Uncertainty Sampling and Transductive Experimental Design for Active Dual Supervision

Vikas Sindhwani VSINDHW@US.IBM.COM
Prem Melville PMELVIL@US.IBM.COM
Richard D. Lawrence RICKLAWR@US.IBM.COM
Mathematical Sciences, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

Abstract

Dual supervision refers to the general setting of learning from both labeled examples as well as labeled features. Labeled features are naturally available in tasks such as text classification where it is frequently possible to provide domain knowledge in the form of words that associate strongly with a class. In this paper, we consider the novel problem of active dual supervision, or, how to optimally query an example and feature labeling oracle to simultaneously collect two different forms of supervision, with the objective of building the best classifier in the most cost effective manner. We apply classical uncertainty and experimental design based active learning schemes to graph/kernel-based dual supervision models. Empirical studies confirm the potential of these schemes to significantly reduce the cost of acquiring labeled data for training high-quality models.

1. Introduction

The performance of learning algorithms is typically bounded by the amount and quality of labeled examples. To streamline the costly process of acquiring labeled data, it is often worthwhile to turn to active learning. Traditional active learning schemes query a human for labels of intelligently chosen examples. However, human effort may also be profitably expended in collecting alternative forms of supervision. Consider, for instance, the text-categorization task of classifying movie reviews as expressing positive or negative opinion. One could read through several reviews labeling them based on the sentiment expressed therein. Alternatively, one could look though a list of words (features) commonly occurring in movie reviews, such as *mesmerizing* and *boring*, and label them as positive or negative. We refer to the latter form of supervision as *feature labels.* This kind of labeling is qualitatively different from reading and labeling documents – it requires a human to condense prior linguistic experience with a word into a sentiment label that reflects the net emotion that the word evokes. Ideally, one should be able to learn a classifier from a combination of both kinds of supervisory inputs. In general, these forms of supervision are not mutually redundant, have different acquisition costs, human annotation quality and degrees of utility towards learning a dual supervision model.

Recent papers (Melville et al., 2009; Sindhwani et al., 2008; Druck et al., 2008) have demonstrated that the availability of labeled features can significantly reduce the number of labeled examples required to build high-quality classifiers. Having algorithms that can incorporate dual supervision, gives rise to the novel task of active learning in this setting – i.e., how can we select the most informative examples and/or features to be labeled, so as to build the best model at the lowest cost. In this paper we explore this task of Active Dual Supervision. While traditional active learning focuses on estimating the value of acquiring the label of unlabeled examples, in active dual supervision, it is important to also estimate the value of information of feature labels – this has not been attempted before to the best of our knowledge. Furthermore, an ideal scheme should be able to trade-off the costs and benefits of the different forms of labels.

In this paper, we apply classical uncertainty and experimental design based schemes for feature-side active learning and active dual supervision. At their core, our methods utilize a new transductive bipartite graph approach for dual supervision closely related to

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009. Copyright 2009 by the author(s)/owner(s).

the framework introduced in (Sindhwani et al., 2008). We report extensive empirical results to emphasize the value of feature labels, and to demonstrate the hitherto untapped potential of feature-side active learning and active dual supervision.

2. Learning with Dual Supervision

Consider a document classification problem where we are given an $n \times d$ document-term matrix $\mathbf{X}$, expressing n documents as bag-of-words feature vectors over a vocabulary of d words. Let $\mathbf{y}$ denote the associated $n \times 1$ document labels vector with $y_i \in \{+1, 0, -1\}$ where 0 implies that the document is unlabeled. In a dual supervision setting, we also have access to a sparse $d \times 1$ feature labels vector $\boldsymbol{z}$ where $z_i = 1$ suggests that the corresponding word has strong affinity with the positive class and $z_i = -1$ similarly suggests affinity with the negative class. Given minimal dual supervision in the form of highly sparse vectors $\mathbf{y}, \boldsymbol{z}$, the goal is to construct an accurate classification model. Note that since text classification is the primary application considered in this paper, we use documents/words interchangeably with examples/features.

Our dual supervision model is a graph-based tranductive model inspired by the methods proposed in (Sindhwani et al., 2008). The vectors $(\mathbf{y}, \boldsymbol{z})$ are treated as partial labels on the vertices of a bipartite graph that represents the data matrix. Documents form one set of vertices and words the other, with edges representing occurrence of a word in a document. Starting from the partial labeling, the key intuition behind this approach is to effectively *diffuse* label information to unlabeled data *from both sides* of the data matrix. The overall setup of this paper is schematically depicted in Figure 1, showing labeled words for the sentiment classification application mentioned in Section 1. In the active learning setting, we also have access to *document* and *word oracles*. An active dual supervision scheme attempts to identify the most useful labels – for documents and/or words – to acquire from these oracles, with the goal of building the best predictive model. We discuss such schemes in the next section. For the rest of this section, we focus on the static setting where a fixed partial labeling $(\mathbf{y}, \boldsymbol{z})$ is available over examples and features.

