Active Learning for Probabilistic Structured Prediction of Cuts and Matchings

Sima Behpour [1 2] **Anqi Liu** [3] **Brian D. Ziebart** [2]

Abstract

Active learning methods, like uncertainty sampling, combined with probabilistic prediction techniques have achieved success in various problems like image classification and text classification. For more complex multivariate prediction tasks, the relationships between labels play an important role in designing structured classifiers with better performance. However, computational time complexity limits prevalent probabilistic methods from effectively supporting active learning. Specifically, while non-probabilistic methods based on structured support vector machines can be tractably applied to predicting cuts and bipartite matchings, conditional random fields are intractable for these structures. We propose an adversarial approach for active learning with structured prediction domains that is tractable for cuts and matching. We evaluate this approach algorithmically in two important structured prediction problems: multi-label classification and object tracking in videos. We demonstrate better accuracy and computational efficiency for our proposed method.

1. Introduction

In many real-world applications, obtaining labeled instances for training is expensive. This is particularly true for multivariate prediction tasks, in which many labels are required for each training example. For example, an image can require many tags (e.g., mountain, sky, tree) as part of a multi-label prediction task, and video tracking has many pairs of bounding boxes between consecutive frames (Figure 1). Exhaustively annotating datasets for these tasks is extremely burdensome. Active learning (Settles, 2008; 2012) seeks to reduce this annotation burden by requesting the most useful annotations for learning. In this paper, we specifically consider the multivariate active learning setting in which each single variable's value can be separately solicited.

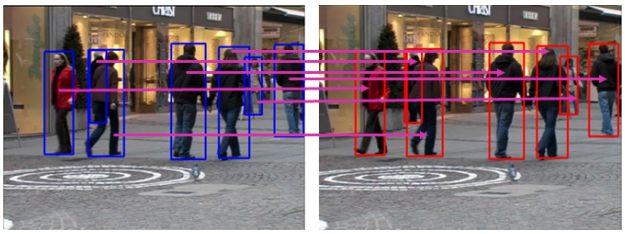


Figure 1. An example of a bipartite matching in a video tracking application. When some assignments are uncertain, choosing the most informative one can significantly reduce any uncertainty in the remaining assignments.

Uncertainty sampling (Lewis & Gale, 1994; Settles, 2012), the most popular active learning strategy, solicits annotations of variables for which the predictor is most uncertain. It can be naively applied to each variable independently in a multivariate learning task. However, leveraging inherent label correlations is critical in multivariate domains for better predictive performance. For instance, certain labels may co-occur in many training samples, while other labels may be mutually exclusive in multi-label learning (Ye et al., 2015). Unfortunately, many important types of these correlations reside on the boundaries of computational efficiency where prevalent probabilistic structured prediction methods, such as conditional random fields (Lafferty et al., 2001), are intractable, while margin-based (non-probabilistic) methods, such as structured support vector machines (Tsochantaridis et al., 2005), are tractable. Margin-based classifiers have been used for active learning in univariate settings by interpreting distance to margin as uncertainty (Tong & Koller, 2001) or using Platt scaling (Platt, 1999). However, extending these interpretations of margin for single variable uncertainty to multivariate settings is not well-defined, making margin-based structured prediction ill-suited for active learning with single variable label solicitation.

We propose a novel approach to multivariate active learning by leveraging adversarial structured prediction methods with rich structural constraints (Behpour et al., 2018; Fathony et al., 2018) to construct the worst-case probabilistic distri-

[1]University of Pennsylvania [2]Department of Computer Science, University of Illinois at Chicago [3]California Institute of Technology. Correspondence to: Sima Behpour <sbehpour@seas.upenn.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

butions of unknown variables. Using these distributions, we assess the multivariate value of information expected from different unlabeled variables to make solicitation choices. We illustrate this approach on two prediction applications: multi-label classification and object tracking in video. We formulate these prediction problems as learning minimum cuts and bipartite matchings—two settings for which efficient probabilistic structured prediction methods have only recently been developed. We demonstrate the benefits of our approach against existing active learning methods that can be efficiently applied to these learning settings.

The paper is organized as follows: Firstly, we introduce background and related work, which include previous efforts on univariate and multivariate active learning problems. We then cover the main methodology of adversarial structured prediction, Adversarial Robust Cut (ARC) and Adversarial Bipartite Matching (ABM) methods, followed by the active learning algorithm details for each method. We compare our proposed method with state-of-the-art approaches on real-world data sets in experiments before we conclude.

2. Background and Related Works

2.1. Univariate Active Learning

Active learning approaches have primarily considered single variable prediction problems. Pool-based active learning (Lewis & Gale, 1994) selects examples from an unlabeled dataset $D_u = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, to form a labeled dataset D_l for training. Here $x \in \mathcal{X}$ is an input/feature variable and $y \in \mathcal{Y}$ is a label variable that we seek to predict. $\mathcal{U}$ denotes the index set for the unlabeled data. A sequence of examples are chosen from D_u based only on input values, x, and moved to the labeled data D_l once the label, y, is revealed. The goal of the active learner is to choose a dataset so that the resulting classifier $f : \mathcal{X} \to \mathcal{Y}$ learned from the available data has minimal error on additional testing data—either from all available data, the remaining unlabeled pool or a separate sample. *Query by committee* (Seung et al., 1992), uncertainty sampling (Lewis & Gale, 1994), and *active learning with the support vector machine* (Tong & Koller, 2001) are among the earliest approaches tackling active learning in binary classification problems. Uncertainty sampling (Lewis & Gale, 1994) works by selecting the unlabeled training instance from the pool of unlabeled data D_u with the label that the predictor, $\hat{P}$, is currently most uncertain about:

$$\underset{i \in U}{\operatorname{argmax}} H(Y_i | x_i, D_l). \tag{1}$$

using the conditional Shannon entropy,

$$H(Y_i | x_i, D_l) = -\sum_{y_i} \hat{P}(y_i | x_i, D_l) \log \hat{P}(y_i | x_i, D_l), \tag{2}$$

of the unknown label Y_i under the predictor's distribution, $\hat{P}$, trained from available labeled data.

