Salience Assignment for Multiple-Instance Regression

Kiri L. Wagstaff KIRI.WAGSTAFF@JPL.NASA.GOV
Machine Learning and Instrument Autonomy Group, Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109 USA

Terran Lane TERRAN@CS.UNM.EDU
Dept. of Computer Science, University of New Mexico, Albuquerque, NM 87131 USA

Abstract

We present a Multiple-Instance Learning (MIL) algorithm for determining the salience of each item in each bag with respect to the bag's real-valued label. We use an alternating-projections constrained optimization approach to simultaneously learn a regression model and estimate all salience values. We evaluate this algorithm on a significant real-world problem, crop yield modeling, and demonstrate that it provides more extensive, intuitive, and stable salience models than Primary-Instance Regression, which selects a single relevant item from each bag.

1. Introduction

Classical machine learning operates on individual items, each represented by a feature vector and assigned a label, which is either categorical (for classification) or real-valued (for regression). However, some learning problems do not fit this model. There are situations in which observations are instead *bags* of items, with a single label applied to the bag. These tasks require a *multiple-instance learning* (MIL) approach. For example, the problem that first motivated this area of research was to predict a drug's activity ("active" or "inactive") given observations of multiple structural conformations of the drug molecule (Dietterich et al., 1997). Only some of the observed conformations contributed to the label of the molecule, and it was not known which ones were relevant. While much work in MIL has focused on obtaining high-accuracy classifications from such bag data, there has been relatively less focus on the task of determining which items are the relevant ones. Two efforts in this direction are Primary-Instance Regression (Ray & Page, 2001), which selects a single relevant item from each bag as its exemplar (primary instance), and the MILES algorithm (Chen et al., 2006), which selects a subset of items from each bag as relevant to the bag label.

In this paper, we present a novel method for inferring the *salience* of each item with respect to a real-valued bag label (i.e., a regression target). In this problem setting, the goal is not to provide predictions over new (unlabeled) data. Instead, we seek to better understand the contents of (labeled) bags of data. The regression target provides the necessary leverage to determine which items within a bag are most relevant to that target. We also introduce the method of *alternating projections* (AP): a feasibility algorithm that is well-known within the optimization community but which is less familiar to machine learning audiences.

Further, we have identified an important agricultural problem that is naturally cast as a multiple-instance learning problem: *crop yield modeling from remote sensing data.* The identification of pixels that are highly salient for corn, wheat, etc., yields can provide an automated survey of where those crops are being grown. These estimates, in turn, can inform precision agriculture efforts and USDA (United States Department of Agriculture) policies. This is an MIL problem in that we have thousands of individual pixel-level multispectral observations for each county, but only one target yield value per county (per crop) and no guidance as to which pixels are relevant to the target. We have found that the salience estimates produced by our algorithm reflect important spatial structures of crop distributions (Section 4).

Like prior work (Ray & Page, 2001), we find that our formulation requires solution of an NP-hard optimization problem (Section 3.2). Thus, we cannot expect to achieve an exact optimum. However, our

Appearing in the *ICML '2007 Workshop on Constrained Optimization and Structured Output Spaces*, Corvallis, OR, 2007.

approach produces a more informative model by returning salience estimates for each data item. The extra degrees of freedom require a more expensive optimization, so our work can be seen as expending computational effort for increased descriptive power.

2. Related Work

A significant amount of work has been devoted to methods for multiple-instance classification, including axis-parallel rectangles (Dietterich et al., 1997), diverse density (Maron & Lozano-Pérez, 1998), voting by the k nearest neighbor bags (Wang & Zucker, 2000), graph spectral methods (Rahmani & Goldman, 2006), and others. Amar et al. (2001) extended the k nearest neighbor and diverse density approaches to MIL problems in which each bag has a real-valued label that indicates its proximity to the target concept. Likewise, Goldman and Scott (2003) interpreted real-valued labels to be "the degree to which the example satisfies the target concept" and used axis-aligned rectangles to learn the target concept.