In practice, to study this setting, we need to simulate partial labellings on a collection of datasets. It is convenient for us to describe upfront the 5 popular binary text classification datasets used in various studies throughout this paper. The movies dataset (2000 examples, 24841 features), popular in the sentiment analysis literature, poses the task of classifying the sentiment of movie reviews into positive or negative. ibm-mac (1937 examples, 9822 features), baseball-hockey (1988 examples, 12148 features) and med-space (1972 examples, 17084 features) datasets are drawn from the 20-newsgroups text collection where the task is to assign messages into the newsgroup in which they appeared. The financial-healthcare (1364 examples, 8956 features) dataset is drawn from the Industry-sector collection where the task is to distinguish between webpages associated with financial versus healthcare industry.

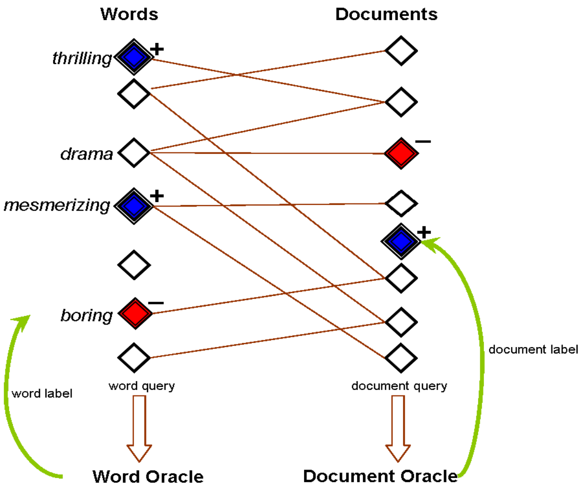

Figure 1. Graph based Active Dual Supervision

Since these datasets come with labels for all documents, some of these labels can be suppressed in order to generate a partial labeling over documents in a straightforward manner. For words, however, we do not have a gold-standard set of labels. We therefore construct a word oracle in the following manner (also see (Druck et al., 2008)). The information gain of words is computed using binary word representations with respect to the known true class labels in the training splits of a dataset. Next, out of the total vocabulary, only the top few words as ranked by information gain are assigned a label. This label is the class in which the word appears more frequently. The oracle returns a “dont know” response (0-valued label in our formulation) for the remaining words. Thus, this oracle simulates a human domain expert who is able to recognize and label the most relevant task-specific words, and also reject a word that falls below the relevance threshold. For instance, in sentiment classification, we would expect a “dont know” response for non-polar words such as “drama”. This oracle is then used to generate fixed partial word labellings, and for active querying in Section 3.

2.1. Graph-Based Dual Supervision

We now outline some technical details for our model. If a word j occurs in document i, there is an undirected weighted edge between the associated nodes of the bipartite graph, with weight X_{ij}. The adjacency matrix, $\mathbf{W}$, and the associated normalized Laplacian, $\tilde{\boldsymbol{L}}$, of this graph are given by, $\mathbf{W} = \begin{bmatrix} \mathbf{0} & \mathbf{X} \\ \mathbf{X}^T & \mathbf{0} \end{bmatrix}$, $\tilde{\boldsymbol{L}} = \begin{bmatrix} \mathbf{I} & \tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}^T & \mathbf{I} \end{bmatrix}$. Here $\tilde{\mathbf{X}}$ is a normalized version of the data matrix defined by $\tilde{\mathbf{X}} = \mathbf{D}_1^{-\frac{1}{2}}\mathbf{X}\mathbf{D}_2^{-\frac{1}{2}}$ where $\mathbf{D}_1, \mathbf{D}_2$ are diagonal matrices of appropriate size defined by $D_{1_{ii}} = \sum_{j=1}^{d} X_{ij}$ and $D_{2_{jj}} = \sum_{i=1}^{n} X_{ij}$. Let $\boldsymbol{f}_d$ denote the $n \times 1$ prediction variables over the n document vertices and $\boldsymbol{f}_w$ denote the $d \times 1$ prediction variables over d word vertices. Then, Graph transduction (see e.g., (Belkin et al., 2004)) solves the following optimization problem to find $\boldsymbol{f}^\star = (\boldsymbol{f}_d, \boldsymbol{f}_w)$, thereby completing the labeling of the bipartite graph,