Guo and Greiner (2007) employ an active learning strategy that selects the unlabeled instance whose label would provide maximum mutual information about the labels of the remaining unlabeled instances, given the labeled data:

$$\underset{i \in U}{\operatorname{argmax}} H(Y_U | X_u, D_l) - H(Y_{U \setminus i} | X_{U \setminus i}, D_l, (x_i, y_i)).$$

The first term does not depend on the selected instance i, therefore we can re-write it as:

$$\underset{i \in U}{\operatorname{argmin}} H(Y_{U \setminus i}, D_l, (x_i, y_i)). \tag{3}$$

Thus, maximizing the conditional mutual information is equivalent to minimizing the classification uncertainty (entropy) of the unlabeled data set. This approach has been successful in semi-supervised learning (Guo & Greiner, 2007; Grandvalet & Bengio, 2005). We employ this heuristic in the selection strategy of our proposed adversarial active learning structured prediction framework.

Active learning with support vector machines uses a non-probabilistic notion of uncertainty based on the distance of each example to the decision boundary. This is used directly to solicit labels from examples closest to the boundary (Tong & Koller, 2001) or by using Platt scaling (Platt, 1999) with uncertainty sampling. Unfortunately, the unreliability of Platt scaling (Lambrou et al., 2012) and the complications of interpreting decision boundary distances of support vector machines for multiclass and structured prediction tasks makes this approach difficult to generally apply.

2.2. Multivariate Active Learning

We consider multivariate active learning in this work. Instead of a univariate label, y, each example has a vector-valued label $\mathbf{y}$. We assume that the active learner can solicit single variables in this label vector, y_i, instead of soliciting the entire vector, $\mathbf{y}$, at each iteration of learning.

The simplest approaches to multivariate active learning reduce each example to a set of univariate examples and apply univariate active learning methods. Binary Relevance (BR) (Godbole & Sarawagi, 2004) decomposes a multi-label classification problem with N labels to N binary classification problems. The decomposition in BR discards the joint information between different labels. The joint information is modeled in many extensions of BR, including methods that model multi-label classification as structured prediction and consider the relevance information (Yang et al., 2009). However, they suffer from a lack of reliability since they employ SVM as the classifier and the final output is required to be transferred to a probabilistic format using methods like SVM Platts (Lambrou et al., 2012; Platt, 1999). *Binary*

version space minimization (BinMin; (Brinker, 2006)) solicits the closest instance to the decision boundary for one of the binary classifiers within BR. The sample selection strategies for BR rely on the main assumption that there is only one label for each instance. These methods are not applicable in active learning for structured prediction since label relevance information is not modeled. The selection procedure of uncertainty sampling based on BR only captures the uncertainty produced by binary classifiers on the labeled set and does not utilize the information relating labeled instances and the remaining unlabeled instances. Ignoring the label and instance relation in this manner limits the application of uncertainty approaches based on single output variables in structured prediction domains. This motivates uncertainty measures that capture marginal or joint information over unlabeled data.

Mutual Information (MI) is a dependency measure between two variables and defines the amount of information that is held in a random variable. The mutual information of two discrete random variables Y_a and Y_b is:

$$I(Y_a; Y_b) = \sum_{y_a \in A} \sum_{y_b \in B} P(y_a, y_b) \log \left(\frac{P(y_a, y_b)}{P(y_a)P(y_b)} \right),$$

where $P(y_a, y_b)$ is the joint probability function of y_a and y_b, and $P(y_a)$ and $P(y_b)$ are the marginal probability distribution functions of y_a and y_b respectively (Cover & Thomas, 2012). The MI can be equivalently expressed as:

$$I(Y_a; Y_b) = H(Y_a) + H(Y_b) - H(Y_a, Y_b), \quad (4)$$

where $H(Y_a)$ and $H(Y_b)$ are the marginal entropy, and $H(Y_a, Y_b)$ is the joint entropy of Y_a and Y_b. Inspired by the success of applying MI as the solicitation strategy in recent studies like (Khodabandeh et al., 2017; Sun et al., 2015), we leverage MI in the multivariate case as our active learning uncertainty measurement in this paper.

Another line of work focusing on active learning for classification on graphs starts with a partially labeled graph and utilizes a deterministic search strategy to query the next example (Dasarathy et al., 2015). We instead take a probabilistic approach using from adversarial structured prediction methods and leverage resulting uncertainty measures.

3. Adversarial Structured Prediction

3.1. Minimax Game Formulation

Rather than seeking a predictor that minimizes the (regularized) empirical risk,

$$\min_{\theta} \mathbb{E}\left[\text{loss}(Y, f_\theta(X))\right] + \lambda ||\theta||_2, \quad (5)$$

adversarial prediction methods (Topsøe, 1979; Grünwald & Dawid, 2004) instead introduce an adversarial approximation of the training data labels, $\check{P}(\check{y}|x)$, and seek a predictor, $\hat{P}(\hat{y}|x)$, that minimizes the expected loss against the worst-case distribution chosen by the adversary:

$$\min_{\hat{P}} \max_{\check{P}} \mathbb{E}_{x \sim \tilde{P}; \check{y}|x \sim \check{P}; \hat{y}|x \sim \hat{P}} \left[\text{loss}(\hat{Y}, \check{Y})\right] \quad (6)$$
$$\text{such that: } \mathbb{E}_{x \sim \tilde{P}; \check{y}|x \sim \check{P}} \left[\phi(X, \check{Y})\right] = \tilde{\mathbf{c}},$$

where the adversary is constrained by certain measured statistics (i.e., based on feature function ϕ) of the training sample $\tilde{\mathbf{c}}$—either with equality constraints, as shown, or inequality constraints. $\tilde{P}$ represents the empirical distribution of $\mathbf{X}$ [and $\mathbf{Y}$] in the training data set.