As noted above, methods for determining the relevance of each item have been limited. Ray and Page (2001) offered the pioneering work in this area by contributing a Primary-Instance Regression (PIR) method. This approach assumes that the label of a bag is determined by exactly one *primary instance* and that the rest of the items in the bag are noisy observations of the primary instance. They proposed an EM-based solution to alternately estimate the most likely primary instance for each bag and then to maximize the fit of a linear regression through the primary instances. Chen et al. (2006) proposed a method for multiple-instance classification that represents each bag by its similarity to each item in the data set, then uses an SVM to select relevant features (those with an SVM weight magnitude > 0). Since each feature implicitly stands for an item, a subset of relevant items is also identified.

3. Salience Assessment for MIR Data

In this work, we generalize these instance-selection methods: rather than making a binary decision about each item's relevance, we assign a continuous *salience* value to each one. Following Ray and Page, we aggregate bags to single exemplars, but our exemplars are weighted averages of the contents of the bag. This approach offers three advantages: first, it provides additional degrees of freedom in locating a high-quality regression fit. Second, it subsumes a number of common aggregators (e.g., `mean` or `max`), while not requiring a domain expert to specify the aggregator a priori. Third, an item's weight can be interpreted as its salience: its relevance, with respect to all other points in the bag, for predicting the bag label. As we show in Section 4, our algorithm successfully computes different saliences for the same bags given different labels. Thus, it is learning information about the structure of the data *in the context of the target of prediction*, not deriving salience from the bag contents alone.

3.1. The AP-Salience Algorithm

In this section, we describe the alternating projections algorithm (AP-Salience) algorithm that we use to optimize item salience. AP algorithms (Bauschke & Borwein, 1996) are closely related to EM algorithms, though they are motivated by geometric considerations rather than statistical ones. We give additional background on AP and an analysis of the performance of AP-Salience in Section 3.2. The core of our algorithm is an iterative re-estimation process: first, we compute the best assignment of salience values to items under a fixed regression model. Then, given the fixed saliences, we update the regressor that maps bag exemplars to bag labels.

Let $\mathbf{B}^i$ denote bag i from a data set D, which is a collection of m bags. Bag i consists of n^i data points, $B^i_j \in \mathbb{R}^d$. (Note that, although bags may contain different numbers of points, every data point must be of the same dimension.) Thus, each bag can be represented as a $d \times n^i$ real matrix. In the multiple-instance classification framework, bags may be positive, $\mathbf{B}^{i+}$, or negative, $\mathbf{B}^{i-}$. For regression (as in this paper), each $\mathbf{B}^i$ instead has an associated label, $y^i \in \mathbb{R}$.

Each bag can be thought of as a cloud of points that share the same y coordinate. That is, a bag is a bounded region of a hyperplane orthogonal to the y axis. We seek a regression surface (e.g., hyperplane) that passes through each bag and predicts the bag's true label from at least one point (the exemplar H) within that bounded region. Each item in bag i, B^i_j, is assigned an (unobserved) real value, α^i_j, that indicates how salient item B^i_j is to predicting y^i for bag $\mathbf{B}^i$. The vector $\alpha^i \in \mathbb{R}^{n^i}$ gives the saliences for all items in the bag.

The exemplar itself is a convex combination of the items in its bag: $H^i = \sum_j \alpha^i_j B^i_j$, where $\sum_j \alpha^i_j = 1$ and all $\alpha^i_j \geq 0$. This forces H^i to fall within the convex hull of the points in $\mathbf{B}^i$. Essentially, this constraint guarantees that bag $\mathbf{B}^i$ will influence the regressor. Without some constraint on α^i, H^i might fall arbitrarily far from the bag data, and $\mathbf{B}^i$ would not influence the final model.

