Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 Keller Hall
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 17-007

Multiple Instance Learning for bags with Ordered instances

Guruprasad Nayak, Varun Mithal, Vipin Kumar

June 7, 2017
Multiple Instance Learning for bags with Ordered instances

Guruprasad Nayak, Varun Mithal, Vipin Kumar
Department of Computer Science and Engineering
University of Minnesota - Twin Cities
{nayak, mithal, kumar}@cs.umn.edu

Abstract

Multiple Instance Learning (MIL) algorithms are designed for problems where labels are available for groups of instances, commonly referred to as bags. In this paper, we consider a new MIL problem setting where instances in a bag are not exchangeable, and a bijection exists between every pair of bags. We propose a neural network based MIL algorithm (MILOrd) that leverages the existence of such a bijection when learning to discriminate bags. MILOrd has an input node for each instance in the bag, an output node that captures the bag level prediction, and a hidden layer that captures the output from an instance level classifier for each instance in the bag. The bag level prediction is obtained by combining these hidden layer values using a function that models the importance of each instance, unlike the traditional schemes where each instance is considered equal. We demonstrate the utility of the proposed algorithm on the problem of burned area mapping using yearly bags composed of multispectral reflectance data for different time steps in the year. Our experiments show that MILOrd outperforms traditional MIL schemes that don’t account for the presence of a bijection.

1 Introduction

In the traditional setting of classification, one is provided with features and labels for all training instances and task at hand is to learn a function that maps a given instance to a class. However, in some cases, instance labels might be difficult or expensive to obtain and it might only be possible to label instances collectively in groups. This is the premise of the Multiple Instance Learning (MIL) setting [Dietterich et al., 1997], where the training set includes features for all instances but labels are available only for groups of instances, called bags. The task is then to learn a classifier that can learn to classify a new bag into any of the predefined classes. Figure 1 shows the problem setting for traditional MIL problems. Each dot represents an instance and instances within the same bag are enclosed in a bag shown by dotted lines. The instances are color coded based on their class labels. The red dots correspond to positive instances, while the black dots denote negative instances.

One application where we have to work with this form of weak supervision is the problem of producing yearly burned area maps using remote sensing data. Given the sequence of reflectance data observed at every location for every time step of a particular year, the goal is to determine if the location was burned in that year. Such an algorithm can be used to generate burned area maps for the forests in the tropical belt (e.g. Amazon and South-East Asia) where fires are often also associated with expanding deforestation fronts and linked to illegal activities, such as clearing land for plantations, which have severely endangered the wild life in the region, caused considerable air pollution and accelerated climate change. To achieve this, ideally, we would want a training set with labels for every time step in the year to learn a classifier that determines whether the location appears burned at every time step or not. However, there are two challenges in this setting -

Challenge C1. Unavailability of per time step labels
- Maintaining accurately dated records of burned area is an expensive ordeal. In many cases where fire events are dated, a single date is recorded for every spatially contiguous fire event. These fire events can span up to hundreds of sq.km area and hence there is a lot of variation in the time when every pixel within the spatial event burns. Moreover, the date when a location starts to burn does not necessarily correspond to the date when the ground starts showing a burn scar in the satellite imagery due to the smoke occluding the satellite imagery post-fire. Hence, it is difficult to obtain per time step labels for the burned locations in our training set and labels are instead available at yearly scales.

Thus, this problem can be viewed as a MIL problem where
every location is a bag described by a collection of feature vectors corresponding to the reflectance values observed at the location for each time step in the year. However, traditional MIL algorithms assume that the presence of a positive instance in a bag makes the bag positive and that every instance in a negative bag is negative. This assumption does not hold true for our application because of challenge C2.

**Challenge C2. Class confusion in sequences of unburned locations.** There exists confusion between burned and unburned pixels in the spectral features used to separate them. As an example, wetlands often get confused with burn scars in the spectral space. This class confusion in the spectral space leads to some unburned locations being assigned to burned class. Even though such confused unburned locations are only a small fraction of unburned locations, since burned locations are a rare class, these spurious positives can significantly impact precision. Moreover, there can be a large variety of sources of these kind of confusions that can be infeasible to exhaustively account for. Thus, one cannot simply assume that if a location looks burned anywhere in the year, it is a truly burned location.