While the empirical risk cannot be tractably optimized for many natural loss functions of interest (e.g., the 0-1 loss or Hamming loss), adversarially minimizing the loss measure is often tractable (Asif et al., 2015). This adversarial minimization aligns the training objective to the loss measure better than surrogate losses (e.g., the hinge loss), providing better performance in practice for both classification (Asif et al., 2015) and structured prediction (Behpour et al., 2018; Fathony et al., 2018).

A key advantage of this adversarial approach for multivariate active learning is that the adversary chooses a joint probability distribution, which provides correlations between unknown label variables, $P(y_i, y_j)$, that are useful for estimating the value of information for different annotation solicitation decisions. The benefit of this uncertainty in structured predictions is most pronounced for settings in which other probabilistic methods—namely, conditional random fields (Lafferty et al., 2001)—are computationally intractable, while adversarial structured prediction methods can be efficiently employed. We focus on active learning for two such structured prediction tasks in this paper: learning to make cuts in graphs and learning to make bipartite matchings. In the remainder of this section, we review the adversarial structured prediction methods for these settings.

3.2. Adversarial Robust Cuts

The Adversarial Robust Cut (ARC) approach (Behpour et al., 2018) learns binary associative Markov networks with attractive pairwise relationships. It is formulated as an adversarial structured prediction problem with the Hamming loss as payoffs, inequality constraints over properties of variable pairs, and equality constraints over unary properties:

$$\min_{\hat{P}(\hat{\mathbf{y}}|\mathbf{x})} \max_{\check{P}(\check{\mathbf{y}}|\mathbf{x})} \mathbb{E}_{x \sim \tilde{P}; \check{\mathbf{y}}|x \sim \check{P}; \hat{\mathbf{y}}|x \sim \hat{P}} \left[\sum_i (\hat{Y}_i \neq \check{Y}_i)\right] \quad (7)$$
$$\text{such that: } \mathbb{E}_{\mathbf{x} \sim \tilde{P}; \check{\mathbf{y}}|\mathbf{x} \sim \check{P}} \left[\Phi_u(\check{Y}, \mathbf{X})\right] = \tilde{\mathbf{c}}_u \text{ and}$$
$$\mathbb{E}_{\mathbf{x} \sim \tilde{P}; \check{\mathbf{y}}|\mathbf{x} \sim \check{P}} \left[\Phi_p(\check{Y}, \mathbf{X})\right] \leq \tilde{\mathbf{c}}_p.$$

After introducing Lagrangian multipliers, θ_u and θ_p, to enforce the unary and pairwise constraints on the adversary,

the optimization can be re-written as:

$$\min_{\theta_u,\theta_p\leq 0} \mathbb{E}_{\tilde{P}(\mathbf{x})}\left[\left(\max_{\check{\mathbf{p}}_\mathbf{X}}\min_{\hat{\mathbf{p}}_\mathbf{X}} \hat{\mathbf{p}}_\mathbf{X}^\mathrm{T}\, \mathbf{C}_{\theta,\mathbf{x}}\, \check{\mathbf{p}}_\mathbf{X}\right) - \theta_u^\mathrm{T}\tilde{\mathbf{c}}_u - \theta_p^\mathrm{T}\tilde{\mathbf{c}}_p\right],$$

with the matrix defined as $C_{\theta,X}(\hat{y},\check{y}) = \sum_i(\hat{y}_i \neq \check{y}_i) + \psi(\check{y}, X)$, where $\psi(\mathbf{y},\mathbf{x}) = \sum_i \theta_u \cdot \phi(y_i,\mathbf{x}) + \sum_{i\neq j}\theta_p \cdot \phi(y_i,y_j,\mathbf{x})$. An example is shown in Table 1. Note that the inner saddlepoint corresponds to a zero-sum game between predictor and adversary over this game matrix $\mathbf{C}$.

Table 1. Augmented Hamming loss matrix for n=3 samples.

	$\check{y}$=000	$\check{y}$=001	$\check{y}$=010	$\check{y}$=011	$\check{y}$=100	$\check{y}$=101	$\check{y}$=110	$\check{y}$=111
$\hat{y}$=000	0+ ψ_{000}	1+ ψ_{001}	1+ ψ_{010}	2+ ψ_{011}	1+ ψ_{100}	2+ ψ_{101}	2+ ψ_{110}	3+ ψ_{111}
$\hat{y}$=001	1+ ψ_{000}	0+ ψ_{001}	2+ ψ_{010}	1+ ψ_{011}	2+ ψ_{100}	1+ ψ_{101}	3+ ψ_{110}	2+ ψ_{111}
$\hat{y}$=010	1+ ψ_{000}	2+ ψ_{001}	0+ ψ_{010}	1+ ψ_{011}	2+ ψ_{100}	3+ ψ_{101}	1+ ψ_{110}	2+ ψ_{111}
$\hat{y}$=011	2+ ψ_{000}	1+ ψ_{001}	1+ ψ_{010}	0+ ψ_{011}	3+ ψ_{100}	2+ ψ_{101}	2+ ψ_{110}	2+ ψ_{111}
$\hat{y}$=100	1+ ψ_{000}	2+ ψ_{001}	2+ ψ_{010}	3+ ψ_{011}	0+ ψ_{100}	1+ ψ_{101}	1+ ψ_{110}	2+ ψ_{111}
$\hat{y}$=101	2+ ψ_{000}	1+ ψ_{001}	3+ ψ_{010}	2+ ψ_{011}	1+ ψ_{100}	0+ ψ_{101}	2+ ψ_{110}	1+ ψ_{111}
$\hat{y}$=110	2+ ψ_{000}	3+ ψ_{001}	1+ ψ_{010}	2+ ψ_{011}	1+ ψ_{100}	2+ ψ_{101}	0+ ψ_{110}	1+ ψ_{111}
$\hat{y}$=111	3+ ψ_{000}	2+ ψ_{001}	2+ ψ_{010}	1+ ψ_{011}	2+ ψ_{100}	1+ ψ_{101}	1+ ψ_{110}	0+ ψ_{111}