$$\boldsymbol{f}^\star = \underset{\boldsymbol{f}:\sum_i f_i = 0}{\operatorname{argmin}} \ (\boldsymbol{f} - \boldsymbol{t})^T C (\boldsymbol{f} - \boldsymbol{t}) + \mu \boldsymbol{f}^\top \tilde{\boldsymbol{L}} \boldsymbol{f} \qquad (1)$$

where μ is a real-valued regularization parameter (set to its default value of 0.01 for all experiments in this paper), $\boldsymbol{t} = [\mathbf{y}^\top \ \boldsymbol{z}^\top]^\top$ is the concatenated label vector, and C is a diagonal cost matrix, $C = diag(\frac{1}{l_{\mathbf{y}}}|\mathbf{y}|; \frac{1}{l_{\boldsymbol{z}}}|\boldsymbol{z}|)$, that makes the loss terms measure average squared loss over $l_{\mathbf{y}}$ labeled documents and $l_{\boldsymbol{z}}$ labeled words separately. The solution is obtained by solving a linear system, where the high sparsity of the data matrix allows scaling upto very large datasets using iterative techniques (e.g., conjugate gradient). For convenience, we henceforth refer to this approach as **Gra**ph-based **D**ual **S**upervision, abbreviated as GRADS .

Kernel Formulation and Connection to SVD: It is well-known (Smola & Kondor, 2004) that a graph regularizer $\tilde{\boldsymbol{L}}$ can be associated with a kernel matrix $\mathbf{K} = \tilde{\boldsymbol{L}}^\dagger$, where † denotes pseudo-inverse, such that the graph transduction solution obtained from Eqn. 1 can instead be recovered from standard kernel regularized least squares training over just the labeled entities, i.e. labeled examples and labeled features, $f_i^\star = \sum_j K(i,j)\alpha_j^\star$ where $\boldsymbol{\alpha}^\star = (\mathbf{K}_{LL} + \mu C_L^{-1})\ t_L$ where L indexes labeled examples and labeled features, and t and C are the same as in the context of Eqn. 1. The kernel matrix can also be plugged into other non-linear models such as SVMs, Logistic Regression and Gaussian Processes. In practice, it may be computationally preferable to solve Eqn. 1 exploiting the sparsity of $\tilde{\boldsymbol{L}}$ instead of dealing with a dense $\mathbf{K}$ as in this formulation. Note that the kernel matrix $\mathbf{K}$ is a $(n+d) \times (n+d)$ similarity matrix *over both documents and words.* The eigen-decomposition of the pseudo-inverse of the normalized Laplacian $\tilde{\boldsymbol{L}}$ of a bipartite graph can be explicitly computed in terms of the SVD of $\tilde{\mathbf{X}}$. This is a non-trivial computation and we point the reader to Corollary 2 of (Ho & Dooren, 2005). From this calculation (not shown here), one can explicitly construct a "semantic" feature map $\psi : \mathcal{D} \cup \mathcal{W} \mapsto \mathcal{R}^{n+d}$ defined in terms of the left and right singular vectors of $\tilde{\mathbf{X}}$ that maps the set of documents $\mathcal{D}$ and the set of words $\mathcal{W}$ to points in the same euclidean space such that $K(i,j) = \psi(i)^T\psi(j)$. From this viewpoint, labeled features in this framework are simply additional labeled data points that augment the point cloud of labeled documents in this semantic feature space. Other mechanisms for euclidean embedding of co-occurrence data are naturally pertinent to this discussion (Globerson et al., 2007).

Higher-order Graph Regularizers: We point the reader to (Smola & Kondor, 2004) for typical alternative choices of graph regularizers generated by $\tilde{\boldsymbol{L}}$ and the form of smoothness they impose. In this paper, we use iterated regularized Laplacians of the form $(\tilde{\boldsymbol{L}}^p + \epsilon I)$ where p is an integer parameter and ϵI is a small ridge term with $\epsilon = 10^{-8}$. Effectively, the parameter p modulates the semantic feature map, ψ, discussed above towards the dominant singular subspace of $\mathbf{X}$. The empirical behavior of p is reported in sub-section 2.2.