One approach to avoid such false positives is to leverage the temporal pattern present in the positive instances in positive bags. Burn activity is associated with a fire season, and most burns occur in these specific time steps. Therefore, burn signatures occurring in other time steps are more likely due to spurious positives than actual burn events. As an example, this challenge is illustrated in figure 2. For a small region in California, we built a classifier to distinguish reflectance value on a burn date from the reflectance value on a date of healthy vegetation. To do this, we hand crafted a training set consisting of reflectance values from the precise date of burn event for burned locations and reflectance values for healthy vegetation through different time steps of the year from the unburned locations. Note that determining the precise time of fire required careful manual inspection of each burned location and is impractical to do every time. For features, we used reflectance data from MODIS available every 8 day (46 observations in a year). Figure 2 shows the probability of burning for 4 locations for every time step in the year. The time series shown in 2a, 2b and 2c are for unburned locations while the time series in 2d corresponds to a burned location. This region experienced fire in the second half of the year (time steps 30-45). As expected, the burned location in figure 2d shows a high probability in this time and close to zero probability for the remaining time steps. We would expect unburned sequences to show close to zero probability for all time steps. However, note how the unburned locations shown here have high probability values in a number of time steps throughout the year. Especially, the unburned location corresponding to figure 2c shows a high probability value for a contiguous time window that includes the true time window of burn activity in the region. In fact the number of time steps where this location shows high probability is more than that of a true burned location. Hence, even if we had limited our analysis to the fire season (time steps 30-45), we would still have incorrectly classified locations in figure 2b and figure 2c as burned.

Thus, because of C2, traditional MIL algorithms that work on the presence based assumption will incur a number of false positives and will suffer from poor precision. Traditional MIL algorithms treat instances within a bag independently and do not pay attention to the inherent ordering that exists among them in this application. In this paper, we propose to use the pattern of occurrence of positive instances with respect to the instance ordering in addition to the feature values for each instance while classifying bags. More formally, we consider a MIL problem setting with the following characteristics:

1. The instances inside a bag are not exchangeable, i.e. the instances within each bag follow an order that defines the composition of the bag in addition to the features that define those instances. In the burned area mapping problem, this ordering is defined by the time step associated with each observation (instance) for a location (bag).

2. There exists a bijection between every pair of bags, i.e. there is a one-to-one correspondence between instances in one bag to the instances in the other (meaning that instances in the same position in the ordering across two bags are comparable).

3. There is a fixed pattern of occurrence of positive instances (with respect to the instance ordering) in a positive bag, i.e., the signature of a positive bag depends on this instance ordering in addition to the features of it’s constituent instances. In the context of the burned area mapping problem, this tries to capture the right seasons for fire occurrence.

We propose a neural network based solution to this MIL problem. There are two key benefits to this formulation-

**B1. Increased precision** - By accounting for the right positions for the positive instances to occur in the instance ordering, this classifier is able to handle false positives described in C2.

**B2. Reduced sample complexity** - Since all bags in our problem contain the same number of instances, one approach that we could take is to concatenate the feature vectors for each time step in the year and learn a traditional classifier like SVM on this concatenated feature space. However, due to the high dimensionality of the concatenated feature space, we would require a large number of training samples to attain good performance. Since a MIL algorithm would use the same instance level classifier across all time steps in the year, it has lower model complexity and hence has a smaller requirement on the number of training samples.

In this paper, we make the following contributions

1. We propose a MIL algorithm (MILOrd) to classify bags with ordered instances. By leveraging the structure in the data, MILOrd outperforms existing MIL schemes in datasets where the discriminating information lies not just in the instance level features but also in their relative ordering.

2. We demonstrate the utility of the proposed scheme on the problem of burned area detection which has profound societal and environmental importance.
Probability of a time step being positive

0.2
0.4
0.6
0.8
1

Probability of a time step being positive

0.2
0.4
0.6
0.8
1

Probability of a time step being positive

0.2
0.4
0.6
0.8
1

Probability of a time step being positive

0.2
0.4
0.6
0.8
1

Figure 2: Probability estimates of an instance level (per time step) classifier for negative and positive bags corresponding to the problem of burned area classification for the 2008 forest fires in California.

