Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)

Multi-Task Active Learning with Output Constraints

Yi Zhang
Machine Learning Department
Carnegie Mellon University

Abstract

Many problems in information extraction, text mining, natural language processing and other fields exhibit the same property: multiple prediction tasks are related in the sense that their outputs (labels) satisfy certain constraints. In this paper, we propose an active learning framework exploiting such relations among tasks. Intuitively, with task outputs coupled by constraints, active learning can utilize not only the uncertainty of the prediction in a single task but also the inconsistency of predictions across tasks. We formalize this idea as a cross-task value of information criteria, in which the reward of a labeling assignment is propagated and measured over all relevant tasks reachable through constraints. A specific example of our framework leads to the cross entropy measure on the predictions of coupled tasks, which generalizes the entropy in the classical single-task uncertain sampling. We conduct experiments on two real-world problems: web information extraction and document classification. Empirical results demonstrate the effectiveness of our framework in actively collecting labeled examples for multiple related tasks.

1 Introduction

Many real-world problems exhibit the same property: multiple prediction tasks are related in the sense that their outputs need to satisfy certain constraints. In information retrieval and text mining, classifying documents and web pages into a set of predefined categories are treated as multiple tasks, but these categories are usually defined by a taxonomy of inheritance semantics. In information extraction, recognizing different entities and relations are separate prediction tasks, but their outputs are often related, e.g., x is the mayor of a city tells that x is also a politician, a person, but not food.

These constraints couple the task outputs and provide valuable information. As a result, researchers have recently proposed to leverage constraints of task outputs to improve supervised and semi-supervised learning (Chang, Ratinov, and Roth 2007; Chang et al. 2008; Carlson et al. 2010), where constraints are used to regularize either the estimation of model parameters or the inference on unlabeled data.

In this paper, we study active learning on multiple tasks when their outputs are coupled by constraints. Intuitively, with tasks coupled by a set of constraints, active learning can utilize not only the uncertainty of the prediction in a single task but also the inconsistency of predictions across tasks. For example, if two models classify an object as positive examples for two mutually exclusive classes, we can be certain that at least one model is making an incorrect prediction. This kind of certainty can not be achieved in traditional single-task active learning due to the lack of true labels.

The value of information in decision theory offers a systematic treatment for active learning and observation selection problems (Kapoor, Horvitz, and Basu 2007; Krause and Guestrin 2009), where different choices of reward functions give different active learning heuristics. We formalize our idea of constraint-driven active learning across tasks as a value of information framework, in which each possible labeling assignment for a task is first propagated to all relevant tasks reachable through constraints and the reward is measured over all relevant tasks. A specific choice of reward function in our framework leads to the cross entropy measure on the predictions of coupled tasks, which generalizes the entropy measure used in the classical single-task uncertainty sampling and also highlights the role of task inconsistency in multi-task active learning.

We conduct experiments on two real-world problems: web information extraction and document classification. Regardless of the choice of reward functions, our multi-task active learning approaches with constraints consistently outperform the corresponding single-task selection methods. In this sense, we suggest that the proposed framework should be considered as a standard procedure for actively collecting labeled examples for large numbers of coupled tasks, e.g., classification for a taxonomy of categories.

The rest of the paper is organized as follows. In Section 2, we review the notion of value of information for active learning. In Section 3 we propose our framework of constraint-driven multi-task active learning. Empirical studies are presented in Section 4. We discuss related work in Section 5. In Section 6 we conclude the paper and discuss future work.

2 Value of Information for Active Learning

In this section we review the notion of value of information for active learning. Suppose our goal is to build a prediction model $\hat{p} = \hat{p}(Y|\mathbf{x})$: given an example $\mathbf{x}$ from the input space $\mathcal{X}$, we can predict the conditional probability of its label Y,

i.e., $p(Y = y|\mathbf{x}), y \in Dom(Y)$. In traditional pool-based active learning (for a single task), we have a set of unlabeled samples $\mathbf{U}$. We want to actively choose unlabeled samples from this pool for labeling requests, so that the prediction performance of the model learned from labeled examples is maximized. The key is to measure how useful labeling a sample $\mathbf{x} \in \mathbf{U}$ will be for improving the current model $\hat{p} = \hat{p}(Y|\mathbf{x})$. This can be viewed as measuring the value of information (Krause and Guestrin 2009) for requesting the unknown label Y on each unlabeled sample $\mathbf{x} \in \mathbf{U}$:

$$VOI(Y, \mathbf{x}) = \sum_y P(Y = y|\mathbf{x}) R(\hat{p}, Y = y, \mathbf{x}) \quad (1)$$