Out-of-Sample Extension: It is straightforward to observe that the special structure of the bipartite Laplacian (for $p = 2$) implies the following: $\|\tilde{L}\boldsymbol{f}\|^2 = \|\boldsymbol{f}_d - \tilde{\mathbf{X}}\boldsymbol{f}_w\|^2 + \|\boldsymbol{f}_w - \tilde{\mathbf{X}}^T\boldsymbol{f}_d\|^2$. Thus, the regularization penalty Eqn. 1 enforces a least squares fit between $\tilde{\mathbf{X}}\boldsymbol{f}_w$ and the transductive predictions $\boldsymbol{f}_d$ over documents. This suggests that one can treat $\boldsymbol{f}_w$ as parameters of a linear model and make out-of-sample predictions on completely unseen test points as follows,

$$f(\boldsymbol{x}) = \boldsymbol{f}_w^\top \tilde{\boldsymbol{x}} = \boldsymbol{f}_w^\top D_2^{-\frac{1}{2}} \boldsymbol{x} / \sqrt{\sum_{i=1}^{d} x_i} \qquad (2)$$

In sub-section 2.2, we compare this inductive formula with the alternative approach of retraining the model after incorporating test data as nodes in the bipartite graph.

2.2. Learning from Labeled Features

A dual supervision model goes beyond traditional learning from labeled examples, by also incorporating feature labels explicitly. Our active learning schemes build on GRADS which is itself a new model. In this section, we address a natural empirical question: how

Table 1. Comparison with GE-FL, effect of parameter p and quality of out-of-sample prediction.

Data set	*#labels*	GE-FL	GRADS	GRADS	*#labels*	GE-FL	GRADS	GRADS
	(IG)		$p = 1$	$p = 5$	*(LDA)*		$p = 1$	$p = 5$
			in, out	in, out			in, out	in, out
movie	43.7	79.7	79.2, 79.3	80.0, 77.5	4.6	62.3	63.1, 62.5	66.5, 67.6
med-space	50.0	95.2	95.4, 94.8	95.0, 95.0	14.3	92.7	95.1, 95.8	94.8, 94.8
ibm-mac	43.7	85.5	85.9, 87.6	86.1, 84.4	10.4	81.7	80.9, 83.6	83.9, 82.3
baseball-hockey	50.0	95.4	95.0, 95.7	95.9, 94.9	10.8	91.5	90.5, 91.8	94.9, 94.2
financial-healthcare	50.0	58.3	58.8, 53.6	53.7, 42.0	9.4	58.8	54.0, 43.7	51.0, 44.5

well does GRADS perform with respect to competing alternatives for learning from labeled features? To isolate the effectiveness of feature-side supervision, we assume that no labeled examples are available. (Druck et al., 2008) report state-of-the-art performance on such tasks with their Generalized Expectation based model (abbreviated GE-FL) that outperforms several baselines (e.g., voting amongst labeled features), and is shown to be much more cost-effective than standard example-side semi-supervised learning. In Table 1, we benchmark GRADS under exactly the same experimental setting as used by (Druck et al., 2008) for feature labeling experiments: Datasets were divided into training and test splits in the ratio 3:1, and results reported are F1-measures on the test set averaged over 10 such random splits. For each training split, an information gain based feature labeling oracle is learnt (see (Druck et al., 2008) for more details on its construction). A candidate list of 50 words is then generated for labeling by this oracle. This is done in two different ways to get rough bounds on expected performance: (a) picking the top 50 words by information gain, to simulate the situation where the oracle is nearly completely known and (b) clustering documents using LDA (latent Dirichlet allocation) and then picking top 50 features. In the latter case, the unsupervised construction of the candidate list ends up eliciting many "dont-know" responses and the net number of labeled words extracted from the oracle is much fewer (e.g, only 4.6 out of 50 in the case of movies). Table 1 tabulates results under these two settings ("IG" and "LDA" setting) in the two sets of columns respectively.

We report GRADS performance under different settings: (i) Setting the parameter p to 1 versus 5 to explore the effect of higher-order graph regularization, and (ii) obtaining transductive predictions on the test data by including it as part of the bipartite graph ("in"), versus treating it as completely unseen data ("out") on which predictions are made using the out-of-sample prediction formula Eqn. 2. For all experiments in this paper, we used $\mu = 0.01$. From Table 1, we can conclude the following:

- In the IG setting GE-FL and GRADS have similar performance but GRADS outperforms GE-FL on 4 of the 5 datasets in the LDA setting.
- Higher order graph regularization (choosing $p = 5$ instead of $p = 1$) boosts performance on 3 of the 5 datasets in the case of LDA features.
- Eqn. 2 returns very high quality out-of-sample prediction as evidenced by the small performance difference between "in","out".

We conclude that GRADS is competitive with the state-of-the-art methodologies for learning from labeled features. In practice, in the initial stages of a feature-side active learning session, only a few, moderately relevant feature labels may have been extracted from the oracle. In such situations, as the LDA setting results in Table 1 indicate, the co-clustering assumptions, and the ability to control its strength via p, are expected to be beneficial aspects of GRADS .

