SEMI-SUPERVISED LEARNING FOR TEXT CLASSIFICATION USING FEATURE AFFINITY REGULARIZATION

Bin Zhang, Mari Ostendorf

University of Washington, Seattle, WA 98195

ABSTRACT

Most conventional semi-supervised learning methods attempt to directly include unlabeled data into training objectives. This paper presents an alternative approach that learns feature affinity information from unlabeled data, which is incorporated into the training objective as regularization of a maximum entropy model. The regularization favors models for which correlated features have similar weights. The method is evaluated in text classification, where feature affinity can be computed from feature co-occurrences in unlabeled data. Experimental results show that this method consistently outperforms baseline methods.

Index Terms— semi-supervised learning, text classification, maximum entropy, feature affinity matrix, regularization

1. INTRODUCTION

Semi-supervised learning techniques are used in many problems to leverage unlabeled data [1]. Applying semi-supervised learning to text classification is desirable due to limited labeled text and increasing amount of unlabeled text. A popular approach is to use the expectation-maximization (EM) algorithm on generative models such as Naive Bayes, treating unlabeled data as data with missing information [2]. Discriminative models are found to be more effective than generative models in many text classification tasks [3, 4, 5], because they do not need to estimate the joint probability distributions. The maximum entropy (MaxEnt) model [6] is a widely used discriminative classifier due to its computational tractability and straightforward optimization. However, because MaxEnt only computes the conditional probability of a label given features, the EM algorithm cannot be directly applied.

Various alternatives for semi-supervised learning have been explored for MaxEnt models. Suppose we are given a set of label data $\mathcal{L}$ and unlabeled data $\mathcal{U}$ that come from the same distribution, and the goal is to learn a classifier based on both $\mathcal{L}$ and $\mathcal{U}$. Grandvalet et al. proposed entropy minimization in [7], which adds an entropy regularization term to the traditional MaxEnt training objective. This regularizer encourages lower entropy of predicted labels on unlabeled data. The intuition behind entropy minimization is the principle used by many semi-supervised learning methods, that is, classification boundaries should lie in low density regions. By favoring lower entropy (thus more certain) predictions on unlabeled data, the decision boundaries are effectively pushed away from high density regions. As with other methods that use this principle, it may not work well if the data has significant overlaps between different classes, where the classifier may get over-confident for the incorrect decisions.

Mann et al. [8] proposed another semi-supervised MaxEnt learning method named expectation regularization (XR). Like entropy minimization, XR also adds a new regularization term to the training objective, specifically the Kullback-Leibler divergence $D(\tilde{p}\|\hat{p})$ between the model-expected label distribution $\hat{p}$ of unlabeled data and a prior label distribution $\tilde{p}$. It prefers models that produce a label distribution closer to the prior, which may be estimated empirically from labeled data or specified by experts. It was shown that XR outperforms entropy minimization on a few tasks [8]. Other work has invetigated l_2 posterior label distribution regularization for conditional randome fields [9]. XR has also been extended to weakly-supervised MaxEnt learning by annotating salient features, from which a label distribution can be estimated [10].

In this paper, we attempt to use unlabeled data in semi-supervised learning in a different way. Instead of regularizing the labels or posteriors on the unlabeled data, we regularize the feature weights based on the structure of features learned from unlabeled data. This approach is motivated by the fact that the feature space in text classification problems is typically very high dimension, so many features are not observed in a limited labeled training data set. Since the posterior predictions based on a more limited feature set are weak, it may be more productive to focus on leveraging information about the features in the unlabeled data. Specifically, we propose feature affinity regularization (FAR), which learns feature affinity information from unlabeled data and incorporates it in the training objective in the form of regularization, such that we favor models which assign similar weights to correlated features. With FAR, even if we only observe a small number of features on labeled data, the model will still be aware of other unseen features that are related to the observed features. Therefore, we can train a model that generalizes better to new data. The method applies to both inductive and transductive classifiers, depending on whether the test data

can be incorporated in the unlabeled set when learning feature correlation.

A similar algorithm is proposed as regularizing with networks of features, in work by Sandler et al. [11], also motivated by the challenges of high-dimensional features in text classification. They show that the Gaussian form of their regularization is equivalent to the locally linear embedding (LLE) method for transforming high-dimensional data [12], without the dimensionality reduction, and achieves better results for two text classification problems. A key difference in our work is the mechanism for representing feature affinity. In [11], the feature similarity measure is application-dependent. In contrast, our work represents feature affinity purely based on co-occurrence, which is computationally cheap to obtain and (we conjecture) generalizes reasonably across tasks. It is shown to be helpful in our experiments on two tasks.