Algorithm 1 AP-Salience

1: **Inputs:** $\{\mathbf{B}^k\}_{k=1\cdots m}; Y; \epsilon_1; \epsilon_2$ // Bag data; vector of bag labels; regularization coefficients for α and W
2: **Outputs:** $W; \{\alpha^k\}_{k=1\cdots m}$ // Regression coefficients; per-bag salience vectors
3: $W = \texttt{random}(d+1, 1)$ // Initialization: random $d+1$ column vector
4: **repeat**
5: **for** $k = 1$ to m **do** // Project from W space to α^k simplex
6: $\alpha^k \leftarrow$ solution of the QP: $\begin{cases} \min_{\alpha^k}: & {\alpha^k}^\mathsf{T}\left({\mathbf{B}^k}^\mathsf{T} W W^\mathsf{T}\mathbf{B}^k + \epsilon_1\mathbf{I}\right)\alpha^k - y^k W^\mathsf{T}\mathbf{B}^k\alpha^k \\ \text{subject to:} & {\alpha^k}^\mathsf{T}\mathbf{1} = 1; \quad \alpha_i^k \geq 0 \quad \forall i \end{cases}$
7: $H^k \leftarrow \mathbf{B}^k\alpha^k$
8: **end for**
9: $W \leftarrow (\mathbf{H}\mathbf{H}^\mathsf{T} + \epsilon_2\mathbf{I})^{-1}\mathbf{H}Y$ // Project from $\{\alpha^k\}$ onto W space
10: **until** convergence

For this paper, we assume a linear regression: $\widehat{y}(H^i) = W^\mathsf{T}H^i$, where W is the $d+1$ vector of regression weights and H^i is a $d+1$ column vector in homogeneous coordinates (prepending a 1 to each data vector). We take the usual L_2 loss, with regularization terms ϵ_1 and ϵ_2 on α^k and W (which are otherwise underdetermined). Altogether, we have the following optimization problem:

$$\begin{aligned} \text{min: } & f(W, \{\alpha^k\}_{k=1\cdots m}) \qquad (1) \\ &= \sum_{k=1}^{m}\left[\left(y^k - W^\mathsf{T}H^k\right)^2 + \epsilon_1\|\alpha^k\|^2\right] + \epsilon_2\|W\|^2 \\ &= \sum_{k=1}^{m}\left[\left(y^k - W^\mathsf{T}\mathbf{B}^k\alpha^k\right)^2 + \epsilon_1\|\alpha^k\|^2\right] + \epsilon_2\|W\|^2 \end{aligned}$$

$$\text{subj to: } \alpha_i^k \geq 0 \quad \forall i,k; \quad \sum\nolimits_{i=1}^{n^k}\alpha_i^k = 1 \quad \forall k,$$

where the minimization is with respect to both W and $\{\alpha^k\}_{k=1\cdots m}$. The core part of the objective, $f()$, is the squared-error loss term: $(y^k - W^\mathsf{T}\mathbf{B}^k\alpha^k)^2$. The factor $\mathbf{B}^k\alpha^k$ represents the aggregation of the data in bag k to a single exemplar (H^k), and the factor $W^\mathsf{T}H^k$ is the linear regression estimate of exemplar H^k. Analytic minimization of Equation 1 over both W and $\{\alpha^k\}$ simultaneously is difficult. But if either W or $\{\alpha^k\}$ is known, then the other can be found via well understood least-squares techniques. This motivates the AP approach of Algorithm 1.

In the AP-Salience algorithm, m represents the number of bags; $Y \in \mathbb{R}^m$ is the vector of all bag labels; $\mathbf{H}$ denotes the $(d+1) \times m$ matrix of all exemplar points; $\mathbf{I}$ and $\mathbf{1}$ denote the identity matrix and vector of all ones, respectively, each of the appropriate size; and ϵ_1 and ϵ_2 are regularization coefficients. In line 3, we initialize the weight vector to a random vector. The core of the algorithm is an alternation between two projection steps. In the first step (lines 5–8), we solve for each α^k assuming a fixed W. This step can be seen as a projection from the space of all possible W vectors ($\mathbb{R}^{d+1}$) onto the n^k-simplex. The simplex constraint necessitates a quadratic program to solve the least-squares problem. Next, we fix all of the α^k vectors and project back onto the W space (line 9). We alternate between these two steps until convergence.

