
Prime	6.1±0.6	6.1±0.6	36.0±45.0
GlobalPruning	6.0±0.4	6.0±0.4	5.6±3.5
BalancedPruning	**5.2±0.3**	**5.2±0.3**	5.6±3.0

Table 1: Linear regression comparison of results on MIR-Gaussian$(100, 0.05, 0.1)$

Experiments on nonlinear regression using neural networks are shown in Table 2 and Figure 9. It could be seen that both Pruning algorithms achieved high ac-

Algorithms	RMSE×100		Iterations
	Mean	Median	
Instance	6.5±0.4	6.2±0.4	No
Aggregated	5.9±0.6		No
Prime	6.3±0.9	6.7±0.8	4.5±1.4
GlobalPruning	6.1±0.5	5.9±0.5	4.0±1.6
BalancedPruning	6.1±0.4	**5.7±0.3**	4.5±1.5

Table 2: Nonlinear regression comparison of results on MIR-Gaussian$(100, 0.05, 0.1)$

Algorithms	RMSE×100		Iterations
	Mean	Median	
Instance	14.9±1.0	14.7±1.0	No
Aggregated	6.8±0.6		No
Prime	13.6±2.6	13.3±2.9	12.6±21.3
GlobalPruning	7.4±0.4	6.3±0.4	14.4±4.2
BalancedPruning	6.9±0.5	**5.4±0.3**	17.0±0.8

Table 3: Linear regression comparison of results on MIR-Outlier1$(100, 0.05, 0.1)$

curacy, while accuracy of Aggregated slightly deteriorated, which is possibly attributable to the small training data set available to Aggregated that increased the likelihood of neural network overfitting. These overfitting problems are evident upon closer inspection of Figure 9 that shows that the Aggregate-MIR produced the most "wavy" function that indicates problems with overfitting.

It appears that median prediction averaging is slightly advantageous to the mean averaging. Overall, it is worth emphasizing that a casual look at Figures 6 and 9 shows that differences between all the competing algorithms are rather small.

Experiments on MIR-Outlier1 Data. In Tables 3 and 4 and Figures 7 and 10 we show the results on MIR-Outlier$(100, 0.05, 0.1)$ data. The first observation is that the median averaging is a significantly better choice than the mean averaging. Another important result is that BalancedPruning is consistently the best algorithm and that it is closely followed by GlobalPruning. Interestingly, Instance and Prime are significantly less accurate in the linear regression case, while the difference is not as large in the nonlinear case. The explanation can be found by looking at Figure 7 that shows severe attenuation in the learned regression function. This result is expected when strong outliers are present in the bags.

The difference between Pruning algorithms and Aggregate is further increased, which was expected due to presence of instance outliers and the more complex attribute noise. The number of iterations in Pruning algorithms increased as compared to MIR-Gaussian experiments, which is the expected behavior due to the increased instance noise.

Experiments on MIR-Outlier2 Data. In Tables 5 and 6 and Figures 8 and 11 we show the result on MIR-Outlier$(100, 0.05, 0.1)$ data. BalancedPruning and GlobalPruning are still the most accurate, while the improvement over Aggregated further increased, which is especially indicative by inspection of Figure 11. By looking at Figure 8, it could be seen that the Aggregated linear predictor is very similar to the Pruning algorithm. The reason for the difference in the prediction accuracy is the way the predictors are used on an unseen bag - attribute aggregation prior to prediction is a poor choice in presence of outlying instances. Compared to the previous two data sets, the accuracy of all algorithms further decreased due to increased target noise.

Algorithms	RMSE×100		Iterations
	Mean	Median	
Instance	12.8±1.0	10.0±1.1	No
Aggregated	7.9±0.8		No
Prime	10.2±1.5	8.1±1.5	5.5±1.7
GlobalPruning	9.3±1.1	6.9±1.2	6.0±2.0
BalancedPruning	9.7±1.0	**6.8±0.8**	6.8±2.2

Table 4: Nonlinear regression comparison of results on MIR-Outlier1$(100, 0.05, 0.1)$