2 Related Work

Multiple Instance Learning (MIL) refers to the class of learning algorithms designed for cases where class labels are available only for groups of instances (called bags) while features are available for every instance in the training set. The goal typically, is to learn a classifier that can predict the label for a new bag of instances. The first instance of this paradigm proposed in [Dietterich et al., 1997] was motivated by the problem of drug activity prediction. Subsequent works in this area have applied it to several other domains including image classification ([Chen et al., 2006]), text categorization ([Kotzias et al., 2015]), spam filtering ([Quadrianto et al., 2009]), fraud detection ([Juszczak et al., 2009] et cetera. For a detailed reading on MIL, the reader is referred to surveys [Amores, 2013], [Foulds and Frank, 2010], [Cheplygina et al., 2015].

MIL algorithms can be broadly divided into two categories - techniques that use instance level discrimination information and techniques that use bag level discrimination information, referred to as instance space algorithms and bag space algorithms respectively in [Amores, 2013]. Instance space algorithms assume that the instances comprising each bag have hidden labels associated with them and the bag level labels are generated by aggregating these instance level labels. Methods that fall in this category usually predict both instance and bag level labels. Most of the early MIL algorithms ([Dietterich et al., 1997], [Zhang and Goldman, 2001], [Andrews et al., 2002], [Ray and Craven, 2005]) used the max aggregation function i.e., a bag would be considered positive if any of its constituent instances are positive. This is referred to as the presence based assumption in [Weidmann et al., 2003]. The presence based assumption can be generalized to threshold based assumption wherein at least a certain fraction of instances in the positive bag has to be positive while all instances in the negative bag are negative. Examples in this category include [Duan et al., 2011], [Li et al., 2011], [Hajimirsadeghi et al., 2013]. These methods encode proportion of positive instances in a bag as a parameter which is either priorly known or fixed using cross validation. Most of these methods assume some kind of relationship between the labels of instances in a bag to the label of bag based on the number of instances in the bag marked positive. For instance, a bag might be marked positive if at least one instance in it is positive or at least $k$ or exactly $k$ of the $T$ instances are marked positive. We make no such explicit assumption about the relationship between instance level labels and bag level labels in this work.

On the other hand, algorithms that use bag level discrimination information [Amores, 2013] directly compute the bag level label without computing the instance level labels. This is either done by transforming each bag into a single feature vector ([Zhou et al., 2009], [Nowak et al., 2006]) or defining a kernel/similarity measure on bags ([Gürtner et al., 2002], [Wang and Zucker, 2000]), followed by application of any standard instance level classifier like logistic regression, SVM and K-NN. A simple bag space algorithm is to construct a concatenated feature vector of all instance features in the bag and learn an instance level classifier like a neural network or logistic regression on it (we call these schemes stackedTS ANN and stackedTS LR respectively and we will use them as baselines later). While it might be possible to define a bag level feature transformation or a kernel that accounts for the instance ordering in bags, to the best of our knowledge, currently, there are no methods that take this approach.

In the context of time series classification, there are techniques that use temporal autocorrelation in the data to handle time steps with noisy values in the feature space. However, in our case, since the noise itself is often temporally autocorrelated, these kind of methods are not expected to perform well.

3 Method

Problem Setting

Consider a data set consisting of bags with non exchangeable instances that have a one-to-one correspondence across bags. More formally, the training set is a collection of $N$ bags of instances with features $\{X_i^T\}_{i=1}^T$ corresponding to each of the $T$ instances in the $i$th bag. Moreover, the $i$th instances $(i \in \{1, \ldots, T\})$ across any two bags are comparable. Given this problem setting, the goal is to predict the label for an unseen test bag of instances. We aim to solve this problem in the case where class labels for bags and instances are both binary ($y_i \in \{0, 1\}$ and $Y_I \in \{0, 1\}$ for all instances $i$ and bags $I$).