3.2. Alternating Projections Methods

In this section, we show that the AP-Salience algorithm converges to a critical point of the objective function (1). We also find that exact optimization of this objective is NP-hard, so perfect minimization is impractical for large data sets.

Alternating projections is a powerful class of methods for the *convex feasibility problem*: finding the intersection of convex sets (Bauschke & Borwein, 1996). AP-based methods are widely used in the optimization community, where they appear in the solution of convex optimization problems, but they are less well known in the machine learning community. They can be used for function minimization by choosing the convex sets to be the zeros of the partial derivatives of the objective function. Then an intersection point is a simultaneous zero of all partial derivatives and is, thus, a critical point of the function. Algorithm 1 uses this formulation to find a critical point of (1).

There is an immense body of literature on AP algorithms that dates back to von Neumann (which we will not attempt to review here), but the essential idea is to start with a group of convex sets, $C_0, \ldots, C_n$, in a Hilbert space, X, and a corresponding group of projection operations $T_0, \ldots, T_n$. The goal is to find a point in the intersection of the C_i. To do so, start with an arbitrary point, $x \in X$, and iteratively apply the projection operations in sequence. T_i updates the representation of x in C_i, moving it closer (in the metric of X) to the corresponding representations in $C_{j\neq i}$. Because the C_i are convex, the sequence of T_i form a contraction mapping and the sequence of xs converges in norm to a fixed point: the desired point

in the intersection.

In our case, the Hilbert space is $X = \mathbb{R}^{d+1} \times \prod_{i=1}^{m} \mathcal{S}^i$, the Cartesian product of the $\mathbb{R}^{d+1}$ space of W with the α^i simplexes. A point $x \in X$ is an assignment of W and all α^i. The C_i are defined to be:

$$C_0 \equiv \{W : \tfrac{\partial}{\partial W} f(W, \alpha^i) = 0\} \times \alpha_{\text{fixed}}$$
$$C_i \equiv \{\alpha^i : \tfrac{\partial}{\partial \alpha^i} f(W, \alpha^i) = 0\} \times W_{\text{fixed}} \quad ,$$

the subsets of $\mathbb{R}^{d+1}$ and $\mathcal{S}^i$ corresponding to zeros of the derivative of the objective function, $f()$, with respect to each free variable.

These sets are convex because they are solutions to linear equations (the derivatives of the quadratic $f()$). The projections are lines 6 and 9 of Algorithm 1, which return points in the C_i sets. The Hessian matrix in line 6 is positive definite by construction, so the QP possesses a global optimum, which is a zero of $\frac{\partial}{\partial \alpha^k} f$ under fixed W and $\alpha^{i \neq k}$. Similarly, line 9 is the standard least-squares solution that minimizes $f()$ under fixed $\{\alpha^i\}$. Thus, the AP iteration is guaranteed to find *some* point in the intersection of the C_i, i.e., a critical point of $f()$ with respect to all variables simultaneously.

Unfortunately, the complete Hessian of $f()$ (with respect to all variables) is indefinite, so the critical point cannot be classified *a priori* (derivation omitted for space). As a corollary, optimization of Equation 1 is NP-hard (Pardalos & Vavasis, 1991), so we cannot expect to find the global optimum.

AP-Salience is more expensive than Ray and Page's PIR, but it remains polynomial in n^k and d on each iteration. The solution to the QPs in line 6 can be found in time polynomial in n^k, and the linear regression in in line 9 requires time $O(d^3)$ for the matrix inversion. Unfortunately, like EM algorithms, AP algorithms do not provide general guarantees on the rate of convergence. However, we experienced reasonable runtimes and convergence rates in practice, even using a basic QP implementation in Matlab.