Algorithms	RMSE×100		Iterations
	Mean	Median	
Instance	15.6±1.2	15.5±1.2	No
Aggregated	9.1±8.4		No
Prime	14.7±2.3	14.4±2.3	10.4±6.0
GlobalPruning	8.3±4.4	7.3±6.1	16.0±2.8
BalancedPruning	8.0±3.5	**6.7±5.1**	15.5±3.0

Table 5: Linear regression comparison of results on MIR-Outlier2$(100, 0.05, 0.1)$

Result summary. The experimental results clearly illustrate the success of the proposed Pruning algorithms on all three types of synthetic data sets. Their advantage over the alternatives grows when attribute and target noise deviate from the Gaussian distribution. Performance of Instance algorithm is consistently worst, while Aggregated suffers due to decreased training data size and when noise deviates from Gaussian. Performance of Prime resembles performance of Instance algorithm. The explanation could be found in the inherent property of Prime that makes it difficult to move from the initial solution which is given by the Instance algorithm.

Algorithms	RMSE×100		Iterations
	Mean	Median	
Instance	13.0±0.9	10.6±0.8	No
Aggregated	10.8±0.7		No
Prime	10.0±1.1	8.6±1.0	5.5±1.2
GlobalPruning	10.1±1.0	7.9±0.9	6.8±2.0
BalancedPruning	9.9±1.1	**7.7±0.9**	6.7±2.3

Table 6: Nonlinear regression comparison of results on MIR-Outlier2$(100, 0.05, 0.1)$

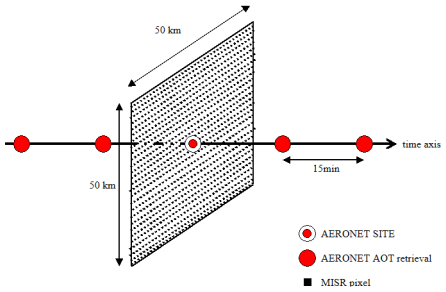

Figure 4: Spatial-temporal collocation of MISR observations and AERONET AOD prediction

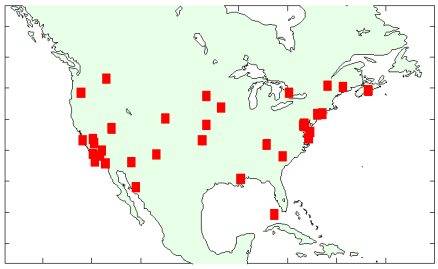

Figure 5: The location of AERONET sites used for this study

5 MIR Experiments on Aerosol Data Set

5.1 Aerosol data. We constructed the aerosol data set by merging measurements from ground-based and satellite instruments. Ground based measurements were obtained from Aerosol Robotic Network (AERONET) which is a global remote sensing network of about 540 ground-based radiometers that accurately measure AOD several times an hour under clear-sky conditions. Satellite measurements come from Multi-angle Imaging SpectroRadiometer (MISR), aboard NASA's Terra satellite, which is one of the major instruments for satellite-based AOD prediction. MISR instrument measures reflected solar radiation at four spectral bands from nine view angles along the direction of flight with high spatial resolution (one pixel is of the size from $275 \times 275m^2$ to $1.1 \times 1.1km^2$ depending on the viewe angle and wavelength) and global Earth coverage in every nine days.

Each time the satellite passes over an AERONET site and when the conditions are right (low cloud coverage), there is an opportunity to merge AERONET and MISR measurements and create a labeled bag of instances. Illustration of the merging is given in Figure 4. For this study, we used 800 bags over 35 AERONET sites in the continental U.S. (Figure 5) measured between 2002 - 2004. Each bag encompasses a region of size $50 \times 50km^2$ centered at the AERONET site. Each region contains around 2,200 $1.1 \times 1.1km^2$ pixels, and each pixel is represented as an instance. In our experiments, we selected 100 non-cloudy pixels from each region to the bag. The instance attributes were taken to be the 12 MISR reflectances from the middle 3 MISR cameras and solar and view zenith angles, while the bag target value was taken to be AERONET AOD measurement.