3. Active Dual Supervision

Since GRADS is no different from a standard supervised kernel method with a particular graph-based choice of the kernel over documents and words, many well-developed intuitions around active learning can immediately be brought to bear on the dual supervision setting. In this paper, we probe two classical methodologies for active learning: uncertainty sampling and experimental design. These schemes are pool-based, i.e., they have access to both unlabeled examples and unlabeled features, and can score both dimensions with respect to their respective criterion for measuring expected gain from acquiring a label. While our core dual supervision model and the active learning wrappers around them conceptually treat the two dimensions symmetrically, we point out that feature-supervision brings new considerations to the discussion.

3.1. Classical Active Learning Schemes

We first adapt some classical schemes to our setting.

Uncertainty-based Sampling: In uncertainty-sampling, data points whose classification is most ambiguous according to the current model are chosen for labeling. For SVMs (Tong & Koller, 2001) this corresponds to querying the document that is closest to the current separating hyperplane, i.e., points where the model output is closest to 0. For the graph-based dual supervision model, we have access simultaneously to current predictions over examples and features, i.e., $\boldsymbol{f}_d$ and $\boldsymbol{f}_w$ respectively. Thus, given the current model, $\text{argmin}_i \, |\boldsymbol{f}_{d_i}|$ is the best example to label, while $\text{argmin}_j \, |\boldsymbol{f}_{\mathbf{w}_j}|$ is the best feature to label according to this heuristic.

As a traditional active learning scheme, uncertainty sampling is a standard baseline for active learning researchers. It is popular due to ease of implementation and decent empirical performance on real-world problems. It is therefore an appealing strategy for acquiring example labels. However, acquisition of feature labels with the same strategy may not necessarily lead to better models. If the model is uncertain about a feature, it may have insufficient information and may indeed benefit from learning its label. On the other hand, it is also quite likely that a feature has a low score because it does not carry much discriminative information about the classes. In such cases, a query is likely to come back with a "dont-know response". The active learner has to also ensure that queries are not wasted with such responses, for all but the most relevant words.

Empirical results presented in sub-section 4.1 show that *certainty sampling*, i.e choosing $\text{argmax}_j \, |\boldsymbol{f}_{\mathbf{w}_j}|$ for querying, is far more effective than uncertainty sampling for feature label acquisition. It is often able to pass a sizable number of relevant words to the oracle. By contrast, uncertainty sampling has a much lower hit-rate than even random sampling. In the early stages of active learning, feature certainty sampling serves to confirm or correct the orientation of model weights on different words. It is worthwhile, however, to keep in mind that while feature certainty may practically be very useful, it is not an optimal strategy either, in that queries may be wasted simply confirming confident predictions, overall providing limited benefit to the model.

Transductive Experimental Design: In the statistics community, active learning has been most throughly studied under the umbrella of classical experimental design for linear least squares models. Consider the data matrix $\mathbf{X}$ as a pool of n examples, from which a subset $\mathbf{X}_{\mathcal{A}}$ is selected for training a least squares classifier, where $\mathcal{A} \in \{1, 2 \ldots n\}$ denotes a subset of indices. Under the usual assumptions on the distribution of prediction errors (zero mean, equal variance σ^2), it is well known that the trained least squares model has an estimation error with zero mean and covariance matrix $H(\mathcal{A}) = \sigma^2(\mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}})^{-1}$. The goal of experimental design methods is to select the optimal subset $\mathcal{A}$ of a pre-determined number of data points, that "minimizes" H in some sense: D-optimal design minimizes $logdet(H(\mathcal{A}))$, E-optimal design minimizes $\|H(\mathcal{A})\|_{fro}^2$ and A-optimal design minimizes $trace(H(\mathcal{A}))$. (Yu et al., 2006) discuss the notion of *Transductive* Experimental design which instead focuses on minimizing the variance in predictions over a given set of unlabeled examples, which is of more direct interest. This method often outperforms uncertainty sampling in traditional active learning tasks. Moreover, (Yu et al., 2006) outline a kernelized version which can be immediately applied to our active dual supervision setting.