MILord

Figure 3 shows the proposed neural network model. The neural network is trained to output the bag label given instance features as input. Each of the input nodes is a feature corresponding to each instance in the bag. At the hidden layer, features for each instance are combined using an activation function to produce an instance level score. These scores are then aggregated using the activation function at the output layer to give the bag level label. Note that unlike the standard neural network with one hidden layer, in MILord, there is a correspondence between a hidden layer node and an input node. Edges connecting the features for one instance to the
The constant $\lambda$ controls the relative weight of the two terms. This optimization problem can be solved with batch gradient descent using the standard back propagation algorithm. The parameter $\lambda$ is fixed using cross validation.

4 Evaluation

Datasets

We demonstrate the proposed method on the application of burned area detection using multitemporal multispectral reflectance data. This reflectance data is captured by NASA’s Moderate Resolution Imaging Spectroradiometer (MODIS) sensor aboard earth orbiting satellites Aqua and Terra. The raw data is spatially stitched together, corrected for atmospheric errors and made available through NASA archives as 8-day spectral composites at a spatial resolution of 0.25 sq.km globally. The goal here is to take this reflectance data at every vegetated location for a calendar year and determine if the location was burned in that year. Vegetated locations are determined based on MODIS land cover product, also available through the NASA archives [LPDAAC, 1]. We evaluate our method and baselines on four case studies across the country. The characteristics of the four data sets $^1$ are summarized in table 1.

Evaluation metrics

As is evident from table 1, burned area classification is a rare class problem. The number of burned locations is less than 5% of the total number of locations for any given region and land cover. In such skewed data scenarios where the objective is to detect the rare class, an evaluation metric that places importance on precision (user’s accuracy) and recall (producer’s accuracy) rather than accuracy is more appropriate [Tan and others, 2006], [Giglio et al., 2009]. We choose the harmonic mean of precision and recall (F-measure) while reporting our results in this work.

Baselines

We evaluate the proposed algorithm on each of the test cases in table 1 and compare the performance with five baseline approaches. For the first two baselines, we use the stacked

$^1$All data and code used here can be downloaded from https://goo.gl/eOIint
feature vector obtained by concatenating all instances in the ordering. The first baseline StackedTS LR is a logistic regression classifier learned on the stacked feature space. The second baseline StackedTS ANN is an Artificial Neural Network containing one hidden layer with $T$ nodes (size of each bag) learned on the stacked feature space. The third baseline MI/LR is a MIL scheme proposed in [Ray and Craven, 2005] that uses a logistic regression classifier at the instance level and generates bag level labels with the max aggregation. The fourth baseline RMIMN proposed in [Hajimirsadeghi et al., 2013] is a MIL scheme that implements the threshold based assumption. The fifth baseline Sliding MI/LR is a time series MIL approach similar to that used by the authors in [Stikic et al., 2011]. For each time step, a feature is constructed by concatenating the features for a temporal window of 5 time steps around it. Results are reported for the MI/LR algorithm run on these new features. Finally, we use MIOrdHet as the sixth baseline.

R1. Classification performance - Table 2 shows the comparative results measured as F-measure for the three test cases. In every test case, a set of 1000 samples comprising of randomly chosen 500 burned and 500 unburned locations were used for training and the remaining samples are used for testing. Every algorithm being evaluated outputs a continuous value for every test sample. Depending on the threshold chosen to binarize this output, the resulting F-score on the test set will vary. The results are reported based on the best attainable F-measure on the test set over the entire range of thresholds. Table 2 shows the average F-measure over 10 random partitions of the entire range of thresholds. Table 2 shows the average F-measure over 10 random partitions of train and test sets. The standard deviation in each case is reported within parentheses. The number of training samples are varied on one of the test cases. It can be seen that when the number of training samples is low, MIOrd has better F-measure and lower standard deviation. MIOrd also reaches it’s optimum performance in fewer samples. However, given a large number of training samples, MIOrdHet starts to approach the performance of MIOrd. Since, MIOrdHet and stackedTS ANN are more complex models, given enough training samples, these schemes will eventually perform at least as good as MIOrd that assumes a certain structure to hold true in the data. Note that, it may not always be possible to find a large number of training samples for every landcover around all geographies around the globe. This indicates that (a) it is useful to model the instance ordering in the MIL algorithm for increased precision and that (b) it is useful to model the multiple instance nature of the data to reduce the sample complexity requirement for optimal performance.