This formula shows that the value of information for a labeling request $(Y, \mathbf{x})$ is the sum of the *reward* of each possible labeling outcome $Y = y$ for the current model $\hat{p}$, denoted as $R(\hat{p}, Y = y, \mathbf{x})$, weighted by the probability of this outcome $P(Y = y|\mathbf{x})$. The true label probability $P(Y = y|\mathbf{x})$ is unknown, and in most cases[1], it is replaced by the estimate from the current model $\hat{p}$:

$$VOI(Y, \mathbf{x}) = \sum_y \hat{p}(Y = y|\mathbf{x}) R(\hat{p}, Y = y, \mathbf{x}) \quad (2)$$

The reward function is the key part of value of information in eq. (2). A reasonable heuristic is to measure how surprising is the labeling outcome $(Y = y, \mathbf{x})$ given current model $\hat{p}$. Two reward functions following this heuristic are:

$$R(\hat{p}, Y = y, \mathbf{x}) = -\log_2 \hat{p}(Y = y|\mathbf{x}) \quad (3)$$

$$R(\hat{p}, Y = y, \mathbf{x}) = 1 - \delta(y, \operatorname*{argmax}_{y'} \hat{p}(Y = y'|\mathbf{x})) \quad (4)$$

The function in eq. (3) is the optimal code length of the outcome if the distribution is given by the current model $\hat{p}$. In this case, an impossible outcome (with $\hat{p} = 0$) has an infinite reward, and an already known outcome (with $\hat{p} = 1$) has no value. The second function in eq. (4) takes the value 0 if the labeling outcome y coincides with the most likely predict y' (i.e., no surprise and no reward) and 1 otherwise. We can view eq. (3) as the log reward for $\hat{p}(Y|\mathbf{x})$ if the true label is y, and eq. (4) as the $0/1$ reward (Roy and Mccallum 2001).

Incorporating the reward eq. (3) into eq. (2), we have:

$$VOI(Y, \mathbf{x}) = -\sum_y \hat{p}(Y = y|\mathbf{x}) \log_2 \hat{p}(Y = y|\mathbf{x}) \quad (5)$$

which is the entropy of the predicted distribution used in uncertain sampling. Similarly, the reward in eq. (4) leads to:

$$VOI(Y, \mathbf{x}) = 1 - \max_y \hat{p}(Y = y|\mathbf{x}) \quad (6)$$

which is the criteria used in least-confident sampling (Settles 2009). Many active learning heuristics are equal to maximizing the value of information with different reward functions $R(\hat{p}, Y = y, \mathbf{x})$, e.g., estimated error reduction on the whole unlabeled set (Roy and Mccallum 2001).

[1]There are exceptions: Guo and Greiner (2007) replace the true label probability with an *optimistic* guess maximizing the reward.

3 Multi-Task Active Learning with Constraints

In this section, we propose our framework for multi-task active learning with output constraints. In Section 3.1 we introduce the key component of our method: the notion of cross-task value of information. In Section 3.2 we analyze the proposed criteria and show how inconsistency of predictions among tasks is captured. In Section 3.3 we discuss a few extensions to our framework, e.g., launch new tasks, manage supplementary tasks and handle tasks with different input spaces. In Section 3.4 we give the complete algorithm.