Another method that explores feature similarity (or more precisely, feature correspondence) from unlabeled data is structural correspondence learning (SCL) [13]. SCL can be applied to domain transfer, where source and target domains are unmatched, and labeled data is only available in the source domain. SCL defines a set of pivot features which are common across domains, and constructs linear predictors to predict those pivot features from unlabeled data in both domains. A feature in the source domain is said to correspond to a feature in the target domain if they are both useful in predicting the pivot features, and a linear combination of them is used as a new feature in supervised learning. FAR is similar in motivation, but is different from SCL in a few aspects: it does not require hand-picking of pivot features, and it is designed for within-domain semi-supervised learning (vs. domain transfer).

2. FEATURE AFFINITY REGULARIZATION

Assume that, for each N-dimensional feature vector $\mathbf{x}$, the MaxEnt model computes the posterior probability for class $k = 1, 2, \ldots, K$ based on the following formula,

$$p(y = k|\mathbf{x}) = \frac{\exp\left(\sum_{i=1}^{N} \lambda_{ik} x_i\right)}{\sum_{j=1}^{K} \exp\left(\sum_{i=1}^{N} \lambda_{ij} x_i\right)},$$

where λ_{ik}'s are real-valued feature weights which describe how indicative feature i is with respect to class k. They form an $N \times K$ weight matrix $\mathbf{\Lambda}$.

The FAR regularizer is defined as

$$J_{\text{FAR}} = \frac{1}{2} \sum_{k=1}^{K} \sum_{i,j,\ i<j} w_{ij} (\lambda_{ik} - \lambda_{jk})^2,$$

where w_{ij}'s are elements of a symmetric matrix $\mathbf{W}$. We call $\mathbf{W}$ a *feature affinity matrix* as it encodes affinity information between pairs of features. The larger w_{ij} is, the more correlated feature i and feature j are. All w_{ij}'s are non-negative.

To use FAR in MaxEnt model learning, we add J_{FAR} to the training objective function,

$$J = -\sum_{(\mathbf{x},y)\in\mathcal{L}} \log p(y|\mathbf{x}) + \sum_{i,k} \frac{\lambda_{ik}^2}{2\sigma^2} + \gamma J_{\text{FAR}}.$$

The first two terms are negative loglikelihood and an l_2 regularizer, respectively. l_2 regularization is a standard technique to perform weight smoothing for MaxEnt, and it improves the classification accuracy in testing. σ and $\gamma > 0$ are hyperparameters that are tuned on development data.

It can be shown that J_{FAR} is convex with respect to $\mathbf{\Lambda}$. Therefore, overall J is still convex with respect to $\mathbf{\Lambda}$, and we can safely use gradient algorithms to minimize J without the risk of being stuck at local optima. We use L-BFGS in our experiments.

The gradient of J_{FAR} can be computed as follows. In this derivation, δ_{ij} is the Kronecker delta.

$$\begin{aligned}
\frac{\partial J_{\text{FAR}}}{\partial \lambda_{mn}} &= \sum_{i,j,\ i<j} w_{ij}(\lambda_{in} - \lambda_{jn}) \frac{\partial(\lambda_{in} - \lambda_{jn})}{\partial \lambda_{mn}} \\
&= \sum_{i,j,\ i<j} w_{ij}(\lambda_{in} - \lambda_{jn})(\delta_{im} - \delta_{jm}) \\
&= \sum_{j,\ m<j} w_{mj}(\lambda_{mn} - \lambda_{jn}) \\
&\quad - \sum_{i,\ i<m} w_{im}(\lambda_{in} - \lambda_{mn}) \\
&= \sum_{i} w_{im}(\lambda_{mn} - \lambda_{in}).
\end{aligned}$$

Intuitively, the gradient pushes λ_{mn} towards λ_{in} if $w_{im} > 0$, therefore correlated features get similar weights.

Unlike XR, where specifying a prior distribution over a small set of labels is relatively easy, coming up with an $N \times N$ matrix $\mathbf{W}$ for FAR can be a daunting task if it is done by a human. In this work, we propose to learn $\mathbf{W}$ automatically from unlabeled data, including test data. This approach has a few benefits. First, information from unlabeled data is brought into the training objective indirectly through $\mathbf{W}$, thus it becomes semi-supervised. Second, deriving feature affinity information automatically from data makes it easy to apply to a variety of tasks.

In this paper, we investigate a simple way to compute $\mathbf{W}$, that is, using feature co-occurrence. If feature i and feature j co-occur many times on a sufficiently large data set, we can conjecture that feature i and feature j are likely to have similar predictive power; thus, the affinity value w_{ij} should be high. Once FAR is applied to MaxEnt, it will push the feature weights λ_i and λ_j closer.