Though these zero-sum games (Table 1) can be solved using a linear program in time that is polynomial in the size of the game, the game size (played over the set of all possible label assignments as actions) grows exponentially with the number of predicted variables. The double oracle algorithm (McMahan et al., 2003) is a constraint generation method that uncovers a (sparse) set of strategies for each player that supports the equilibrium. It operates by repeatedly finding the game's equilibrium for the current set of strategies and then alternatively adds each player's best response against the other player's equilibrium distribution as a new row or column for the game. For ARC, the adversary's best response against the predictor's distribution, $\hat{P}(\hat{y}|x)$, is:

$$\operatorname*{argmax}_{\check{y}} \mathbb{E}_{\hat{P}(\hat{\mathbf{y}}|\mathbf{x})}\left[\sum_i(\hat{Y}_i \neq \check{y}_i)\right] + \sum_i \psi_u(\check{y}_i,\mathbf{x}) + \sum_{i\neq j}\psi_p(\check{y}_i,\check{y}_j). \quad (8)$$

This reduces to finding the minimum s-t cut for a graph (Figure 2) with edges between predicted variables Y_i and Y_j weighted by pairwise potentials, and edges from a source node and to a sink node weighted by the expected loss against the predictor's distribution and the unary potentials. The predictor's best response is obtained independently for each variable based on the expected loss and the marginal

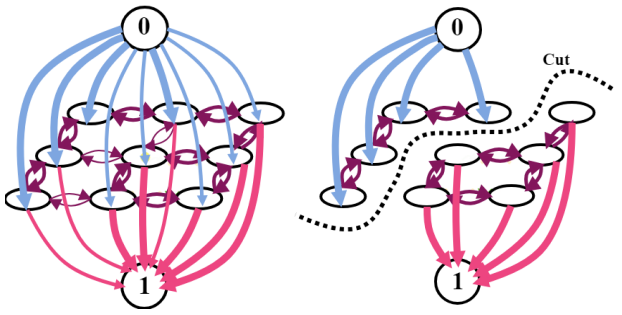


Figure 2. A directed graph to augment a Markov network (left) so that the minimum cut (right) provides the most probable assignment of each variable based on its connection to the source node (0) or target node (1).

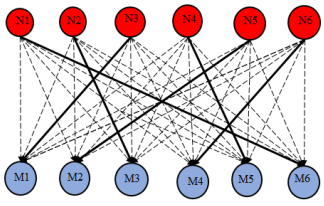
probability of the adversary's distribution:

$$\operatorname*{argmin}_{\hat{\mathbf{y}}} \mathbb{E}_{\check{P}(\check{\mathbf{y}}|\mathbf{x})}\left[\sum_i\left(\hat{y}_i \neq \check{Y}_i\right)\right] = \operatorname*{argmax}_{\hat{y}_i} \check{P}(\hat{y}_i). \quad (9)$$

Lastly, the optimal Lagrangian multipliers values are obtained using AdaGrad (Duchi et al., 2011) with gradients computed as differences between the expected features under the adversary's distribution and the empirical features calculated from the training data: $\mathbb{E}_{\mathbf{x}\sim\tilde{P};\check{\mathbf{y}}|\mathbf{x}\sim\check{P}}[\phi(\check{\mathbf{Y}},\mathbf{X})] - \mathbb{E}_{\mathbf{x},\mathbf{y}\tilde{P}}[\phi(\check{\mathbf{Y}},\mathbf{X})]$. The overall objective is convex, so convergence to a global optimum is guaranteed.

3.3. Adversarial Bipartite Matching

We consider bipartite graphs G = (V, E) (i.e., graphs with two disjoint vertex sets with no edges between nodes in the same set) and perfect bipartite matchings, B, a set of edges in which each vertex is incident to only one edge of B, as illustrated in Figure 3.



	M1	M2	M3	M4	M5	M6
N1	0	0	0	0	0	1
N2	0	0	1	0	0	0
N3	1	0	0	0	0	0
N4	0	0	0	0	1	0
N5	0	1	0	0	0	0
N6	0	0	0	1	0	0

Figure 3. A bipartite matching graph of size six, $|M| = |N| = 6$, and the adjacency matrix.

This matching can be denoted as an assignment (or permutation) π, where $\pi_i \in [n]$ indicates the item in the second set that item i from the first set is paired with.

Adversarial Bipartite Matching (ABM) (Fathony et al., 2018) seeks the predictive distribution over assignments with the lowest expected loss against a constrained adver-

sarial approximation of the training data assignments:

$$\min_{\hat{P}(\hat{\pi}|x)} \max_{\check{P}(\check{\pi}|x)} \mathbb{E}_{x\sim\tilde{P};\hat{\pi}|x\sim\hat{P};\check{\pi}|x\sim\check{P}} \left[\sum_{i=1}^{n} \hat{\pi}_i \neq \check{\pi}_i\right] \quad (10)$$
$$\text{such that } \mathbb{E}_{x\sim\tilde{P};\check{\pi}|x\sim\check{P}} \left[\sum_{i=1}^{n} \phi_i(x, \check{\pi}_i)\right] = \tilde{\mathbf{c}},$$

where $\hat{\pi}, \check{\pi}$ are the node assignments chosen by the predictor and the adversary, respectively. This approach also applies the double oracle method (McMahan et al., 2003) to generate active constraints (rows and columns of the game matrix) supporting the game's equilibrium. Best responses are obtained using the Hungarian matching algorithm, also known as the Kuhn-Munkres algorithm, to find the maximum-weight matchings in a bipartite graph in $O(|V|^3)$ time.