3.1 Cross-Task Value of Information

Consider a set of T tasks, each with a (categorical) response variable Y_i, $i = 1, 2, \ldots, T$. Our goal is to learn a classifier for each task: $\hat{p}_i = \hat{p}_i(Y_i|\mathbf{x}), i = 1, 2, \ldots, T$. Each sample in our training set $\mathbf{x} \in \mathbf{U}$ is associated with T labels. For each sample, we might know some (or none) of its T labels. We use $UL(\mathbf{x})$ to denote the set of unknown labels on a sample $\mathbf{x}$: $UL(\mathbf{x}) = \{Y_i : Y_i \text{ is unknown for } \mathbf{x}\}$.

In multi-task active learning, we need to choose both the sample and the task for labeling. We measure the value of information for the sample-task pair $(Y_i, \mathbf{x})$ as follows:

$$VOI(Y_i, \mathbf{x}) = \sum_{y_i} \hat{p}_i(Y_i = y_i|\mathbf{x}) R(Y_i = y_i, \mathbf{x}) \quad (7)$$

where $\mathbf{x} \in \mathbf{U}$ is a sample from our training set, $Y_i \in UL(\mathbf{x})$ is an unknown label on $\mathbf{x}$, and $R(Y_i = y_i, \mathbf{x})$ is the reward function for a possible labeling outcome for $(Y_i = y_i, \mathbf{x})$. As in eq. (2), the true label probability $P(Y_i = y_i|\mathbf{x})$ is unknown and replaced by the estimate from the current ith model $\hat{p}_i(Y_i = y_i|\mathbf{x})$. Note that we will discuss managing completely new tasks (i.e., $\hat{p}_i$ not available) in Section 3.3 .

The key question is to decide the reward function $R(Y_i = y_i, \mathbf{x})$ for each possible labeling outcome $(Y_i = y_i, \mathbf{x})$. The set of constraints among task outputs, denoted as $\mathbf{C}$, provides important information on what other facts we can infer from a given outcome $(Y_i = y_i, \mathbf{x})$. To formalize this, we define the set of *propagated outcomes* $Prop_{\mathbf{C}}(Y_i = y_i)$ as the labeling outcomes we can *infer* from the assignment $Y_i = y_i$ based on the set of constraints $\mathbf{C}$:

$$Prop_{\mathbf{C}}(Y_i = y_i) = \{Y_j = y_j \mid Y_i = y_i \dashrightarrow_{\mathbf{C}} Y_j = y_j\} \quad (8)$$

Inference of outcomes is based on the rules provided by constraints. For example, the inheritance constraint "Y_j is a derived class of Y_i" provides two rules "$Y_i = 0 \rightarrow Y_j = 0$" and "$Y_j = 1 \rightarrow Y_i = 1$". The mutual exclusion constraint "Y_i and Y_j are mutually exclusive classes" gives two rules "$Y_i = 1 \rightarrow Y_j = 0$" and "$Y_j = 1 \rightarrow Y_i = 0$". The agreement constraint "Y_i and Y_j must agree", which is common when we use multiple views to predict the same target variable, brings the rules "$Y_i = y \rightarrow Y_j = y$" and "$Y_j = y \rightarrow Y_i = y$". Note that even without any rule, i.e., $\mathbf{C} = \emptyset$, we still have $Prop_{\mathbf{C}}(Y_i = y_i) = \{(Y_i = y_i)\}$ since a labeling outcome at least indicates itself. Also, the $\dashrightarrow$ in eq. (8) means that we also include the outcomes we

indirectly infer from $Y_i = y_i$, i.e., by sequentially applying available rules multiple times.

Based on the notion of propagated outcomes, we define the reward function $R(Y_i = y_i, \mathbf{x})$ in eq. (7) as follows:

$$R(Y_i = y_i, \mathbf{x}) = \sum_{\substack{Y_j = y_j \in Prop_{\mathbf{C}}(Y_i = y_i) \\ Y_j \in UL(\mathbf{x})}} R(\hat{p}_j, Y_j = y_j, \mathbf{x}) \quad (9)$$

where $R(\hat{p}_j, Y_j = y_j, \mathbf{x})$ is the single-task reward function discussed in Section 2, two examples of which are given as eq. (3) and eq. (4). Also, we only consider inferred labeling outcomes Y_j that are unknown on sample $\mathbf{x}$: $Y_j \in UL(\mathbf{x})$.