Given the $(n+d) \times (n+d)$ bipartite graph kernel $\mathbf{K}$, let $E = \{1, \ldots, n\}$ denote the indices of documents and $F = \{n+1, \ldots, n+d\}$ denote the indices of words. Suppose we picked a subset of examples and features indexed by $\mathcal{A} \subset E \cup F$, trained a (kernelized) regularized least squares model (with regularization parameter μ) and made predictions on the document collection. Then, the predictive error has a covariance matrix, $H(\mathcal{A}) = \frac{1}{\mu}\left[\mathbf{K}_{EE} - \mathbf{K}_{E\mathcal{A}}(K_{\mathcal{A}\mathcal{A}} + \mu I)^{-1}\mathbf{K}_{\mathcal{A}E}\right]$. To select features for labeling based on the transductive experiment design criterion, we minimize $trace(H(\mathcal{A}))$ over $\mathcal{A} \subset F$, while for selecting examples, the minimization is done over $\mathcal{A} \subset E$. We used a simple matrix deflation procedure proposed in (Yu et al., 2006) to greedily minimize $H(\mathcal{A})$ over appropriate sets. Intuitively, $H(\mathcal{A})$ measures how well the span of documents and/or words in $\mathcal{A}$ represents the entire document collection in the semantic feature space associated with the kernel $\mathbf{K}$. It is worthwhile to note that the objective function, $H(\mathcal{A})$, is *independent* of the labels of $\mathbf{X}_{\mathcal{A}}$. Hence, unlike uncertainty-based schemes that score candidates based on the current model, examples and features in this case can be ranked for querying prior to any training. Note that on some tasks, however, label dependence may be desirable.

3.2. Probabilistic Interleaving

An ideal active dual supervision scheme should be able to gauge the value of acquiring labels for examples and features on the same scale. In addition, it should be able to incorporate differences in domaindependent acquisition costs and asymptotic limits of oracle knowledge for the two different forms of supervision. In our uncertainty-inspired scheme, we use different scores: uncertainty for examples and certainty for features respectively. Similarly, in transductive experimental design, the objective function measures the net predictive variance in predictions over the set of examples not including features. Thus, these scores are not on the same scale. Moreover, in their original form, these schemes are also not cost-sensitive. One way to attempt to align scalings and incorporate costs is by estimating the cost-sensitive expected improvement in the model accuracy for each possible outcome of each possible feature and example label acquisition. Such an *expected utility* approach has been successfully applied to traditional active learning in selecting examples (Saar-Tsechansky et al., 2009). However, applying such an approach to GRADS is computationally very intense, and is fraught with difficult estimations and approximations necessary for the computation of expected utility, particularly in the absence of labeled validation data.

Instead, in this paper, we experiment with fast and simple interleaving schemes where the active learner probabilistically queries the example oracle or the word oracle based on an interleave probability. These schemes allow us to clearly demonstrate the value of active dual supervision, which is the primary contribution of this paper. We benchmark two such schemes: (1) interleaving example uncertainty and feature certainty and (2) interleaving examples and features ranked by transductive experimental design. We study accuracy versus cost tradeoff as a function of the interleave probability. Results in section 4.2 confirm that active dual supervision can often be much more cost effective than one-sided active learning.

4. Empirical Study

We begin our study with feature-side active learning since it impacts choices for active dual supervision, and has not been explored before. In sub-section 4.1, we compare the relative performance of various schemes for feature label acquisitions. In sub-section 4.2, we explore the performance of probabilistic interleaving schemes for active dual supervision.

The experimental setting is different from the setup in sub-section 2.2. To make it computationally feasible to run variance-based active learning schemes, we selected the top 1500 features in each dataset using document frequency, and used the tfidf feature vector representation. Results are averaged over 10 random training-test splits as before. For each split, the feature-label oracle was set to label the top 100 most relevant words based on information gain computed on the training set. For each run, all methods are initialized by a random choice of 6 feature labels. Test sets were completely held out and predictions on it were made using Eqn 2. The active learning session comprised of a fixed total number of queries – 200 queries for all datasets except movies where we experimented with 400. During this session, the accuracy of GRADS on the the test set was monitored after each label acquisition. We used $p = 1$ for med-space and financial-healthcare, and $p = 5$ for other datasets based on results from sub-section 2.2, without any additional tuning.

4.1. Feature-side Active Learning

Figure 4.1 shows active learning curves for uncertainty, certainty, random, and variance-based (transductive experimental design) schemes on 4 datasets (we omit med-space for lack of space; it behaves similar to other 20-newsgroups datasets.). We also show the performance of the model when all the oracle words are fully revealed in the order of information gain. The learning curves marked "oracle" in Figure 4.1 therefore give an upper-bound on the expected performance.

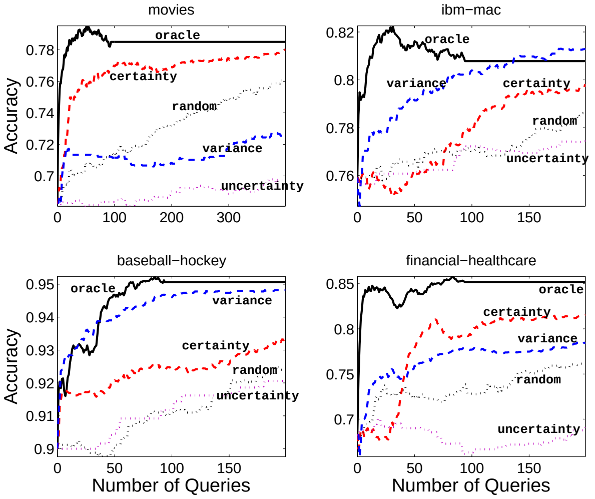

Figure 2. Feature-side Active Learning

We make the following observations.

- Feature-side active learning is generally very effective. On most datasets, by the end of one or the other form of feature-side active learning session, GRADS accuracy comes within 1% of what it can potentially achieve with all oracle word labels.
- Feature uncertainty performs significantly worse

than feature certainty, and very often worse than random sampling. This is because uncertainty sampling is strongly biased towards extracting dont-know responses from the oracle. Table 2 shows the percentage of oracle words retrieved by the end of the session. It is clear that uncertainty tends to extract fewer word labels than random sampling. As Table 2 shows, certainty sampling has very high oracle response rates. Thus, this aspect significantly distinguishes feature-side active learning from traditional example-side active learning. These results support the idea of combining example uncertainty with feature certainty for active dual supervision.

Table 2. Percentage of total oracle words retrieved for random (rand), variance-based (var), uncertainty (unc) and certainty (cert) sampling respectively in 200 queries (400 for movies).

Dataset	rand	var	unc	cert
movies	25.2	13.6	6.9	81.0
ibm-mac	13.2	26.3	2.1	66.6
baseball-hockey	13.5	28.1	1.2	58.2
med-space	11.5	35.3	1.3	69.6
financial-healthcare	13.1	34.3	4.2	60.8

- Querying features based on predictive variance (transductive experimental design) is surprisingly effective on 20-newsgroups datasets. Even though this scheme has lower oracle response rates than certainty sampling (see Table 2), the quality of words picked for labeling appears to be significantly better on these datasets. To the best of our knowledge, this is the first demonstration of the utility of experiment design notions for acquiring feature-side "experiments". On the other hand, on movies the variance scheme performs worse than random. We believe that this is due to the label-independent aspect of transductive experimental design. Many non-polar topical words (e.g., "action", "drama") are picked that do not associate strongly with sentiment classes.

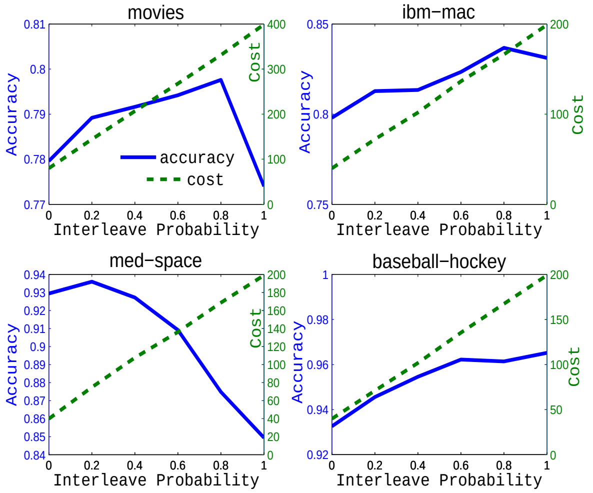

Figure 3. Active dual supervision: Accuracy (solid blue line with y-axis on the left) and Cost (dashed green line with y-axis on the right) as a function of interleave probability.

4.2. Experiments in Active Dual Supervision

In this section, we benchmark active dual supervision schemes based on probabilistic interleaving between example uncertainty and feature certainty. Similar observations (not reported here for lack of space) hold for interleaving examples and features based on transductive experimental design. The interleave probability is varied from 0 to 1 spanning the range from pure feature-based active learning (based on feature certainty scores) to pure example-based active learning (based on example uncertainty scores). We assume that these two forms of supervision incur different costs. For simplicity, we use the cost model suggested in (Raghavan et al., 2007; Druck et al., 2008) where features are roughly 5 times cheaper (i.e., faster to label) than examples. Figure 3 shows, as a function of interleave probability, the total cost incurred (right y-axis) in acquiring labels, assuming unit cost for example labels and 0.2 for feature labels, and the resulting accuracy (left y-axis) returned by GRADS when the active learning session culminates. We omit the plot for financial-healthcare which qualitatively behaves similar to med-space. From Figure 3, it is clear that, in general, the best performing model is obtained by allowing the active learner to be able to query for both forms of supervision. For example, in movies, the accuracy peaks at an interleave probability of 0.8 producing a model better than the one returned by example-only active learning (i.e., when interleave probability = 1.0) or by feature-only active learning (i.e., when interleave probability = 0.0). The effectiveness of interleaving and the optimal interleave probability depends on the dataset. On some datasets (med-space, financial-healthcare) the situation is ideal: pure feature-side active learning returns the best model which also happens to be the cheapest. On other datasets (ibm-mac, baseball-hockey) the cost burden of example-side active learning with GRADS can be significantly lessened by interleaving with feature label requests, yielding an equally accurate model. For example, on baseball-hockey, interleaving feature label requests 40% of the time (i.e., interleave probability 0.6) gives a model – at a cost of around 130 – that is as accurate as the one where all queries are example-label requests (interleave probability 1.0) costing 200 instead. For the same cost range, standard uncertainty sampling with a standard supervised model (regularized least squares classifier) gives accuracies ranging from 0.664 to 0.726 for movies, 0.778 to 0.839 for ibm-mac, 0.886 to 0.947 for med-space, and 0.89 to 0.942 for baseball-hockey. Overall, as Figure 3 shows, active dual supervision with GRADS provides a substantially better accuracy-cost tradeoff. Collectively, these results confirm the potential of active dual supervision.

5. Related Work

Dual supervision is a relatively new area of research. Our methods are motivated by (Sindhwani et al., 2008). In this paper, we compared our base model with that of Druck et al. (2008) who apply a Generalized Expectation criteria to learn a multinomial logistic regression model from labeled features. See references in these papers on prior work on incorporating labeled features via labeled *pseudo-examples*. Recently, (Melville et al., 2009) proposed a classifier for dual supervision based on Pooling Multinomials. Active learning with this classifier has also very recently been explored in (Melville & Sindhwani, 2009) although to the best of our knowledge, there is very little prior work (Raghavan et al., 2007; Godbole et al., 2004) on active learning with feature supervision. In particular, notions from classical experiment design have never previously been explored in this context.

6. Conclusion

In this paper, we developed active learning schemes for dual supervision. Extensive empirical results demonstrate how feature-side active learning and the simultaneous selection of features and examples as queries can reduce the cost of building a high-accuracy model, more effectively than classical one-sided active learning. Developing novel multi-dimensional active learning schemes that can seamlessly interleave between different forms of supervision, take associated label acquisition costs, annotation quality and other oracle properties into account, is a broad topic for future research.

Acknowledgements

We thank Yan Liu, Alexandru Niculescu-Mizil and Kai Yu for helpful conversations. We thank Gregory Druck for providing datasets for comparison with the GE-FL approach.

References

Belkin, M., Matveeva, I., & Niyogi, P. (2004). Regularization and semi-supervised learning on large graphs. *Conference on Learning Theory (COLT)* (pp. 486–500).

Druck, G., Mann, G., & McCallum, A. (2008). Learning from labeled features using generalized expectation criteria. *31st Annual ACM SIGIR Conference* (pp. 595–602).

Globerson, A., Chechik, G., Pereira, F., & Tishby, N. (2007). Euclidean embedding of co-occurence data. *Journal of Machine Learning Research*, *8*, 2265–2296.

Godbole, S., Harpale, A., Sarawagi, S., & Chakrabarti, S. (2004). Document classification through interactive supervision of document and term labels. *Prin. and Prac. of Knowl. Disc. in Databases (PKDD)* (pp. 185–196).

Ho, & Dooren, P. (2005). On the pseudo-inverse of the laplacian of a bipartite graph. *Appl. Math. Letters*, *8*, 917–922.

Melville, P., Gryc, W., & Lawrence, R. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. *15th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining.*

Melville, P., & Sindhwani, V. (2009). Active dual supervision: Reducing the cost of annotating examples and features. *NAACL HLT Workshop on Active Learning for NLP.*

Raghavan, H., Madani, O., & Jones, R. (2007). An interactive algorithm for asking and incorporating feature feedback into support vector machines. *30th Annual ACM SIGIR Conference* (pp. 79–86).

Saar-Tsechansky, M., Melville, P., & Provost, F. (2009). Active feature-value acquisition. *Management Science*, *4*, 664–684.

Sindhwani, V., Hu, J., & Mojsilovic, A. (2008). Regularized co-clustering with dual supervision. *Neural Information Processing Systems (NIPS)* (pp. 976–983).

Smola, A., & Kondor, R. (2004). Kernels and regularization on graphs. *Conf. on Learning Theory (COLT)* (pp. 144–158).

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, *2*, 45–66.

Yu, K., Bi, J., & Tresp, V. (2006). Active learning via transductive experimental design. *International Conference on Machine Learning (ICML)* (pp. 1081–1088).