4. Adversarial Multivariate Active Learning

4.1. Sample Selection Strategy

A label selection strategy that provides the most useful information for learning is needed. The full impact of soliciting a label is the combination of what it reveals about other variables in the structured prediction and what influence updating the model parameters will have on all other variables. Since the latter is very difficult to calculate exactly or even estimate loosly, we focus on the former. The benefit of observing a variable can be measured using information theory. The total expected reduction in uncertainty over all variables, $Y_1, \ldots, Y_n$, from observing a particular variable Y_j given labeled dataset D_l (referred to as V_j) is:

$$V_j = \overbrace{\sum_{i=1}^{n} H(Y_i|D_l)}^{\text{uncertainty before observing } y_j} - \overbrace{\sum_{y_j\in\mathcal{Y}} P(y_j|D_l) \sum_{i=1}^{n} H(Y_i|D_l, y_j)}^{\text{expected uncertainty after observing } y_j}$$
$$= \sum_{i=1}^{n} I(Y_i; Y_j|D_l). \quad (11)$$

These mutual information values can be effectively computed from the adversary's equilibrium distribution using the pairwise marginal probabilities of two variables, $\check{P}(y_i, y_j)$, which was our main motivation for employing adversarial prediction methods for problems that are intractable for other probabilistic structured prediction approaches.

4.2. Active Learning Adversarial Robust Cuts

Our active learning approach for Adversarial Robust Cuts (ARC) solicits single variable values from the pool of unlabeled variables in two steps: it first computes the value of information $V_j(x)$ for each variable j of the sample x and chooses the most informative variable in the label vector from all examples, $\max_{x,i} V_i(x)$ to solicit. After partially labeling this sample, it is returned to both labeled and unlabeled pools. The purpose of returning the sample to the labeled pool is for updating gradients partially (just for labeled variables), and to the unlabeled pool is for providing an opportunity for other variables of the sample to be selected and labeled.

Partially-labeled examples pose challenges to the existing ARC learning method, which is based on fully annotated training examples. We resolve the problem in two ways: Firstly, during label solicitation, we fix the variables that have already been labeled to their true values before we infer an adversarial distribution over cuts that do not violate these assignments. Secondly, when updating model parameters, θ_u and θ_p, we only calculate gradients based on the subsets of variables that have been labeled, even though the adversarial prediction is obtained for all variables in the (partially labeled) example. Equivalently, when calculating the gradients, the expectation of features is considered to only be applied to the subset of variables that have been labeled. Here, we assume the parameters learned from labeled variables should generalize to the partially labeled graph.

4.3. Active Learning Adversarial Bipartite Matching

One of the advantages of our adversarial approach is that it tractably provides meaningful probabilistic predictions that can also be useful for active learning label solicitation. We employ pool-based active learning (Lewis & Gale, 1994) for the bipartite matching problem using Adversarial Bipartite Matching (ABM), as shown in Algorithm 1. It first selects a subset of unlabeled data from D_u and then chooses a subset of edges (N nodes assignments) to be queried. It is notable that the full assignment π may not be solicited and only a subset of N nodes assignment for an instance may be queried. The new labeled data are added to D_l and removed from D_u. The algorithm's goal is to learn the best classifier with the minimum number of labeled instances.

5. Experiments

5.1. Prediction Tasks, Datasets, and Features

Multi-label Classification: In our first application, we consider active learning for multi-label classification. In this setting, the learner chooses a single label variable from which to increase its dataset and update its predictions. We choose eight different datasets covering different domains like text, images, and biology from the Mulan dataset (Tsoumakas et al., 2011) for our experiments. The detailed information for every dataset is presented in Table 2. The total number of labels ranges from 14 (Yeast) to 374 (Corel5K) and the average cardinality (average number of active labels) ranges from 2.402 (Bibtex) to 26.044 (CAL500).

Algorithm 1 Active Learning Algorithm for Adversarial Bipartite Matching (ABM).

Require: Features $\{\phi_i(\cdot)\}$; Initial parameters θ; Initial labels π_{initial}
D_l: a small set of initially labeled examples;
D_u: the pool of unlabeled data for active selection;
Train an initial ABM model f on D_l
repeat
  Make predictions with classifier using parameters θ for all samples in D_u;
  Calculate $v_i(x)$ using Equation 11 for all nodes of samples $x \in D_u$;
  Select the most informative sample according to $X^* = \max_x \sum_i v(x)$;
  Query the sample's node assignments;
  Update the model f according to Eqs. 8-10;
  Move X^* from D_u to D_l;
until stop criterion reached.
return final classifier parameters θ.

Table 2. Multi-label datasets used in the experiments.

Dataset	Domain	# Instances	# Labels	Cardinality
Bibtex	text	7395	159	2.402
Bookmarks	text	87856	208	2.028
CAL500	images	502	174	26.044
Corel5K	images	5000	374	3.522
Enron	text	1702	53	3.378
NUS-WIDE	images	269648	81	2.320
tmc2007	text	28596	22	2.158
Yeast	biology	2417	14	4.237

We use the same feature representation as prior work (Behpour et al., 2018) by defining unary and pairwise features using features from Mulan (Tsoumakas et al., 2011) and word2vec[1] features.

Object Tracking: In our second application, we consider active learning for object tracking between video frames(Mozhdehi & Medeiros, 2017). The active learning selects two consecutive frames to add as labeled data to its training dataset to improve performance. We follow the same problem definition presented by Kim et al. (2012). In this setting, a set of images (video frames) and a list of objects in each image are given. The correspondence matching between objects in frame t and objects in frame $t+1$ is also provided. Figure 1 provides an example of this setup. The number of objects changes when a subset of the objects may enter, leave, or remain in the consecutive frames. To deal with this problem, we double the number of nodes in every frame. We consider the number of objects in frame t as N_t and N^* be the maximum number of objects a frame can have, i.e., $N^* = \max_{t \in T} N_t$. Starting from N^* nodes to present the objects, we consider N^* more nodes as "invisible" nodes to let new objects enter and existing objects to leave. Hence the total number of nodes in each frame doubles to $n = 2N^*$. We follow the joint feature representation in ABM and (Kim et al., 2012) to define the affinities and correlations between node pairs in two consecutive frames.