Computationally, the set of propagated outcomes $Prop_{\mathbf{C}}(Y_i = y_i)$ for each labeling outcome $Y_i = y_i$ can be pre-computed and cached for efficiently access during active learning. To pre-compute this set, we can construct a directed graph, where each label assignment $Y_i = y_i$ ($y_i \in Dom(Y_i), i = 1, 2, \ldots, T$) is a node and each propagation rule is a directed edge. Computing the set of prorogated outcomes from an outcome $Y_i = y_i$ is equivalent to finding the set of reachable nodes from the corresponding node in the graph. The size of the graph only depends on the number of tasks T and the cardinality of response variables $|Dom(Y_j)|_{j=1}^{T}$. Thus this computation is often tractable.

Incorporating eq. (9) into eq. (7), we define the cross-task value of information for a sample-task pair as:

$$VOI(Y_i, \mathbf{x}) = \sum_{y_i} \hat{p}_i(Y_i = y_i|\mathbf{x}) \sum_{\substack{Y_j = y_j \in Prop_{\mathbf{C}}(Y_i = y_i) \\ Y_j \in UL(\mathbf{x})}} R(\hat{p}_j, Y_j = y_j, \mathbf{x}) \quad (10)$$

3.2 Analysis

In this section we analyze the cross-task value of information in eq. (10). We choose the single-task reward function as eq. (3). Analysis is similar for other reward functions.

Including the reward eq. (3) into eq. (10), we have:

$$VOI(Y_i, \mathbf{x}) = \sum_{y_i} \hat{p}_i(Y_i = y_i|\mathbf{x}) \sum_{\substack{Y_j = y_j \in Prop_{\mathbf{C}}(Y_i = y_i) \\ Y_j \in UL(\mathbf{x})}} -\log_2 \hat{p}_j(Y_j = y_j|\mathbf{x}) \quad (11)$$

For a given unlabeled sample-task pair $(Y_i, \mathbf{x})$, the value of information in eq. (11) is determined by: 1) the model $\hat{p}_i$ predicting the probability of each possible assignment $(Y_i = y_i|\mathbf{x})$; 2) propagated facts $(Y_j = y_j) \in Prop_{\mathbf{C}}(Y_i = y_i)$ inferred from each assignment $Y_i = y_i$; 3) other models predicting the probability $\hat{p}_j(Y_j = y_j|\mathbf{x})$ of each inferred outcome $Y_j = y_j$. In particular, $VOI(Y_i, \mathbf{x})$ tends to be high if some assignment $Y_i = y_i$ is highly probable according to $\hat{p}_i$, but the inferred outcome $Y_j = y_j$ is unlikely according to another model $\hat{p}_j$. In this case, both $\hat{p}_i(Y_i = y_i|\mathbf{x})$ and $-\log_2 \hat{p}_j(Y_j = y_j|\mathbf{x})$ will be high and the total value of information $VOI(Y_i, \mathbf{x})$ is significantly increased. In this sense, maximizing the value of information tends to select the sample-task pair on which the current models are contradicting each other given the set of constraints.

The criteria in eq. (11) can be viewed as the sum of cross entropy measures between the model $\hat{p}_i$ and other coupled models $\hat{p}_j$. Recall that the cross entropy of two distributions $P_i(y)$ and $P_j(y)$ is defined as:

$$H(P_i, P_j) = -\sum_{y} P_i(y) \log_2 P_j(y) \quad (12)$$

which is the average coding length of a variable y with distribution P_i when using an optimal code for another distribution P_j. Intuitively, this coding length increases with the discrepancy of P_j from P_i, so it captures the inconsistency of two distributions. Note that P_i and P_j in eq. (12) are defined on the same quantity y, but any two predicted distributions $\hat{p}_i$ and $\hat{p}_j$ in eq. (11) are defined on different target variables. The constraints $\mathbf{C}$ plays a key part in eq. (11): coupling predicted distributions on different variables. As a result, cross-task value of information in eq. (11) is essentially the sum of cross entropy measures between the predicted distribution $\hat{p}_i$ and other coupled predicted distributions.