5.2 Experimental design. We evaluated the 5 proposed MIR algorithms on the described aerosol data. We used several types of regression models with different levels of complexity - linear regression, neural networks with 10 hidden nodes, and ensembles of 5 neural networks with 10 hidden nodes. The exact neural network architecture and training algorithm were the same as used in §4. Accuracy of each algorithm was evaluated using 5-cross-validation (5CV) where the 800 bags were randomly split into 5 groups; 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 5CV was repeated 20 times and the average RMSE and their standard deviation are reported in the tables. Both mean and median averaging methods for bag predictions were evaluated. In addition to the proposed MIR algorithms we measured accuracy of two other algorithms - Mean predictor and MISR operational algorithm.

Mean predictor. Average AOD from the training bags was used as prediction of AOD on test bags. This trivial predictor is useful for benchmarking of other algorithms.

MISR operational algorithm. MISR AOD predictor is a complex deterministic algorithm finely tuned by domain scientists. It is based on matching the atmospheric component of the observed reflectance to the simulated values stored in lookup tables, where the lookup tables are generated by the forward simulation model that estimates atmospheric radiance given the aerosol type and amount. MISR predictor is known as

the best existing AOD predictor from satellite observations.

5.3 Experimental results. Table 7 provides a comparison of the 5 MIR algorithms with linear regression and the Mean predictor (MEAN). It could be seen that all MIR algorithms achieve significant improvement over MEAN. Among them, BalancedPruning with median prediction averaging is the most successful. The number of its iterations indicates that near half of the instances were removed from the bags during the procedure.

Table 8 shows results when neural networks were used as predictors in the MIR algorithms. The accuracies of all algorithms excluding Aggregated significantly improved that indicating that the underlying data generating process is nonlinear. Consistent with the previous results, BalancedPruning is again the most successful MIR algorithm. It is interesting to observe that all 4 instance-as-an-example algorithms performed similarly well and significantly better than Aggregated. The decreased performance of Aggregated can be found in its relatively small size (640 training examples) compared to the 14 attributes used for prediction.

Algorithms	RMSE×100		Iterations
	Mean	Median	
MEAN	18.6		No
Instance	10.2±0.0	9.8±0.0	No
Aggregated	9.4±0.0		No
Prime	10.6±0.2	9.9±0.0	10.3±0.6
GlobalPruning	10.1±0.0	10.0±0.0	4.0±0.2
BalancedPruning	**9.5±0.0**	**9.1±0.0**	8.0±0.5

Table 7: Comparison of accuracies of MIR predictors that use linear regression

Algorithms	RMSE×100		Iterations
	Mean	Median	
MEAN	18.6		No
Instance	8.3±0.1	8.0±0.1	No
Aggregated	9.4±0.0		No
Prime	8.5±0.3	8.3±0.3	10.3±0.6
GlobalPruning	8.1±0.2	7.9±0.2	4.2±0.9
BalancedPruning	8.0±0.1	**7.7±0.1**	3.8±0.5

Table 8: Comparison of accuracies of MIR predictors that use neural networks

Finally, Table 9 provides comparison of different BalancedPruning algorithms with MISR operational predictor. It could be seen that BalancedPruning with an ensemble of neural networks could lead to further accuracy improvements. The ensemble method referred to in Table 10 consisted of bagging with 5 neural networks using random sampling with replacement from the non-pruned instances. The accuracy of this predictor reaches the accuracy of the MISR predictor. This is a highly promising result considering the complexity and the man-power exerted to develop the MISR predictor.

Algorithms	RMSE×100
BalancedPruning (LR)	9.1±0.0
BalancedPruning (NN)	7.7±0.1
BalancedPruning (ensemble)	7.5±0.0
BalancedPruning (NN+ MISR AOD)	6.6±0.2
BalancedPruning (ensemble+MISR AOD)	**6.4±0.0**
MISR Operational Predictor	7.5

Table 9: Comparison of accuracies of several types of BalancedPruning predictors with MISR predictor

We further explored the performance of BalancedPruning by adding AOD prediction provided by the MISR operational algorithm as an additional attribute. This inclusion is quite valid because MISR predictor can be considered as a nonlinear mapping of the original attributes. When this additional attribute is used, BalancedPruning with a neural network ensemble decreases the RMSE of 15% as compared to the MISR predictor alone.

6 Related Work

In this section we give an overview of the work related to multiple instance regression.