We evaluate using a video tracking problem from the MOT challenge dataset (Leal-Taixé et al., 2015). We consider the TUD datasets and the ETH datasets as two different groups of datasets in our experiment. Each dataset differs in the number of samples (i.e., pairs of two consecutive frames to be matched) and the number of nodes (i.e., the number of pedestrian bounding boxes in the frame plus the number of extra nodes to indicate entering or leaving). The detailed information of datasets is described in Table 3. To make the experiment more challenging and to avoid having examples that are too similar in the training set, we combine each pair of datasets that have similar characteristics. In particular, this results in eight mixed datasets that we evaluate.

5.2. Comparison Methods for Multi-label Classification

We consider two sets of multi-label classification experiments based on two solicitation strategies for each method. The first solicitation strategy employs Mutual Information (MI) and the second one is random sampling.

ML-KNN: For our first comparison method for multi-label classification, we consider ML-KNN which is one of the well-known methods in multi-label classification. This method is derived from the K-nearest neighbor (KNN) algorithm. In this method, the K nearest neighbors of each instance are first identified. Then, statistical information of the neighbors' instances are used in maximum a posteriori (MAP) estimation of the label instances (Zhang & Zhou, 2007). Since ML-KNN provides probabilistic output, we can use it in the mutual information (MI) solicitation strategy of active learning.

Robust Bias-aware Prediction: Our second comparison method similarly considers each predicted variable as a separate binary classification problem and applies uncertainty sampling using the robust bias-aware (RBA) learner (Liu et al., 2015). RBA is a probabilistic classifier that adapts to different sample selection biases using the log loss and statistics of the data moments as constraints to find a minimax optimal classifier among conditional label distributions under log loss prediction (Liu et al., 2015). The resulting distribution moderates its uncertainty by taking into account covariate shift between labeled and unlabeled datasets:

$$P(y|x) = e^{\frac{P_l(x)}{P_u(x)}\theta \cdot f(x,y)} / \sum_{y' \in y} e^{\frac{P_l(x)}{P_u(x)}\theta \cdot f(x,y')}.$$

[1] https://code.google.com/p/word2vec/

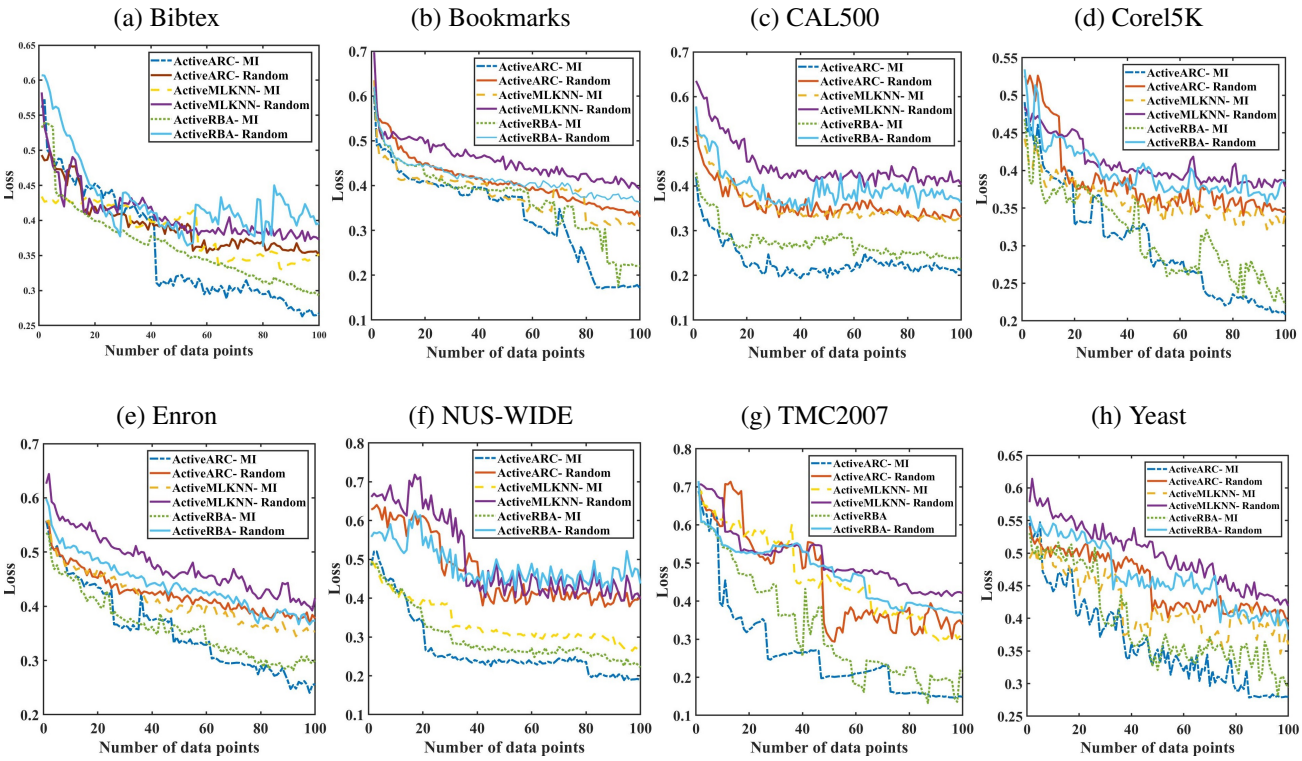


Figure 4. Hamming loss (Loss) values for the first 100 datapoints of active learning averaged over 30 randomized withheld evaluation dataset splits.