3.3 Extensions

We discuss a few potential extensions to our framework that are useful in some real-world systems: launch new tasks, differentiate target and supplementary tasks, and couple prediction models in different input spaces.

The first extension is to launch new tasks: we may start collecting labeled examples for a few tasks on our taxonomy, and new tasks (and constraints) can be continuously added into the taxonomy later. When new tasks are added into the system, we may prefer to collect more labels for them since they have no (or few) labeled examples. This can be naturally handled by our framework. For a completely new task i (without labeled example), we can specify the ith prediction model $\hat{p}_i(Y_i|\mathbf{x})$ in eq. (10) as random guessing (e.g.,, a uniform distribution on all possible assignments). This will lead to more frequent contradictions with other predictions $\hat{p}_j$ on propagated outcomes $Y_j = y_j \in Prop_{\mathbf{C}}(Y_i = y_i)$ and thus large value of $[\hat{p}_i(Y_i = y_i|\mathbf{x}) \cdot -\log_2 \hat{p}_j(Y_j = y_j|\mathbf{x})]$. Similarly, the model for a task with few labeled examples is inaccurate and tends to contradict with other models, and thus the task is more likely to be chosen based on eq. (10).

The second extension is to enable us to specify a subset of tasks as *target* tasks (denoted by $\mathbf{G} \subseteq \{1, 2, \ldots, T\}$) and other tasks as *supplementary* tasks. Our goal is to improve the prediction performance on target tasks. To achieve this, we redefine eq. (9) (and consequently eq. (10)) as:

$$R(Y_i = y_i, \mathbf{x}) = \sum_{\substack{Y_j = y_j \in Prop_{\mathbf{C}}(Y_i = y_i) \\ Y_j \in UL(\mathbf{x}) \wedge j \in \mathbf{G}}} R(\hat{p}_j, Y_j = y_j, \mathbf{x}) \quad (13)$$

in which the reward $R(Y_i = y_i, \mathbf{x})$ includes rewards from propagated outcomes $Y_j = y_j$ only if $j \in \mathbf{G}$. As a result, maximizing the VOI in eq. (10) (redefined with $j \in \mathbf{G}$) will select the sample-task pair which brings significant gains on target tasks. A supplementary task will be chosen only if it gives nontrivial rewards on target tasks. Note that eq. (10) (redefined with $j \in \mathbf{G}$) will contain the entropy term $\sum_{y_i} \hat{p}_i(Y_i = y_i|\mathbf{x}) \cdot -\log_2 \hat{p}_i(Y_i = y_i|\mathbf{x})$ if i is a target task. If a completely new task i is a target task, this entropy

term will be large since $\hat{p}_i$ is a uniform distribution, i.e., we will strongly prefer to label examples for a new target task.

The last extension is to handle tasks defined on different input spaces. This is useful, e.g., for coupling entity recognition tasks and relation extraction tasks. To be more concrete, we may have three tasks: task 1 is to recognize the class "politician", task 2 is to recognize the class "city", and task 3 is to extract the relation "is the mayor of". A constraint is that "x_1 is the mayor of x_2 only if x_1 is a politician and x_2 is a city". To handle these tasks, we need to redefine the notion of propagated outcomes in eq. (8) to include not only the transition of the output variable but also the change of input variables. This part will be studied in future work.

3.4 The Complete Algorithm

In this section we summarize the complete algorithm:

- 1. Choose the base learner $\hat{p}_i$ for each of the T task. In this work, we use multinomial Naive Bayes (McCallum and Nigam 1998) for simplicity and training efficiency.
- 2. Choose the single-task reward, e.g., eq. (3) or (4).
- 3. Specify the set of constraints $\mathbf{C}$ between task outputs, construct the propagation rules and pre-compute the set of propagated outcomes $Prop_{\mathbf{C}}(Y_i = y_i)$ for each labeling outcome $Y_i = y_i$, as discussed in Section 3.1 .
- 4. Choose between the *balance* mode and the *free* mode. In the balance mode, we require that each T labeling requests must be uniformly from T tasks. In the free mode we do not add this restriction.
- 5. Train a model for each task using seed examples. Use random prediction as the initial model if no seed available.
- 6. Compute the value of information for each unlabeled sample-task pair $(Y_i, \mathbf{x})$ according to eq. (10).
- 7. Choose the sample-task pair that maximizes the VOI, without violating the balance restriction (if specified).
- 8. Obtain the labeling outcome for the chosen sample-task pair, and perform propagation according to eq. (8).
- 9. Retrain prediction models with new labeled examples.
- 10. Return to step 6 if more labeling requests are allowed.