Multiple Instance Classification. The previous work can be roughly divided into three approaches. In LOCALIZED-MIL the goal is to find locations in the attribute space that are close to instances from the positive bags and are distant from the negative bag instances. It includes the algorithm from Dietterich et al. [10] that finds a rectangle region, which includes at least one instance from each positive bag and excludes all instances from negative bags. The Diverse Density algorithms [15],[25] find discriminative locations using a probabilistic measure. Recently, an SVM-based algorithm [5] was proposed that extends this idea to multiple discriminative locations. In this approach, given an unseen bag, the bag is positive if at least one instance is near the selected locations.

AGGREGATED-MIL represents bags as meta-instances and relies on measuring the similarity between bags. The Citation-KNN algorithm [22] uses the minimum Euclidean distance between the instances in two bags, while statistic and normalized set distances were proposed in conjunction with the standard SVMs.

MAX-MIL relies on learning a prediction function such that the maximum prediction among instances of

a positive bag is larger than that of a negative bag. One class of algorithms is achieving this by changing the optimization criterion in neural networks [17] or SVMs [3]. Another class relies on iterative retraining of a prediction model where after each iteration the instances from a positive bag are relabeled or removed from the training set [1].

It is worth noting that researchers started addressing extensions of the original multiple instance learning problem. This includes a scenario where all instances in a bag contribute equally and independently to the bag's class label [23] and a scenario where bag labels are noisy [7].

Multiple Instance Regression. The regression side of the problem has been addressed only sparingly. One direction addresses minor extension of the standard multiple instance problem - bag labels are real numbers in the [0,1] range and the bag label is assumed to be equal to the maximum label among its instances [2]. This restrictive assumption is applicable only to problems resembling the standard multiple instance classification, such as drug activity prediction [11]. The linear regression problem where each bag contains a prime instance responsible for the real-valued bag label was addressed by Ray and Page [19] and was already discussed in this paper. Finally, it is worth mentioning the recent effort for determining relative importance of bag instances through constrained optimization [21]. Similarly to the prime instance algorithm, this work does not attempt to use the resulting predictor on the unseen bags.

Learning with measurement errors. In a number of multiple instance learning scenarios, including the aerosol remote sensing problem studied in this paper, bags consist of noisy or distorted versions of the prime instance, which itself is not available to the learner. Such problem can be described as learning in the presence of (possibly repeated [9]) measurement noise and it has been studied in statistics [4]. A well-known result is that applying parametric regression or classification algorithms on data with measurement errors results in biased models with attenuated parameters. A similar behavior has been observed in neural networks where measurement error serves as a form of regularization that could even be useful for reducing overfitting. There are numerous methods developed in statistics for correcting the bias due to measurement errors (e.g. [8]). However, the existing approaches are designed for parametric models and assume familiarity with the data generating process and the type of the measurement noise (most often, additive Gaussian error is assumed). This is quite restrictive for many data mining applications where very little is known about the data or when complex nonparametric methods (e.g. neural networks, decision trees, SVMs) are used for learning.

Supervised Instance Learning. In a recent study [18] it was illustrated that supervised learning where all instances are labeled the same as their bag label often provides competitive results to MIL methods on real-life data. The success of this approach, however, depends on the type and variability of bag instances. For example, if positive bags contain only a few positive instances [3], such an approach is questionable. However, if bag instances are noisy copies of the prime instance, as in many multiple instance regression problems, such an approach might be quite reasonable.

7 Conclusions

In this paper we evaluated several multiple instance regression algorithms on synthetic data sets and on the remote sensing data for aerosol prediction. Among the described algorithms, our experiments showed that the proposed instance pruning approach is highly successful and can lead to very accurate predictions. An interesting question for future research is if domain knowledge, spatial-temporal information, and ancillary attributes can further improve the quality of remote sensing algorithms based on multiple instance regression.

References

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

[2] R. A. Amar, D. R. Dooly, S. A. Goldman and Q. Zhang, *Multiple-instance learning of real-valued data*, in Proc. of ICML'01, 2001, pp. 3–10.

[3] R. C. Bunescu and R. J. Mooney, *Multiple instance learning for sparse positive bags*, in Proc. of ICML'07, 2007.