Omitted Baseline Comparisons: We evaluate two additional sets of baselines for comparison purposes. However, the performance for these is significantly worse than the reported baselines and, thus, we omit these from Figure 4.

The first set of additional baselines that we consider is a standard logistic regression (LR) model that treats each predicted variable as a separate binary classification problem. We follow the same solicitation strategies in these experiments: MI and random sampling. However, the performance of both experiments was very low in comparison with the other baseline methods.

As a second set of baselines, we employ a solicitation strategy that selects the least confident sample for each classifier. The performance is better or close to the random selection strategy but worse than the mutual information strategy.

5.3. Comparison Method for Object Tracking

Structured support vector machines with Platt scaling: For our bipartite matching setting, we consider active learning with structured support vector machines (SSVM) as a baseline to evaluate the performance of our approach. We implement the SSVM model (Taskar et al., 2005; Tsochantaridis et al., 2005) following (Kim et al., 2012) using SVM-Struct (Joachims, 1998). We apply Platt scaling (Platt, 1999) to transform our potential functions to probabilistic outputs under the SSVM. It works by fitting a Sigmoid function to the decision values for each class through optimizing parameters a and b of a Sigmoid function (called the scoring algorithm) parameters, in the following expression: $\frac{1}{1+exp(a \cdot z+b)}$, where z is an input potential value. We first follow a learning algorithm on a subset of data to learn and fit a and b in a multi-class setting. In solicitation step of active learning, we compute the probabilistic value of the bipartite graph edges by passing the potential value of the edge to the Sigmoid scoring function. The entropy of every node is calculated by summing over the entropy of its edges and the entropy of every training sample is computed by summing the nodes entropies. The sample and the node with highest entropy is chosen to be solicited. This matches our active learning ABM solicitation strategy, which is applied to query the most informative node assignment.

5.4. Experimental Results

We report our multi-label classification experiment results in Figure 4. Though the initial performance differences between methods are mixed across datasets, our proposed method, denoted ActiveARC, provides less loss in comparison with other methods (ActiveRBA and ActiveMLKNN) over all datasets after roughly 40 data points are solicited

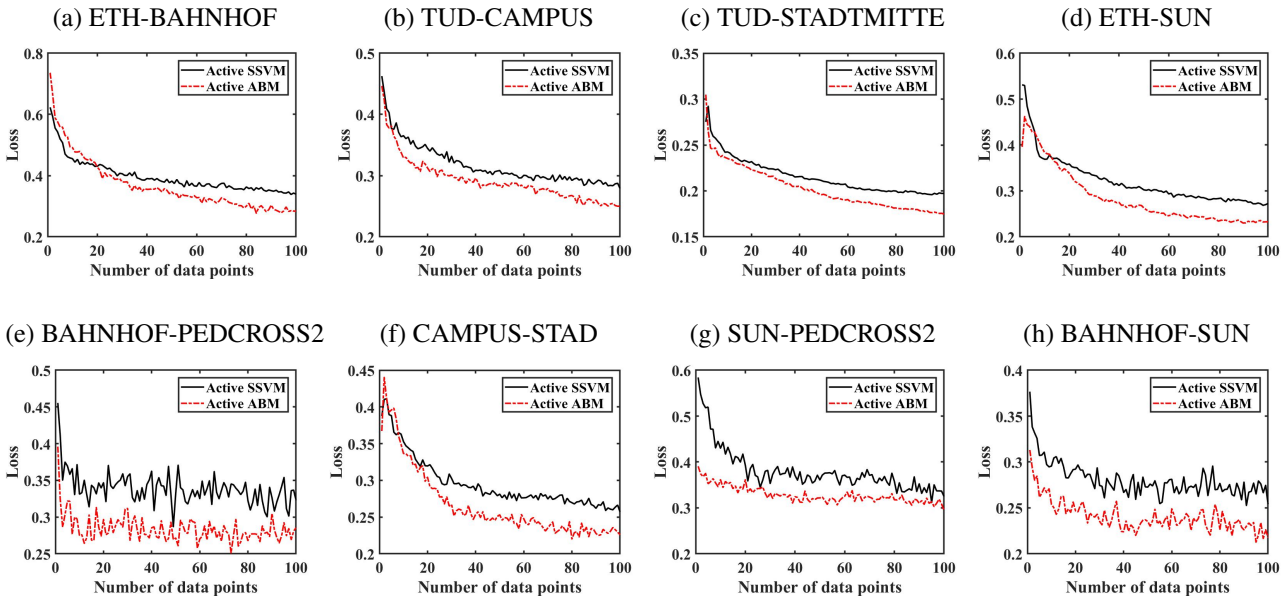


Figure 5. Hamming loss (Loss) values for the first 100 datapoints of active learning averaged over 30 randomized withheld evaluation dataset splits.

Table 3. Dataset properties for tracking experiments.

DATASET	# ELEMENTS	# EXAMPLES
TUD-CAMPUS	12	70
TUD-STADTMITTE	16	178
ETH-SUNNYDAY	18	353
ETH-BAHNHOF	34	999
ETH-PEDCROSS2	30	836

by active learning. We attribute this advantage to ARC's improved ability to incorporate correlations between variables into both its prediction and its label solicitation strategy.

We report the results of our object tracking experiments in Figure 5. Apart from somewhat mixed performance in the early iterations of active learning, our proposed active learning framework (Active ABM) provides better performance compare to Active SSVM. We attribute this advantage to the better uncertainty model that our approach provides compared with the Platt scaling approach used by SSVM.

5.5. Inference Running Time

The key difference for inference under our approach (and advantage when learning) is that it uses multiple mincuts/permutations to construct an equilibrium. The (average) numbers of mincuts/permutations (denoted as Perms in the table) to arrive at an equilibrium for different datasets are presented in Table 4 for both experiments. We can conclude from Table 4 that inference from scratch is roughly 6-20 times slower than SSVM and other methods that use a single mincut or permutation. During training, however, the strategies from the previous equilibria can be cached and reused, making training time comparable to other methods.