4 Empirical Studies

We conduct our empirical studies on two real-world problems: web information extraction and document classification. In this section we discuss our experiments and results.

4.1 Experimental Settings

The first experiment on web information extraction is based the data collected from the CMU Reading the Web project[2]. The project uses semi-supervised learning methods to extract symbolic knowledge (as instances of different entity classes and relations) from the Web. We use 6 classes of 708 named entities (noun phrases) extracted by the system, which include 242 animals, 107 mammals, 200 companies[3], 33 newspaper companies, 171 food and 95 celebrities. Each example is represented by a 99400 dimensional vector, describing its co-occurrence frequencies with the most frequent 99400 contexts on the Web. As a result, the data is a 708×99400 matrix. We define 6 tasks, each to recognize an entity class. The constraints between tasks are: inheritance between "mammals" and "animals", inheritance between "newspaper companies" and "companies", and mutual exclusion between other pairs of classes. We repeat 5 random runs. In each run, we randomly split $2/3$ examples into the training set and the rest $1/3$ examples as the testing set. In the training set, we randomly label 30 seed examples for each task to initialize the multinomial naive Bayes model, and make sure that for each task at least 3 seed examples are positive. We test different active learning methods using the procedure described in Section 3.4 . Details of the methods will be discussed later in this section. We test the performance of each method on the testing set as more labels are requested. We use the average AUC (area under the ROC curve) score over 6 tasks as the performance measure, and the results are further averaged over 5 random runs.

The second experiment on document classification is based on the RCV1 corpus (Lewis et al. 2004), a benchmark collection of over $800,000$ newswire stories from Reuters. In this data set, each article is represented as a 47236 dimensional TF-IDF weight vector, and labeled for 101 topics. In the experiment, we use a subset of the collection[4], which contains 3000 training examples and 3000 testing examples. Some topics are very rare, so we select 36 frequent topics with at least 1% positive examples. We study 36 classification tasks, each to identify a topic. Constraints that couple task outputs are defined by extracting the propagation rules with 100% precision from the corpus. We use these rules to construct sets of propagated outcomes in eq. (8). To start the experiment, we randomly provide 60 labels from the training set for each task to initialize the model. Other settings are same as in the information extraction experiment.

We compare 8 active learning strategies, which result from different choices of three options:

- 1: the log reward in eq. (3) vs. the 0/1 reward in eq. (4).
- 2: the free mode vs. the balance mode (see Section 3.4).
- 3: without constraint vs. with constraints

Note that, without constraint to couple tasks (i.e, $\mathbf{C} = \emptyset$), the cross-task value of information in eq. (10) for a sample-task pair $(Y_i, \mathbf{x})$ is equal to the value of information in eq. (2) for task i. In this case, the balance mode is exactly to perform active learning on each task iteratively, and the free mode may continuously request labeling on the same task if the reward computed as eq. (2) dominates other tasks.

4.2 Results

Empirical results of information extraction experiments are shown in Fig. 1a – Fig. 1d, and results for document classification are presented in Fig. 2a – Fig. 2d. All figures display the change of average AUC score as the number of

[2] http://rtw.ml.cmu.edu/readtheweb.html

[3] We use a subset of companies to not overwhelm other classes.