4. Experimental Results

There are no existing non-artificial data sets for multiple-instance regression problems. In this section, we present a novel MIR problem: crop yield modeling. We find experimentally that AP-Salience assigns meaningful values to pixels across each county.

4.1. Data Sets: Modeling Crop Yield

Each year, the USDA reports the average yield per acre, for each county in the U.S., for a variety of crops. We seek to relate these yields to remote sensing observations from the Multi-angle Imaging SpectroRadiometer (MISR) instrument (Diner et al., 1998). The spatial resolution of each pixel is 250m, so each county is represented by thousands of pixels, but there is only a single yield value (per crop). Some pixels cover fields that were planted with the target crop, while others contain different crops or non-agricultural areas such as cities, lakes, forests, or desert regions. This problem lends itself well to the multiple-instance setting and provides an important, challenging real-world application for this work.

We have collected data sets[1] that cover two states, California and Kansas, over four years (2001–2004). The number of counties that reported yield values for corn and wheat for all four years are as follows:

California		Kansas	
Corn	Wheat	Corn	Wheat
15	20	103	105

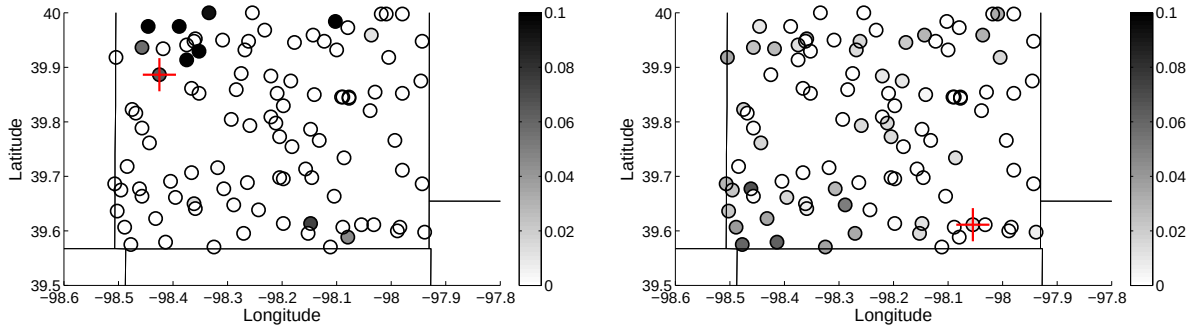
Each county corresponds to a bag, and the remote sensing (pixel) observations in that county are items inside the bag. We subsampled the remote sensing observations for each county to obtain 100 pixels for each bag. Each pixel is a vector comprising a sequence of observations taken every eight days for a year (46 time points per year). At each time point, reflectance values at red and near-infrared wavelengths were recorded, yielding two features per observation. Therefore, each pixel can have up to 92 features (two observations at each time point; 46 times across the year), depending on how much of the time series is used.

4.2. Experimental Methodology

We compared AP-Salience (Algorithm 1) to PIR in terms of their ability to identify salient pixels in the crop yield data (recall that the goal is not to make predictions for new bags). We did not empirically compare these methods to the MILES algorithm, since it was designed for use with MIL classification problems, and we have not yet extended it to use a regression SVM in place of the 1-norm classification SVM it currently employs. This is an area for future work.

As recommended by Ray and Page, we ran PIR ten times with different random initializations for the initial regression parameters. AP-Salience was run only once, with a single, randomly selected initial W vector. We have analyzed these data sets using anywhere from two to 92 features (time series) for each year of observation.

[1] Available at `http://harvist.jpl.nasa.gov/`.



(a) Corn (b) Wheat

Figure 1. Salience values (α) obtained by AP-Salience for 100 pixels in Jewell County, KS. Darker pixels are more salient. The pixel selected by PIR is marked by a cross. Both methods used observations covering January 1–February 18 of 2001.