Table 4. Inference running times for tracking experiments.

ACTIVE ABM		**ACTIVE ARC**	
DATASET	# PERMS	DATASET	# MINCUTS
ETH-BAHN	11.3	BIBTEX	20.4
TUD-CAMP	8.7	BOOKMARKS	14.5
TUD-STAD	9.4	CAL500	12.4
ETH-SUN	10.3	COREL5K	20.2
BAHN-PED2	8.3	ENRON	8.6
CAMP-STAD	10.9	NUS-WIDE	12.8
SUN-PED2	15.1	TMC2007	16.8
BAHN-SUN	5.6	YEAST	9.4

6. Conclusion

In this paper, we investigated active learning for two structured prediction tasks: learning to make cuts in graphs and learning bipartite matchings. Though structured support vector machines can be efficiently employed for these tasks, they are not very useful for guiding label solicitation strategies. Conditional random fields, which do provide useful correlation estimates for computing value of information cannot be applied efficiently for these tasks (e.g., #P-hard for bipartite matchings). We introduced active learning based on adversarial structured prediction methods that enjoy lower computational complexity than existing probabilistic methods while providing useful correlations between variables. We demonstrated the benefits of this approach on two structured prediction tasks: multi-label image tagging as a cut learning task and object tracking through consecutive video frames as a bipartite matching task.

Acknowledgements

This work was supported, in part, by the National Science Foundation under Grant No. 1652530.

References

Asif, K., Xing, W., Behpour, S., and Ziebart, B. D. Adversarial cost-sensitive classification. In *UAI*, pp. 92–101, 2015.

Behpour, S., Xing, W., and Ziebart, B. Arc: Adversarial robust cuts for semi-supervised and multi-label classification. In *AAAI Conference on Artificial Intelligence*, 2018.

Brinker, K. On active learning in multi-label classification. In *From Data and Information Analysis to Knowledge Engineering*, pp. 206–213. Springer, 2006.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2012.

Dasarathy, G., Nowak, R., and Zhu, X. S2: An efficient graph based active learning algorithm with application to nonparametric classification. In *Conference on Learning Theory*, pp. 503–522, 2015.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, July 2011. ISSN 1532-4435.

Fathony, R., Behpour, S., Zhang, X., and Ziebart, B. Efficient and consistent adversarial bipartite matching. In *International Conference on Machine Learning*, pp. 1456–1465, 2018.

Godbole, S. and Sarawagi, S. Discriminative methods for multi-labeled classification. In *Pacific-Asia conference on knowledge discovery and data mining*, pp. 22–30. Springer, 2004.

Grandvalet, Y. and Bengio, Y. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pp. 529–536, 2005.

Grünwald, P. D. and Dawid, A. P. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, 32:1367–1433, 2004.

Guo, Y. and Greiner, R. Optimistic active-learning using mutual information. In *IJCAI*, volume 7, pp. 823–829, 2007.

Joachims, T. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pp. 137–142. Springer, 1998.

Khodabandeh, M., Deng, Z., Ibrahim, M. S., Satoh, S., and Mori, G. Active learning for structured prediction from partially labeled data. *arXiv preprint arXiv:1706.02342*, 2017.

Kim, S., Kwak, S., Feyereisl, J., and Han, B. Online multi-target tracking by large margin structured learning. In *Asian Conference on Computer Vision*, pp. 98–111. Springer, 2012.

Lafferty, J. D., McCallum, A., and Pereira, F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289. Morgan Kaufmann Publishers Inc., 2001.

Lambrou, A., Papadopoulos, H., Nouretdinov, I., and Gammerman, A. Reliable probability estimates based on support vector machines for large multiclass datasets. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 182–191. Springer, 2012.

Leal-Taixé, L., Milan, A., Reid, I., Roth, S., and Schindler, K. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015.

Lewis, D. D. and Gale, W. A. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 3–12. Springer-Verlag New York, Inc., 1994.

Liu, A., Reyzin, L., and Ziebart, B. D. Shift-pessimistic active learning using robust bias-aware prediction. In *AAAI*, pp. 2764–2770, 2015.

McMahan, H. B., Gordon, G. J., and Blum, A. Planning in the presence of cost functions controlled by an adversary. In *Proceedings of the International Conference on Machine Learning*, pp. 536–543, 2003.

Mozhdehi, R. J. and Medeiros, H. Deep convolutional particle filter for visual tracking. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3650–3654. IEEE, 2017.

Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Settles, B. *Curious machines: Active learning with structured instances*. PhD thesis, University of Wisconsin–Madison, 2008.

Settles, B. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

Seung, H. S., Opper, M., and Sompolinsky, H. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 287–294. ACM, 1992.

Sun, Q., Laddha, A., and Batra, D. Active learning for structured probabilistic models with histogram approximation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3612–3621, 2015.

Taskar, B., Chatalbashev, V., Koller, D., and Guestrin, C. Learning structured prediction models: A large margin approach. In *Proceedings of the International Conference on Machine Learning*, pp. 896–903. ACM, 2005.

Tong, S. and Koller, D. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

Topsøe, F. Information-theoretical optimization techniques. *Kybernetika*, 15(1):8–27, 1979.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. Large margin methods for structured and interdependent output variables. *Journal of machine learning research*, 6(Sep):1453–1484, 2005.

Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., and Vlahavas, I. Mulan: A java library for multi-label learning. *Journal of Machine Learning Research*, 12:2411–2414, 2011.

Yang, B., Sun, J.-T., Wang, T., and Chen, Z. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 917–926. ACM, 2009.

Ye, C., Wu, J., Sheng, V. S., Zhao, P., and Cui, Z. Multi-label active learning with label correlation for image classification. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pp. 3437–3441. IEEE, 2015.

Zhang, M.-L. and Zhou, Z.-H. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40 (7):2038–2048, 2007.