[4] http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

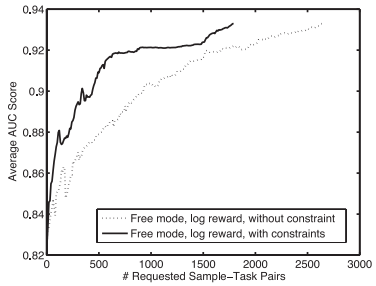
requested labels increases. Each subfigure makes a comparison between active learning with and without constraints. Subfigures differ in the choice of reward functions (log reward or 0/1 reward) and modes (free mode or balance mode). From the results we can see that active learning based on cross-task value of information consistently outperforms active learning using the single-task counterpart, regardless of the choice of reward functions and algorithm modes. Prediction performance is improved significantly faster when labeling requests are chosen according to a cross-task selection criteria, and much less labeling requests are needed for the system to achieve good AUC scores. In addition, the 0/1 reward function in eq. (4) is more effective than the log reward function in eq. (3). Also, the balance mode and the free mode perform comparably well. Overall speaking, the combination of cross-task value of information, the 0/1 reward fucntion, and balance mode (Fig. 1d and Fig. 2d) achieves top performance in both web information extraction and document classification experiments.

5 Related Work

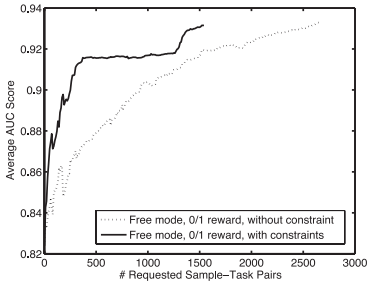
Co-testing (Muslea, Minton, and Knoblock 2006) is an active learning method proposed for multi-view problems, where examples receiving different predictions from multiple views are selected for labeling. Multi-view learning can be viewed as a special case of multi-task learning with output constraints, where tasks are to predict the same variable from different views, and constraints among tasks are *agreement* constraints: predictions should agree. In this sense, our work is a non-trivial extension from multi-view to multi-task active learning. Multi-task (or multi-label) active learning has recently been studied. Reichart et al. (2008) present some heuristics such as iteratively selecting samples from different tasks or aggregating the selection scores from tasks. Qi et al. (2008) propose to estimate the correlation of labels directly from training examples and use the resulting joint label distribution to guide active learning. Another related direction is active learning for structured prediction. Roth and Small (2006) study the tradeoff between querying the labels for an entire structured instance and querying the labels for subcomponents of instances.

6 Conclusion and Future Work

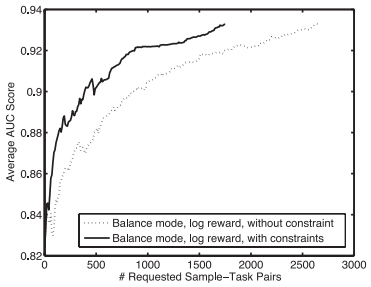
In this paper, we present a systematic framework to incorporate output constraints into the multi-task (-label) active learning process. With task outputs coupled by constraints, a cross-task value of information criteria is designed to measure both the uncertainty and inconsistency of predictions over tasks. A specific example of our framework leads to the cross entropy measure on the predictions of coupled tasks, which generalizes the single-task uncertain sampling using entropy. We conduct experiments on two real-world problems: web information extraction and document classification. Results on both problems demonstrate the effectiveness of our framework in actively collecting labeled examples for multiple related tasks. The proposed framework may be considered as a natural choice to replace standard active learning strategies in any learning system where multiple prediction problems are coupled by output constraints.



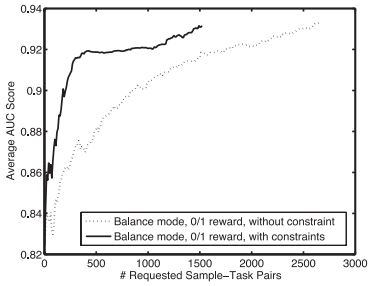
(a) Free mode, log reward function



(b) Free mode, 0/1 reward function



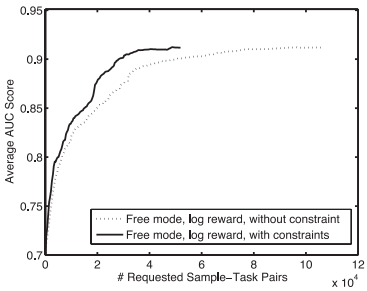
(c) Balance mode, log reward function



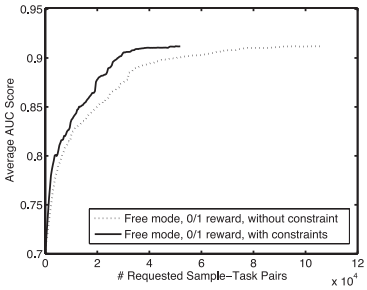
(d) Balance mode, 0/1 reward function

Figure 1: Performance of 8 active learning strategies in web information extraction experiments.