[4] R. J. Carroll, D.Ruppert, C. Crainiceanu and L. A. Stefanski, *Measurement Error in Nonlinear Models: A Modern Perspective*, Second Edition, London: CRC Press, 2006.

[5] Y. Chen, J. Bi, and J. Wang, *MILES: multiple-instance learning via embedded instance selection*, IEEE Trans Pattern Anal Mach Intell, (28)12 2006, pp. 1931–1947.

[6] Y. Chen and J. Z. Wang, *Image categorization by learning and reasoning with regions*, J. Machine Learning Research, (5) 2004, pp. 913–939.

[7] P. M. Cheung and J. T. Kwok, *A regularization framework for multiple-instance learning*, in Proc. of ICML'06, 2006, pp. 193–200.

[8] J. R. Cook and L. A. Stefanski, *Simulation-extrapolation estimation in parametric measurement error models*, J. American Statistical Association, (89) 1994 pp. 1314–1328.

[9] M. Davidian and D. M. Giltinan, *Nonlinear Models for Repeated Measurement Data*, Chapman and Hall, 1995.
[10] T. Dietterich, R. Lathrop and T. Lozano-Pérez, *Solving the multiple-instance problem with axis-parallel rectangles*, Artificial Intelligence, (89) 1997, pp. 31–71.
[11] D. R. Dooly, S. A. Goldman and S. S. Kwek, *Real-valued multiple-instance learning with queries*, J. Comput. Syst. Sci., (72)1 2006, pp. 1–15.
[12] C. Elachi, *Introduction to the Physics and Techniques of Remote Sensing*, John Wiley & Sons, New York, 1987.
[13] C. Ichoku, D. A. Chu, et al., *A spatio-temporal approach for global validation and analysis of MODIS aerosol products*, Geophys. Res. Lett., 29(12), 2002, pp. 1–4.
[14] O. Maron, *Learning from ambiguity*, Dept. of Electrical and Computer Science, Massachusetts Inst. of Technology, Cambridge, 1998.
[15] O. Maron and T. Lozano-Pérez, *A framework for multiple-instance learning*, in Advances in Neural Information Processing Systems, 2003, pp. 570–576.
[16] O. Maron and A. L. Ratan, *Multiple-instance learning for natural scene classification*, in Proc. ICML'98, 1998, pp. 341–349.
[17] J. Ramon and L. De Raedt, *Multi instance neural networks*, in Proc. ICML'00 Workshop Attribute-Value and Relational Learning, 2000.
[18] S. Ray and M. Craven, *Supervised versus multiple instance learning: an empirical comparison*, in Proc. ICML'05, 2005, pp. 697–704.
[19] S. Ray and D. Page, *Multiple-instance regression*, in Proc. ICML'01, 2001, pp. 425–432.
[20] G. Ruffo, *Learning single and multiple decision trees for security applications*, PhD Dissertation, Dept. of Computer Science, Univ. of Turin, Italy, 2000.
[21] K. L. Wagstaff and T. Lane, *Salience Assignment for Multiple-Instance Regression*, In Proc. ICML'07 Workshop on Constrained Optimazation and Structured Output Spaces, 2007.
[22] J. Wang and J. D. Zucker, *Solving the multiple-instance problem: a lazy learning approach*, in Proc. ICML'00, 2000, pp. 1119–1125.
[23] X. Xu and E. Frank, *Logistic regression and boosting for labeled bags of instances*, in Proc. of the Paciffic-Asia Conf. on Knowledge Discovery and Data Mining, 2004, pp. 272–281.
[24] C. Yang and T. Pérez, *Image database retrieval with multiple-instance learning techniques*, in Proc. IEEE Int'l Conf. Data Eng., 2000, pp. 233–243.
[25] Q. Zhang and S. A. Goldman, *EM-DD: an improved multiple-instance learning technique*, in Advances in Neural Information Processing Systems, 2002, pp. 1073–1080.
[26] Q. Zhang, S. A. Goldman, W. Yu, and J. Fritts, *Content-based image retrieval using multiple-instance learning*, in Proc. ICML'02, 2002, pp. 682–689.