4.3. Experimental Results

Finding 1: Salience depends on the bag targets. Figure 1 shows the salience values (α) obtained by AP-Salience for a single county in Kansas. When using corn yield as the desired target value, we find that the most salient pixels are located in the northwestern corner of the county (Figure 1a). However, when using wheat yield as the target, pixels in the southwestern part of the county are far more salient (Figure 1b). The information contained in the target label strongly influences the assignment of salience to items, as desired.

The pixel selected by PIR as the primary instance is indicated with a cross. The primary instance for corn yield received a high salience value from AP-Salience, indicating agreement on its relevance. In contrast, the PIR pixel selected for wheat yield had a much lower salience. In both cases, since PIR only identifies a single pixel as relevant, nothing is known about the remaining pixels.

Finding 2: High-salience pixels cluster spatially. Figure 1 also illustrates that the salience values assigned by AP-Salience exhibit a strong spatial correlation. AP-Salience is not given any information about the physical location of each pixel, yet highly salient values tend to cluster together spatially. That is, we can infer that corn fields tend to occur in the northwest and wheat fields in the southwest of this county. We found similar spatial distribution results in a number of other Kansas and California counties.

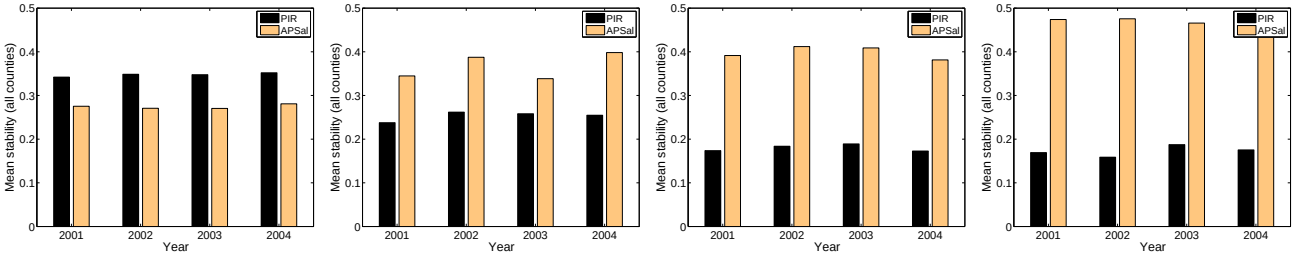
Finding 3: AP-Salience tends to produce more stable results than PIR. We expect approximately the same pixels to remain salient throughout the growing season. In our evaluation, PIR and AP-Salience are run individually for each time point; i.e., a separate regression model is learned for time points $[1, t]$ for $t = [1, 46]$ (using the bags from all four years). To assess the *stability* of both methods, we examined the distribution of max-salience instances across all 46 time points. We define stability for bag i as the entropy of this distribution normalized by the minimum possible entropy (if the same pixel were selected at each time point):

$$S_i = \frac{-\sum_{j=1}^{n_i} p_j \log(p_j)}{-T \log(T)},$$

where p_j is the number of times pixel j was selected as the primary instance (PIR) or max-salience instance (AP-Salience) and T is the number of time points. Figure 2 shows the mean stability (across all counties) obtained for both states and both crops. With the exception of CA/corn, AP-Salience tends to have higher stability than PIR. The results are consistent across years, for a given state and crop. There is, in fact, an inverse trend; PIR is more stable when analyzing the CA data, while AP-Salience is more stable on the KS data (and consistently more stable when modeling wheat versus corn yield). When more bags are available ($\sim$400 for KS versus $\sim$70 for CA), AP-Salience is able to produce a more accurate model, since it can distribute weight across all of the pixels in each bag. PIR tends to find it more difficult to produce a good model with more bags, since it must fit a linear regression through just one item from each bag.