4.1 Burned area detection in tropical forests

Mapping tropical forest fires is challenging because of two reasons. Acquiring accurately labeled samples corresponding to burned and unburned locations is an expensive and time taking process because of which most areas outside North America including the tropics don’t have any labeled training data to build classification models even at the bag level. Moreover, due to the large variation in the spectral signatures of fires across geographies and biomes, classification models trained to map forest fires in North America cannot be directly applied to tropics. The presence of a thermal anomaly signal is used as a proxy to generate labels to train burned area classification model in these regions [Giglio et al., 2009]. However, the thermal anomaly signal in not accurate in pinpointing the date of burn scar visibility at the location. There is typically a time lag between when the burn scar becomes visible in the satellite image and the time step when the sig-

Table 1: Summary of data characteristics for the test locations

<table>
<thead>
<tr>
<th>Region</th>
<th>Land cover</th>
<th>Year</th>
<th>Number of (+) bags</th>
<th>Number of (-) bags</th>
<th>Fire season (approx.)</th>
<th>Number of bags (N)</th>
<th>Number of instances in a bag (T)</th>
<th>Dimensionality of feature space (D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>South California</td>
<td>Forests</td>
<td>2008</td>
<td>2,625</td>
<td>139,275</td>
<td>Jul-Dec</td>
<td>141,900</td>
<td>46</td>
<td>7</td>
</tr>
<tr>
<td>Montana</td>
<td>Forests</td>
<td>2007</td>
<td>6,313</td>
<td>269,997</td>
<td>Aug-Sep</td>
<td>276,310</td>
<td>46</td>
<td>7</td>
</tr>
<tr>
<td>South California</td>
<td>Savannas</td>
<td>2007</td>
<td>3,192</td>
<td>195,095</td>
<td>Jul-Aug</td>
<td>198,287</td>
<td>46</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 2: Bag level evaluation showing the mean of F-measures for each algorithm for different test cases computed over 10 random partitions of training and test sets. The standard deviation in each case is reported within parentheses.
The burned area maps generated using MILOrd have much greater coverage than the current state-of-the-art NASA burned area product MCD64A1 [Giglio et al., 2009] in the tropics. Also, since MILOrd automatically accounts for the fire seasons, it is much more robust to false positives without having to rely on hand-crafted rules to correct for them as existing methods use. An extensive validation of the tropical burned area product using multiple lines of evidence including high resolution satellite imagery and domain specific indices is available in our technical report [citation not provided because of the triple blind review cycle].

5 Discussion

The vector \( w \) captures the weight placed on each instance in the ordering when instance level scores are aggregated to compute the bag level score. We expect that discriminative time steps will be assigned higher magnitude weights while the weights assigned to non-discriminative time steps should be close to zero. Figure 6 shows the plot of the weights connecting the hidden layer to the output node in the network learned for classifying the 2008 forest fires of southern California. As expected, time steps 23-39 in the second half of the year (time steps 223-46) have been assigned higher positive weights since these correspond to the fire season. Some time steps get assigned close to zero weight because they're probably not very discriminating. However, it is interesting to note that some time steps in the first half get assigned high magnitude negative weights. This implies that a positive looking instance at this time step increases the probability of the bag being negative. Modeling the pattern of occurrence of positive instances helps MILOrd get rid of such potential false positives. Also, note that although these error might not happen in a large fraction of instances in the whole region, but since these are rare class problems, even a small false positive rate leads to a low precision.

6 Conclusion

We proposed a new MIL algorithm MILOrd for bags with ordered instances where bags are characterized by specific pattern of occurrence of positive instances with respect to the instance ordering. The results in this paper indicate that leveraging the pattern of occurrence of positive samples (time steps) in positive bags (years) can considerably improve the performance of burned area mapping task.
References