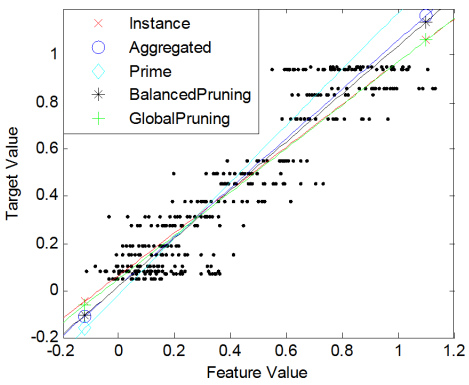

Figure 6: MIR algorithms using linear model on MIR-Gaussian$(100, 0.05, 0.1)$

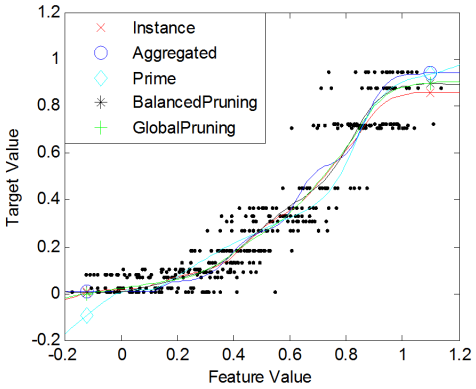

Figure 9: MIR algorithms using nonlinear model on MIR-Gaussian$(100, 0.05, 0.1)$

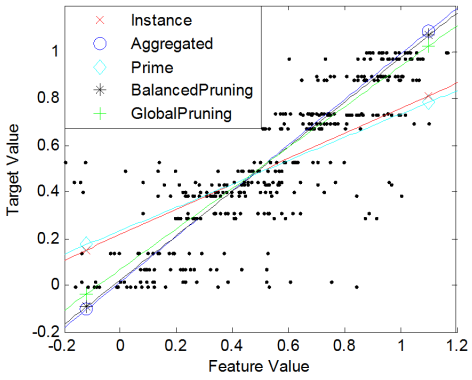

Figure 7: MIR algorithms using linear model on MIR-Outlier1$(100, 0.05, 0.1)$

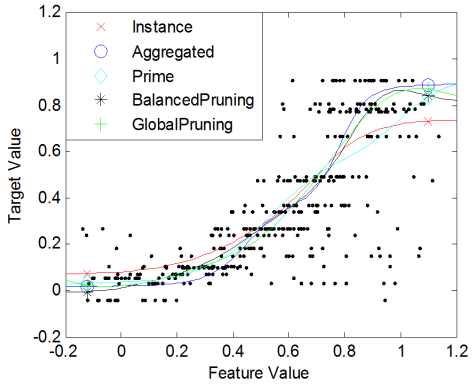

Figure 10: MIR algorithms using nonlinear model on MIR-Outlier1$(100, 0.05, 0.1)$

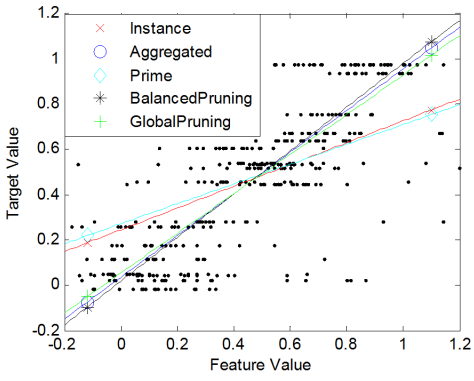

Figure 8: MIR algorithms using linear model on MIR-Outlier2$(100, 0.05, 0.1)$

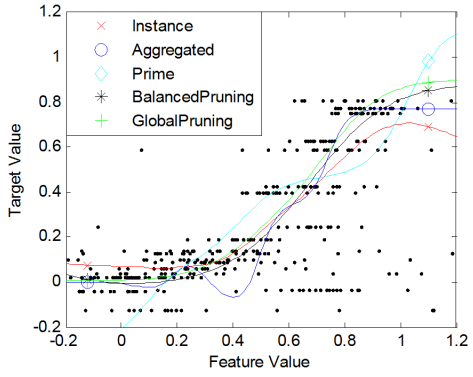

Figure 11: MIR algorithms using nonlinear model on MIR-Outlier2$(100, 0.05, 0.1)$