5. Conclusions and Future Work

In this paper, we have presented a generalized solution to the problem of assessing the salience of items within a bag. In contrast to Primary-Instance Regression (PIR), which selects a single item from each bag to be its exemplar, or MILES, which selects a subset of items as relevant, we permit each item in the bag to contribute a fractional amount to an over-



(a) CA corn (60 bags) (b) CA wheat (80 bags) (c) KS corn (412 bags) (d) KS wheat (420 bags)

Figure 2. Mean (across all counties) stability (across a single year) of the primary instance selected by PIR versus the highest-salience pixel chosen by AP-Salience.

all weighted-sum exemplar. We have demonstrated that our algorithm, AP-Salience, assigns salience values that are sensitive to the target under study. We have also shown that AP-Salience assigns physically plausible (spatially correlated) values to pixels used for crop yield predictions and that it provides max-salient assignments that are generally more stable than those provided by PIR.

We employ linear regression here only as a first step and because of the simplicity of the formulation. To our knowledge, no nonlinear MIR methods exist. We plan to extend our approach to a nonlinear regressor by, for example, employing a nonlinear projection of B_j^i or by kernelizing the entire formulation. We also plan to extend MILES to regression problems to enable a direct comparison with AP-Salience. Finally, we are currently developing MIR methods that can provide predictions for new, unlabeled bags.

Acknowledgments

Dr. Wagstaff's work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration, supported by a grant from NASA's Applied Information Systems Technology Program. Dr. Lane's work was supported, in part, by NIMH grant number 1R01MH076282-01 as part of the NSF/NIH Collaborative Research in Computational Neuroscience Program. Thanks to Sushmita Roy, Robert Granat, and the reviewers for helpful comments on the paper. We also thank the MISR team, without whom this work would not have been possible.

References

Amar, R. A., Dooly, D. R., Goldman, S. A., & Zhang, Q. (2001). Multiple-instance learning of real-valued data. *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 3–10).

Bauschke, H. H., & Borwein, J. M. (1996). On projection algorithms for solving convex feasibility problems. *SIAM Review*, *38*, 367–426.

Chen, Y., Bi, J., & Wang, J. Z. (2006). MILES: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*.

Dietterich, T. G., Lathrop, R. H., & Lozano-Pérez (1997). Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence*, *89*, 31–71.

Diner, D. J., Beckert, J. C., Reilly, T. H., Bruegge, C. J., Conel, J. E., Kahn, R., Martonchik, J. V., Ackerman, T. P., Gordon, H. R., Muller, J.-P., Myneni, R., Sellers, R. J., Pinty, B., & Verstraete, M. M. (1998). Multiangle imaging spectroradiometer (MISR) instrument description and experiment overview. *IEEE Transactions on Geoscience and Remote Sensing*, *36*, 1072–1087.

Goldman, S. A., & Scott, S. D. (2003). Multiple-instance learning of real-valued geometric patterns. *Annals of Mathematics and Artificial Intelligence*, *39*, 259–290.

Maron, O., & Lozano-Pérez, T. (1998). A framework for multiple-instance learning. In M. I. Jordan, M. J. Kearns and S. A. Solla (Eds.), *Advances in Neural Information Processing Systems 10*, 570–576. MIT Press.

Pardalos, P. M., & Vavasis, S. A. (1991). Quadratic programming with one negative eigenvalue is NP-hard. *Journal of Global Optimization*, *1*, 15–22.

Rahmani, R., & Goldman, S. A. (2006). MISSL: Multiple-instance semi-supervised learning. *Proceedings of the Twenty-Third International Conference on Machine Learning*. Pittsburgh, PA.

Ray, S., & Page, D. (2001). Multiple instance regression. *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 425–432).

Wang, J., & Zucker, J. D. (2000). Solving the multiple-instsance learning problem: A lazy learning approach. *Proceedings of the Seventeenth International Conference on Machine Learning* (pp. 1119–1125).