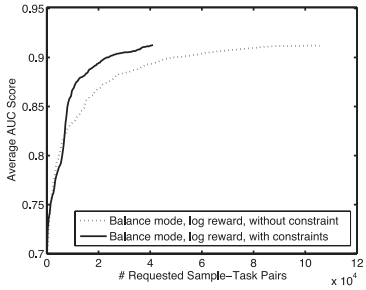
Future work will concentrate on expanding or updating the predefined set of output constraints during the learning process. It is also interesting to include more complex types



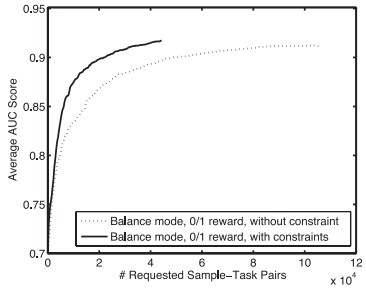
(a) Free mode, log reward function



(b) Free mode, 0/1 reward function



(c) Balance mode, log reward function



(d) Balance mode, 0/1 reward function

Figure 2: Performance of 8 active learning strategies in document classification experiments.

of constraints into the proposed framework, e.g., constraints involving three or more labels. Another direction is to use output constraints in modeling the joint predictive distribution of multiple tasks. The resulting joint label distribution combines the information from both labeled examples and domain knowledge, and can be used to either guide the active learning process or make more accurate predictions.

7 Acknowledgments

We thank Dr. Tom Mitchell for helpful discussions and resources from the CMU Reading the Web Project.

References

Carlson, A.; Betteridge, J.; Wang, R.; Hruschka, E. R.; and Mitchell, T. M. 2010. Coupling semi-supervised learning of information extraction. In *WSDM*.

Chang, M.-W.; Ratinov, L.; Rizzolo, N.; and Roth, D. 2008. Learning and inference with constraints. In *AAAI*.

Chang, M.-W.; Ratinov, L.-A.; and Roth, D. 2007. Guiding semi-supervision with constraint-driven learning. In *ACL*.

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

Kapoor, A.; Horvitz, E.; and Basu, S. 2007. Selective supervision: guiding supervised learning with decision-theoretic active learning. In *IJCAI*, 877–882.

Krause, A., and Guestrin, C. 2009. Optimal value of information in graphical models. *J. Artif. Intell. Res.* 35:557–591.

Lewis, D. D.; Yang, Y.; Rose, T. G.; and Li, F. 2004. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.* 5:361–397.

McCallum, A., and Nigam, K. 1998. A comparison of event models for naive bayes text classification. In *IN AAAI-98 WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION*, 41–48. AAAI Press.

Muslea, I.; Minton, S.; and Knoblock, C. A. 2006. Active learning with multiple views. *J. Artif. Intell. Res.* 27:203–233.

Qi, G.-J.; Hua, X.-S.; Rui, Y.; Tang, J.; and Zhang, H.-J. 2008. Two-dimensional active learning for image classification. In *CVPR*.

Reichart, R.; Tomanek, K.; Hahn, U.; and Rappoport, A. 2008. Multi-task active learning for linguistic annotations. In *Proceedings of ACL: HLT*, 861–869.

Roth, D., and Small, K. 2006. Margin-based active learning for structured output spaces. In *ECML*, 413–424.

Roy, N., and Mccallum, A. 2001. Toward optimal active learning through sampling estimation of error reduction. In *ICML*, 441–448.

Settles, B. 2009. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison.