Simple, Robust, Scalable Semi-supervised Learning via Expectation Regularization

Gideon S. Mann GIDEON.MANN@GMAIL.COM
Andrew McCallum MCCALLUM@CS.UMASS.EDU
Department of Computer Science, University of Massachusetts, Amherst, MA 01003

Abstract

Although semi-supervised learning has been an active area of research, its use in deployed applications is still relatively rare because the methods are often difficult to implement, fragile in tuning, or lacking in scalability. This paper presents *expectation regularization*, a semi-supervised learning method for exponential family parametric models that augments the traditional conditional label-likelihood objective function with an additional term that encourages model predictions on unlabeled data to match certain expectations—such as label priors. The method is extremely easy to implement, scales as well as logistic regression, and can handle non-independent features. We present experiments on five different data sets, showing accuracy improvements over other semi-supervised methods.

1. Introduction

Research in semi-supervised learning has yielded many publications over the past ten years, but there are surprisingly fewer cases of its use in application-oriented research, where the emphasis is on solving a task, not on exploring a new semi-supervised method. This may be partially due to the natural time it takes for new machine learning ideas to propagate to practitioners. We believe it is also due in large part to the complexity and unreliability of many existing semi-supervised methods.

The goal of our work here is to propose a simple semi-supervised learning method that consistently provides accuracy improvements, that is robust across many problem domains without meta-parameter tuning, and scalable to extremely large unlabeled data set sizes.

This paper presents *expectation regularization* (XR), a new method for semi-supervised learning with exponential-family parametric models. Many exponential-family models such as logistic regression and multi-class maximum entropy classifiers are optimized by maximizing the conditional log-likelihood of the true labels given the input features. XR augments this objective function by adding a second term that encourages model predictions on unlabeled data to match certain designer-provided expectations. In particular, the XR term minimizes the KL-divergence between feature/label expectations predicted by the model and human-provided feature/label expectation priors.

In this paper we empirically explore one important special case termed *label regularization*, in which the human provides a label prior distribution, and the XR term encourages the optimization procedure to find parameters that predict a similar label distribution on the unlabeled examples. (Intuitively one can see that this prevents a typical failure case of several alternative semi-supervised methods, in which the learned model predicts the same label for almost all inputs.) Appropriate label distributions are often easily provided by human prior knowledge; alternatively they can be obtained from the limited labeled data, from which they can be estimated far more accurately than sparse input feature distributions. We show below that XR is surprisingly robust to inaccuracies in the provided label distribution prior.

Expectation regularization offers a number of practical advantages over previous semi-supervised learning methods. It is simple to implement and to use—requiring no pre-clustering of unlabeled data, no inverted index for graph construction, no "auxiliary functions" and no "contrastive" examples. It has two meta-parameter terms, both of which require little or

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007. Copyright 2007 by the author(s)/owner(s).

no tuning and are not overly sensitive. It is purely conditional on inputs, and thus can robustly handle arbitrarily overlapping, non-independent feature sets. It is a parametric model, and thus it can be applied quickly to new instances without requiring the storage large quantities of the labeled and unlabeled training data. Not only can XR perform well with many labeled examples, but unlike other methods it also excels at very small levels of labeled data (as little as one per class). Significantly, it scales up to vast numbers of unlabeled points (easily millions). It is quite robust; in our experiments it provided consistent accuracy improvements.

We present experimental results on five different data sets, and compare against seven different alternative supervised and semi-supervised methods. Across the data sets XR outperforms naïve Bayes, SVMs, EM, maximum entropy, entropy regularization (serving also as a stand-in for transductive SVMs), cluster kernels, as well as a graph-based method. The only times when XR under-performs an existing method is (a) a radial-basis-function SVM in the case of large amounts of labeled data, and (b) naïve Bayes EM on a simple, extremely sparse data set, where naïve Bayes outperforms maximum entropy. We also demonstrate robustness to error in prior estimation and across meta-parameter settings.

In future work we will experiment with expectations on features other than labels, and will also apply these methods to structured models, such as conditional random fields (Lafferty et al., 2001; Sutton & McCallum, 2006), which are a natural fit for XR.

2. Related Work

There have been many different approaches to semi-supervised learning over the past decade that have shown various accuracy improvements. Here we discuss some of the most popular methods: generative-models with EM, other "cluster-based" methods, auxiliary-function methods, and graph-based methods.

Generative models trained by expectation maximization (Dempster et al., 1977) have had a long history in semi-supervised machine learning. Nigam et al. (1998) present a semi-supervised naïve Bayes model for text classification, and this method has also been applied to structured classification problems such as part-of-speech tagging (Klein & Manning, 2004). However, while EM sometimes works very well, it can be fragile, finding solutions that are worse than the equivalent supervised model. Cozman and Cohen (2006) discuss the risks of using EM and describe situations where it can fail.

Other "cluster-based" methods are discriminative, directly aiming to place the decision boundary in low-density regions. For example transductive support vector machines (TSVMs) (Joachims, 1999) explicitly model the distance between classes by simultaneously searching over labelings of unlabeled/test instances and margins between regions of similarly-labeled instances. This search can be expensive, and TSVMs have difficulty handling large number of unlabeled instances, with running time $O(n^3)$ as originally described; although Sindhwani and Keerthi (2006) propose a method for speeding up training in some cases. Furthermore, in our experience, TSVMs require extensive and delicate tuning of meta-parameters. We note that Sindhwani and Keerthi report results with meta-parameters tuned on test data.

Another cluster-based method with significantly faster training times is entropy regularization (Grandvalet & Bengio, 2004). Here a traditional conditional label likelihood objective function is augmented with a second term that minimizes the entropy of the label distribution predicted on unlabeled data. Chapelle et al. (2006) give empirical evidence that entropy minimization performs as well as (if not better than) TSVMs, (when the SVM is given a linear kernel). However entropy regularization also requires extremely sensitive tuning of the relative weight between the two terms. Furthermore, when faced with small amounts of labeled data and vast amounts of unlabeled data, entropy minimization is unstable, preferring solutions where all points are assigned the same label. (We note that our label regularization can easily be combined with entropy regularization to avoid this problem.) Another fast cluster-based method is information regularization (Corduneanu & Jaakkola, 2003), which measures distance via the mutual information between a classifier and the marginal distribution $p(x)$.In general, if the cluster assumption is violated (*i.e.* the classes are not widely separable) assigning decision boundaries to low density regions is a poor choice.

Instead of using data clustering directly to position the decision boundary, other methods pre-cluster unlabeled data, and use these clusters as features for supervised training on the labeled data (Miller et al., 2004; Li & McCallum, 2005). These methods can work well when natural unsupervised clusterings are correlated with the supervised task, and when the amount of labeled data is not too small. Auxiliary-task methods (Ando & Zhang, 2005) embed the cluster-discovery into supervised training; contrastive methods (Smith

& Eisner, 2005) perturb the input space. Although these methods have been demonstrated to produce impressive gains, both are quite sensitive to the selection of auxiliary information, and making good selections requires significant insight.[1]

Graph-based methods, also known as manifold methods, have been widely applied to semi-supervised learning, and can be highly accurate. Here a graph (typically with weighted edges) is formed over the labeled and unlabeled points, and points are assigned labels based on the labels of their neighbors. Zhu and Ghahramani (2002) propose label propagation, where labels propagate from labeled instances to unlabeled instances. Szummer and Jaakkola (2002) present a closely related approach which uses random walks through the graph to assign labels. Li and McCallum (2004) examine simultaneous pair-wise distance and classification boundaries, which produces an implicit clustering over points. However, like TSVMs, graph-based methods are slow, requiring time $O(n^3)$ or $O(kn^2)$ where k is the number of neighbors. They also are not compact parametric models—they require that labeled and unlabeled data be stored and used to classify new instances. Sub-sampling unlabeled data can reduce runtime from $O(n^3)$ to $O(m^2n)$ (Delalleau et al., 2006), but subsampling does not take full advantage of available unlabeled data. Other techniques for speeding up training can reduce the time complexity to $O(m^3), m < n$, but may reduce performance (Zhu & Lafferty, 2005). In this paper we compare against a representative graph-based label propagation method called Quadratic Cost Criterion (QC) (Bengio et al., 2006) whose results are reported in Chapelle et al. (2006).

Some semi-supervised learning methods other than our expectation regularization have also used label prior distributions, but in quite different ways. For example, class mean normalization (CMN) (Zhu et al., 2003) employs class priors as a post-processing step to set thresholds on the propagation of a label. Conditional harmonic mixing (Burges & Platt, 2006) is another graph-based method that minimizes over each point the KL-divergence between the currently predicted label distribution and the distribution predicted by its neighbors. Schapire et al. (2002) use a human-generated prior on *model parameters* and minimize the *per-instance* KL-divergence between the label distribution predicted by the prior model and that predicted by the learned model. Schuurmans (1997) uses predicted label distributions on unlabeled data for model structure selection (as opposed to parameter estimation).

There are, of course, cases of semi-supervised learning being used in application settings, however, often with various difficulties. For example, Macskassy and Provost (2006) apply harmonic mixing to classification in relational data, but complain about running time and prefer a simpler method. Niu et al. (2005) apply label propagation to word sense disambiguation, and show that performance is sensitive to choice of metric for constructing graph. Merialdo (1994), in a now famous negative result, attempts semi-supervised learning to improve HMM part-of-speech tagging and finds that EM with unlabeled data reduces accuracy. Klein and Manning (2004) show that with very clever initialization, however, EM can help. Kockelkorn et al. (2003) use transductive SVMs for text classification, but complain that it is computationally costly.

3. Expectation Regularization

Many of the methods discussed above use knowledge of the marginal $p(x)$ either explicitly (Corduneanu & Jaakkola, 2003) or implicitly (Grandvalet & Bengio, 2004) in deciding where to place decision boundaries. Given knowledge of the marginal, these methods formulate regularization criteria which favor decision boundaries that are placed in areas of low density.

Expectation regularization uses an additional source of knowledge: beliefs about the conditional probabilities of labels given features, $\tilde{p}(y|x_j)$. These expectations can be obtained through various means, either from estimation on labeled data or through human prior knowledge. This type of information constitutes a new modality of supervision, where instead of labeled examples, the user provides beliefs about selected conditional probabilities.

Domain knowledge can be supplied to the classifier in a flexible way using expectation regularization. In many domains, class priors, $p(y)$, are a valuable source of information that are often approximately known to the classifier designer. For example, in university web page classification, one might estimate that roughly 60% of the personal home pages belong to students. In other cases, we may have expectations about the relationships between features and labels. For example, in the named-entity recognition, we may estimate that in newswire text 50% of capitalized words are named entities. In gene name tagging, there may be a 75% probability that a word is a gene if it ends with the morpheme "gene." Classifier designers traditionally employ features that they know are correlated to labels. With expectation regularization the classifier

[1] Personal communication, F. Pereira

designers can also supply estimated feature/label expectations. (Experimental results below show that our method is surprisingly robust to a wide range of errors in these estimates.)

Given these expectations, we introduce a regularizer that penalizes classifiers whose conditional probabilities $p_\theta(y|x_j)$ on unlabeled data deviate from the human-provided expectations $\tilde{p}$.

Consider a set of unlabeled data $U = \langle u_1..u_n \rangle$, where each data instance u comprises a feature vector $x^{(u)} = \langle x_1^{(u)}..x_n^{(u)} \rangle$. Since we do not have access to the complete marginal $p(x)$, we use the unlabeled empirical distribution $\hat{p}(x)$ to compute the conditional probabilities $\hat{p}_\theta(y|x_j = 1)$

$$\begin{aligned} \hat{p}_\theta = \hat{p}_\theta(y|x_j = 1) &= \sum_{x_{-j}} \hat{p}(x_{-j}|x_j = 1) p_\theta(y|x_j = 1, x_{-j}) \\ &= \frac{1}{|U_j|} \sum_{x \in U_j} p_\theta(y|x)\,, \end{aligned}$$

where U_j is defined to be $\{x \in U : x_j = 1\}$. Here, the notation x_{-j} is used to indicate $\{x \setminus x_j\}$ (all features apart from x_j). The expectation regularization term added to the objective function is

$$\Delta(\tilde{p}, \hat{p}_\theta)\,,$$

where $\tilde{p}$ is the human-provided conditional probability and $\hat{p}_\theta$ is the model's expected conditional probability, and Δ is a distance metric. In this paper, we explore one particular choice of distance metric: KL-divergence. This choice of Δ is equivalent to augmenting the likelihood with a Dirichlet prior over expectations where values for the priors α are proportional to $\tilde{p}$. KL-divergence can be factored into two parts

$$\begin{aligned} \Delta(\tilde{p}, \hat{p}_\theta) = D(\tilde{p}||\hat{p}_\theta) &= \sum_y \tilde{p} \log \frac{\tilde{p}}{\hat{p}_\theta} \\ &= -\sum_y \tilde{p} \log \hat{p}_\theta + \sum_y \tilde{p} \log \tilde{p} \\ &= H(\tilde{p}, \hat{p}_\theta) - H(\tilde{p})\,. \end{aligned}$$

Since $H(\tilde{p})$ is constant with respect to the model parameters, minimizing the KL-divergence can also be seen as minimizing the cross entropy of a hypothesized distribution and the expected distribution on the unlabeled data, $H(\tilde{p}, \hat{p}_\theta)$. Note that this is distinct from the traditional log-likelihood. The log-likelihood is equivalent to the cross entropy over instances where for each instance only the correct label has non-zero probability. In this regularization term, $\tilde{p}$ and $\hat{p}_\theta$ are the expected distributions averaged over all instances.

We apply expectation regularization to conditionally trained log-linear maximum entropy models, which are also known as multinomial logistic regression models. In these models, the probability of the class label y for a data instance x is calculated by

$$p_\theta(y|x) = \frac{1}{Z(x)} \exp\left(\sum_k \theta_k x_k\right),$$

where $Z(x) = \sum_y \exp(\sum_k \theta_k x_k)$ is the partition function. Given training data $D = \langle d_1..d_n \rangle$, the model is trained by maximizing the log-likelihood of the labels

$$\ell(\theta; D) = \sum_d \log p_\theta(y^{(d)}|x^{(d)})\,.$$

This can be done by gradient methods (Malouf, 2002), where the gradient of the likelihood is

$$\frac{\partial}{\partial \theta_k} \ell(\theta; D) = \sum_d x_k^{(d)} - \sum_d \sum_y p_\theta(y|x^{(d)}) x_k^{(d)}\,.$$

For semi-supervised discriminative training, we augment the objective function by adding regularization terms on the unannotated data. (Here Gaussian prior is also shown.)

$$\ell(\theta; D, U) = \sum_d \log p_\theta(y^{(d)}|x^{(d)}) - \frac{\sum_k \theta_k}{2\sigma^2} - \lambda \Delta(\tilde{p}, \hat{p}_\theta)\,.$$

In practice, we find that λ does not need tuning for each data set. We set it simply to $\lambda = 10 \times$ # labeled examples.

As an important special case of expectation regularization, we examine *label regularization*, in which the features in question are the "default features," where $\forall x : x_j = 1$. In this case, the goal of the regularizer is to match the prior distribution on labels. Note that this useful special case is not available to Schapire et al. (2002) because expectation regularization is a global regularizer as opposed to a local regularizer. If the model exactly matched the label expectation on a per-instance basis, in application it would assign all instances to the majority class.

3.1. Expectation Regularization Gradient

This section presents the gradient for KL-divergence based expectation regularization. First, we define the unnormalized potential

$$\hat{q}_\theta = \hat{q}_\theta(y|x_j = 1) = \sum_{x \in U_j} p_\theta(y|x)\,.$$

After dropping terms in $\frac{\partial}{\partial \theta_k} D(\tilde{p}||\hat{p}_\theta)$ which are constant with respect to the partial derivative, we are left with

$$\begin{aligned}\frac{\partial}{\partial\theta_k}\sum_y \tilde{p}\log\hat{q}_\theta &= \sum_y \frac{\tilde{p}}{\hat{q}_\theta}\sum_{x\in U_j}\frac{\partial}{\partial\theta_k}p_\theta(y|x)\\ &= \sum_y \frac{\tilde{p}}{\hat{q}_\theta}\sum_{x\in U_j}p_\theta(y|x)\left(x_k - \sum_{y'}p_\theta(y'|x)x_k\right)\\ &= \sum_{x\in U_j}\sum_y p_\theta(y|x)x_k\frac{\tilde{p}}{\hat{q}_\theta}\\ &\quad - \sum_{x\in U_j}\sum_{y'}p_\theta(y'|x)x_k\sum_y\frac{\tilde{p}\times p(y|x)}{\hat{q}_\theta}\\ &= \sum_{x\in U_j}\sum_y p_\theta(y|x)x_k\\ &\qquad \times\left(\frac{\tilde{p}}{\hat{q}_\theta} - \sum_{y'}\frac{\tilde{p}\times p_\theta(y'|x)}{\hat{q}_\theta}\right).\end{aligned}$$

When $\tilde{p} \propto \hat{q}_\theta$ (the expected unlabeled distribution matches the labeled distribution) the gradient is 0. This conforms to the intuition behind the development of the regularizer.

3.2. Temperature

Label regularization can occasionally find a degenerate solution where, rather than the expectation of all instances matching the prior distribution, instead, the distribution over labels for *each instance* will match the given distribution on every example. For example, given a three class classification task, if the labeled class distribution $\tilde{p}(y) = \{.5, .35, .15\}$, it will find a solution such that $p_\theta(y) = \{.5, .35, .15\}$ for every instance. As a result, all the test instances will be assigned the same label.

One solution, appealing to 0/1 loss, would be to simply measure and match the expectation over winning class counts, calculating $\hat{p}_\theta$ as $\frac{1}{U_j}\sum_{x\in U_j}\delta(y, \arg\max_{y'} p_\theta(y'|x))$. However, this is not differentiable. So instead, we make $p_\theta(y|x)$ more peaked using a temperature less than 1.

$$p_\theta(y|x) \propto \exp\left(\frac{1}{T}\sum_k \theta_k x_k\right).$$

This is differentiable and thus amenable to many gradient ascent methods. In practice we find that this meta-parameter does not require fine-tuning. Across all data sets we simply use $T = 0.1$ for multi-class problems and $T = 1$ for binary classification problems, and we find this to work well.

4. Experimental Results

We evaluate on five different data sets, and compare against seven different methods (both supervised and semi-supervised). We experiment with varied amounts of data, from one instances per class up to thousands of instances. We also examine the effect of noise on the label priors and present results which support the robustness of the method with respect to varied λ and temperature.

Name	# Points	# features	# classes
SRAA	40k	77,494	4
POS	40k	11,520	44
SecStr	83k	314 (45,436)	2
BIOII	200k	54,958	3
CoNLL03	200k	114,264	9

Table 1. The data sets are complex: they have dramatic class skews, highly inter-dependent features, and large amounts of data. The SecStr data set has 315 atomic features, and 45k features when pairwise feature conjunctions are used.

4.1. Experimental Set-up

Text classification has been a major target of semi-supervised approaches, (Nigam et al., 2006), and we evaluate on the simulated/real auto/aviation (**SRAA**) task. We examine three especially difficult natural language processing tasks: the CoNLL03 named-entity recognition task (**CoNLL03**), Part of speech tagging of the Wall Street Journal (**POS**), and the 2006 BiocreativeII evaluation (**BIOII**), using a sliding window classifier. Finally, we examine a protein secondary structure prediction task (**SecStr**), as extensively evaluated in Chapelle et al. (2006). Table 1 shows characteristics of the various data sets. The tasks are very large in scale, with up to hundreds of thousands of instance and features. They have complex characteristics such as heavily inter-dependent features and highly skewed class distributions.

Across all of the experiments we compare with supervised naïve Bayes and maximum entropy models, and semi-supervised naïve Bayes trained with EM and maximum entropy models trained with entropy regularization. For the tasks where there may be more features per instance than others, we used document length normalization for the naïve Bayes approaches which we have found to sometime significantly improve accuracy. On the secondary structure prediction we additionally compare with a supervised SVM using a radial-basis function (RBF) kernel, a Cluster Kernel (Weston et al., 2006) and a graph based-method, the Quadratic Cost Criterion with Class Mean Normalization (Bengio et al., 2006) trained using various data sub-sampling schemes (Delalleau et al., 2006): a random sampler and two smarter variations.

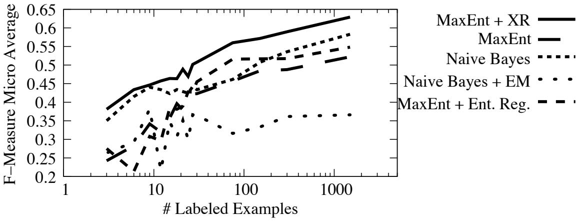

Figure 1. **BIOII**: Label regularization (XR) outperforms all other methods. The x-axis represents increasing numbers of labeled data instances. The y-axis is the F-measure micro average across all classes.

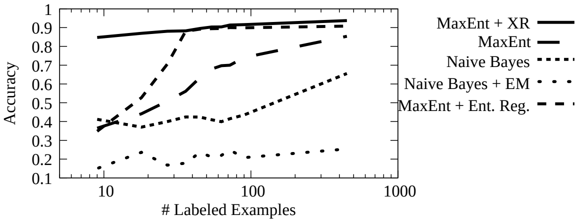

Figure 2. **CoNLL03**: Label regularization (XR) outperforms all other methods. The x-axis represents increasing numbers of labeled instances per class, and the x-axis is accuracy.

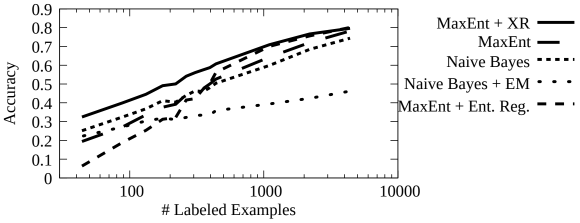

Figure 3. **POS**: Label regularization (XR) outperforms all other methods, though performance improvements over supervised maximum entropy methods appear to level off at 1300 labeled instances.

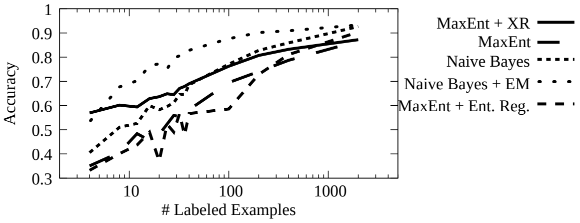

Figure 4. **SRAA**: Label regularization (XR) outperforms its supervised maximum entropy counterpart and entropy regularization and is the winner at one labeled instance per class. After that, naïve Bayes EM is the clear winner.

For **CoNLL03**, **POS**, **BIOII**, and **SRAA**, we run ten trials, splitting the data randomly into two sections, training and test. From the training set, we randomly chose some instances to be labeled and cause the rest to be hidden. We then report results on the test data (in what is commonly called inductive learning). For **SecStr** we use the labeled/unlabeled splits provided by Chapelle et al. (2006) and evaluate on the hidden training data (in what is commonly called transductive learning). In order to provide a somewhat more fair comparison with the RBF kernels used by the other methods on this task, the feature set used by the maximum entropy model and naïve Bayes models is augmented by pairwise feature conjunctions, corresponding to a quadratic kernel.

For the maximum entropy model trained with entropy regularization, after some experimentation, we weighted its contribution to the objective function with λ = # labeled data points / # unlabeled data points.
For the experiments, we use the true label priors estimated from data, corresponding to a use-case where a user gives this knowledge to the system during training. Section 4.3 presents experiments showing robustness to noisy label priors. Across the experiments, we observed that label regularization trains in time linear in the amount of unlabeled data.

4.2. Learning Curves

Figures 1, 2, 3, and 4 show classifier performance as greater amounts of labeled data is added. In **POS**, **BIOII**, and **CoNLL03**, label regularization yields significant benefits over the alternative approaches for all amounts of training data. On **SRAA**, label regularization also shows a benefit over the fully supervised maximum entropy model but its accuracy is not as high as that obtained by the EM-trained naïve Bayes learner.[2] At one instance per class, label regularization is unbeaten and yields improvement when compared to all other approaches considered. Across the experiments, as the tasks become more complicated, with larger feature sets and more unlabeled data, the label regularizer provides increasingly higher accuracy than EM and entropy regularization.

In **SecStr**, label regularization outperforms the other methods at 100 labeled points, and approaches the cluster kernel method on 1000 points. At only 2 labeled data points, it outperforms the *supervised* SVM and maximum entropy model when they are trained with 100 labeled points. In these experiments QC is not run over the complete data (presumably because of

[2] Note here that the baseline performance of the maximum entropy model is much lower than the naïve Bayes model, so that label regularization starts off at a considerable deficit.

	# Labeled Instances		
	2	100	1000
SVM (supervised)		55.41	**66.29**
Cluster Kernel		57.05	65.97
QC randsub (CMN)		57.68	59.16
QC smartonly (CMN)		57.86	59.29
QC smartsub (CMN)		57.74	59.16
Naive Bayes (supervised)	52.42	57.12	64.47
Naive Bayes EM	50.79	57.34	57.60
MaxEnt (supervised)	52.42	56.74	65.43
MaxEnt + Ent. Min.	48.56	54.45	58.28
MaxEnt + XR	**57.08**	**58.51**	65.44

Table 2. Label regularization outperforms other semi-supervised learning methods at 100 labeled data points. At one instance per class, its performance is better than the *supervised* SVM and maximum entropy model at 100.

scalability problems), but operates on a subset, either selected randomly (randsub) or in a smarter fashion (smartonly and smartsub), while the label regularization method uses the complete data. As in the other experiments, label regularization only helps accuracy, while in many of the other methods (EM, entropy regularization, cluster kernels) unlabeled data degrade performance.

We have tried additional experiments combining label regularization and entropy regularization and in most cases, it does not lead to improvements over label regularization alone and sometimes decreases the accuracy of label regularization. The two exceptions are on the **SRAA** and the **SecStr** data sets. Notably, on **SecStr**, combined entropy regularization and label regularization yields a performance of 66.30—matching the performance of the supervised radial-basis SVM and beating all other unsupervised methods.

4.3. Noisy Priors

The previous section assumes that the system has accurate knowledge of the prior distributions over the labels. In this section, we perform a sensitively analysis by gradually smoothing the class distribution until it reaches a uniform distribution. We add noisy counts ν to the true counts $c(y)$:

$$\tilde{p}(y) = \frac{c(y) + \nu}{\sum_{y'} c(y') + \nu}.$$

As more noise is added, the prior distribution converges to uniform.

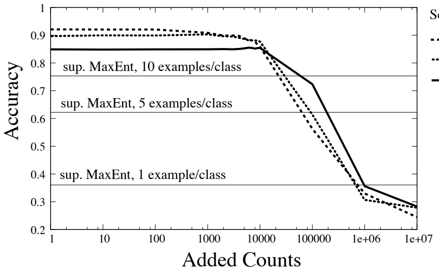

Figure 5. **CoNLL03**: The x-axis represents increasing amount of noise towards a uniform distribution. On this data set, the majority class is 84% of the instances, and so the uniform distribution is an extremely poor approximation. Performance suffers little when the majority class prior is erroneously given as 61%($\nu = 10,000$)

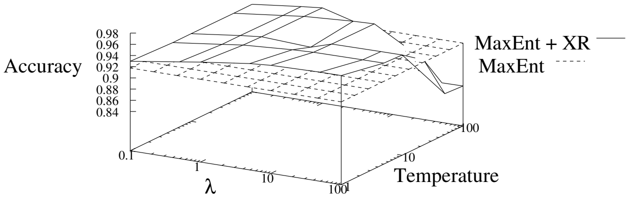

Figure 6. **CoNLL03**: For a wide range of λ and temperature the performance is similar and surpasses the purely supervised performance.

Figure 5 demonstrates the effect of increasing noise in the system. At $\nu = 1,000$, the majority class probability drops from 84% to 80% and there is almost no loss of performance. At $\nu = 10,000$ are added, the majority class probability drops to 61% and there is only a slight loss of performance. At $\nu = 1e07$ the majority class probability has dropped to 11%, a virtually uniform distribution, and performance has leveled off. These results are encouraging as they suggest that relatively large changes (of 20% absolute, 27% relative) can be tolerated without major losses in accuracy. Even when the human has no domain knowledge to contribute, label distribution estimates of sufficient accuracy should be obtainable from a reasonably small number of labeled examples.

4.4. Robustness

Along with robustness in the face of noise from the estimated label priors, the model is robust to changes in λ and temperature. As can be seen in Figure 6, λ and temperature have a wide plateau over which their performance is stable. At some extreme values of λ and temperature, the performance degrades, and can drop below supervised performance. This trend was observed for 500 labeled examples (shown in the figure), as well as in cases when there as little as one labeled example for a number of the data sets. For other semi-supervised techniques such as entropy regularization, extensive tuning is required across for each individual data set and labeled/unlabeled data set sizes in order to improve upon supervised-only performance (Jiao et al., 2006).

5. Conclusion

This paper has presented *expectation regularization*, a new method for semi-supervised learning. This method penalizes models by divergence between the model's expectations over the unlabeled data and conditional probabilities, which can be estimated from labeled data or given as prior knowledge. An important special case, *label regularization* is empirically explored, where we find it to provide accuracy improvements over entropy regularization, naïve Bayes EM, Quadratic Cost Criterion (a representative graph-based method) and a cluster kernel SVM. Our hope is that the simplicity, robustness and scalability of this method will enable semi-supervised learning to be more widely deployed.

In future work we will experiment with more general cases of expectation regularization, in which the human provides expectations on feature/label pairs. We will also ultimately apply these methods to structured models, such as conditional random fields, which, as exponential family models, are also a natural fit for XR, and in which the XR gradient can still be efficiently calculated by dynamic programming.

Acknowledgments

This work was supported in part by DoD contract #HM1582-06-1-2013, in part by The Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant #IIS-0427594, and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010.

References

Ando, R. K., & Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *JMLR*, *6*.

Bengio, Y., Dellalleau, O., & Roux, N. L. (2006). Label propagation and quadratic criterion. In O. Chapelle, B. Schlkopf and A. Zien (Eds.), *Semi-supervised learning*. MIT Press.

Burges, C., & Platt, J. (2006). Semi-supervised learning with conditional harmonic mixing. In O. Chapelle, B. Schlkopf and A. Zien (Eds.), *Semi-supervised learning*. MIT Press.

Chapelle, O., Scholkopf, B., & Zien, A. (2006). Analysis of Benchmarks. In O. Chapelle, A. Zien and B. Scholkopf (Eds.), *Semi-supervised learning*. MIT Press.

Corduneanu, A., & Jaakkola, T. (2003). On information regularization. *UAI*.

Cozman, F., & Cohen, I. (2006). Risks of Semi-Supervised Learning. In O. Chapelle, A. Zien and B. Scholkopf (Eds.), *Semi-supervised learning*. MIT Press.

Delalleau, O., Bengio, Y., & Le Roux, N. (2006). Large-scale algorithms. In O. Chapelle, B. Schölkopf and A. Zien (Eds.), *Semi-supervised learning*.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *J. Royal Stat. Soc.*, *39*, 1–38.

Grandvalet, Y., & Bengio, Y. (2004). Semi-supervised learning by entropy minimization. *NIPS*.

Jiao, F., Wang, S., Lee, C.-H., Greiner, R., & Schuurmans, D. (2006). Semi-supervised conditional random fields for improved sequence segmentation and labeling. *COLING/ACL*.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. *ICML*.

Klein, D., & Manning, C. (2004). Corpus-based induction of syntactic structure: Models of dependency and constituency. *ACL*.

Kockelkorn, M., Luneburg, A., & Scheffer, T. (2003). Using transduction and multi-view learning to answer emails. *PKDD*.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of ICML*.

Li, W., & McCallum, A. (2004). *A note on semi-supervised learning using markov random fields*Computer Science Technical Note). University of Massachusetts, Amherst.

Li, W., & McCallum, A. (2005). Semi-supervised sequence modeling with syntactic topic models. *AAAI*.

Macskassy, S., & Provost, F. (2006). *Classification in networked data* (Technical Report CeDER-04-08). New York University.

Malouf, R. (2002). A comparison of algorithms for maximum enotrpy parameter estimation. *COLING*.

Merialdo, P. (1994). Tagging english text with a probabilistic model. *Computational Linguistics*.

Miller, S., Guinness, J., & Zamanian, A. (2004). Name tagging with word clusters and discriminative training. *ACL*.

Nigam, K., McCallum, A., & Mitchell., T. (2006). Semi-supervised Text Classification Using EM. In O. Chapelle, A. Zien and B. Scholkopf (Eds.), *Semi-supervised learning*. MIT Press.

Nigam, K., McCallum, A., Thrun, S., & Mitchell, T. (1998). Learning to classify text from labeled and unlabeled documents. *AAAI*.

Niu, Z.-Y., Ji, D.-H., & Tam, C. L. (2005). Word sense disambiguation using label propagation based semi-supervised learning. *ACL*.

Schapire, R., Rochery, M., Rahim, M., & Gupta, N. (2002). Incorporating prior knowledge into boosting. *ICML*.

Schuurmans, D. (1997). A new metric-based approach to model selection. *AAAI*.

Sindhwani, V., & Keerthi, S. S. (2006). Large Scale Semi-supervised Linear SVMs. *SIGIR*.

Smith, N., & Eisner, J. (2005). Contrastive estimation: Training log-linear models on unlabeled data. *ACL*.

Sutton, C., & McCallum, A. (2006). An introduction to conditional random fields for relational learning. In L. Getoor and B. Taskar (Eds.), *Introduction to statistical relational learning*. MIT Press.

Szummer, M., & Jaakkola, T. (2002). Partially labeled classification with markov random walks. *NIPS*.

Weston, J., Leslie, C., Ie, E., & Noble, W. S. (2006). Semi-supervised protein classification using cluster kernels. *Semi-Supervised Learning*.

Zhu, X., & Ghahramani, Z. (2002). *Learning from labeled and unlabeled data with label propagation* (Technical Report CMU-CALD-02-107). CMU.

Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using gaussian fields and harmonic mixtures. *ICML*.

Zhu, X., & Lafferty, J. (2005). Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. *ICML*.