We propose the following way to compute $\mathbf{W}$.

$$w_{ij}^{\text{raw}} = \begin{cases} C(f_i, f_j) & \text{if } C(f_i, f_j) > \theta \\ 0 & \text{otherwise} \end{cases},$$

where $C(f_i, f_j)$ is the number of unlabeled instances in which f_i and f_j co-occur. Co-occurrence counts tend to have very long tails, and the majority are small numbers. For large datasets, this poses challenges to computation and storage. In this case, setting a non-zero threshold θ is necessary, which leads to a sparser $\mathbf{W}$. We use the smallest θ that satisfies our computational constraints. The impact of FAR to MaxEnt learning is also controlled by the hyper-parameter γ, which will be less dependent on the size of unlabeled data $\mathcal{U}$, if we compute $\mathbf{W}$ as

$$w_{ij}^{\text{normalized}} = \frac{w_{ij}^{\text{raw}}}{|\mathcal{U}|},$$

in which the size of unlabeled data is taken into account.

Some degenerative cases further motivate FAR. Consider the case where $\mathbf{W}$ only has two non-zero elements w_{uv} and w_{vu}, $u \neq v$. That is to say, only feature u and feature v are correlated. If only u is observed in the training data, the model trained using FAR will have $\lambda_{uk} = \lambda_{vk}$, $k = 1, 2, \ldots, K$ (ignoring the effect of l_2 regularization for now). In other words, the model will treat feature v the same as feature u, even though feature v does not appear in the training data. This enables MaxEnt to extend the model to *unseen features*. Conventional MaxEnt will have $\lambda_{vk} = 0$, effectively ignoring unseen feature v.

If both feature u and feature v appear in the training data, with FAR, the distance $|\lambda_{uk} - \lambda_{vk}|$ will be less than that obtained by conventional MaxEnt. How close they become depends on the strength of the regularizer, controlled by γ. This is effectively performing *feature weight smoothing* based on feature affinity. With an appropriate way to derive $\mathbf{W}$, this smoothing effect can help improve model generalizability.

In practice, it is almost impossible to obtain a degenerative $\mathbf{W}$ as mentioned above. There will usually be many more non-zero elements. Each non-zero element corresponds to an edge in the graph of all features. Most likely, a feature will be connected to many other features. If we put the feature weights into the corresponding graph nodes, FAR enforces weight smoothness over the entire feature graph. Features with more direct or indirect connections tend to have similar weights in a MaxEnt model trained with FAR.

3. EXPERIMENTS

We carry out semi-supervised text classification experiments on two tasks, including movie review classification and newsgroups topic classification. We compare FAR with a supervised baseline (l_2 regularized MaxEnt) and two semi-supervised baselines (MaxEnt with XR and MaxEnt with self-training). Self-training [1] is an alternative to EM that can be applied to a variety of models including MaxEnt. In order to test the performance of semi-supervised learning under different conditions, we downsample the training data randomly using different sampling rates while maintaining the original class prior. The unlabeled data used to learn feature affinity matrix include all the ones that are not sampled for training, as well as all the development and evaluation data with labels removed.

3.1. Movie Review Classification

The task of movie review classification was introduced in [5] as an application of sentiment analysis. The authors showed that MaxEnt with unigram features outperforms other methods. In this paper, we use the movie review data set described in [14]. This data set is comprised of 5000 positive and 5000 negative sentences extracted from `www.rottentomatoes.com` reviews.

We use ten-fold cross validation in our experiments, and we keep training-development-evaluation ratio to 8:1:1. The results reported are classification accuracies.

Table 1. Movie review classification results

Downsampling	l_2	XR	self	FAR
1%	54.4%	54.7%	51.8%	**56.2%**
2%	57.6%	57.3%	54.2%	**59.2%**
5%	62.0%	61.8%	59.2%	**63.7%**
10%	65.6%	65.6%	64.2%	**67.2%**
20%	69.0%	69.0%	68.7%	**70.4%**
50%	72.9%	72.9%	72.9%	**74.1%**

3.2. Newsgroups Topic Classification

We use the 20 newsgroups data set [15], which is a collection of approximately 20,000 newsgroup documents, partitioned nearly evenly across 20 different newsgroups as classes. We employ the "bydate" version of the data as recommended by the author.

This data set has been divided into training and test sets. When downsampling the training set for semi-supervised learning, we also create a development set using the data not sampled for training, and the size of the development set is about 10% of the original training set.

TF-IDF feature is extracted from each document, as what is typically done in the text categorization literature. We evaluate the performance using classification accuracy.

3.3. Results and Discussion

From the experimental results in Tables 1 and 2, we see that FAR consistently outperforms other three baseline methods. The experiments also show that FAR not only works on binary features used in movie review classification, but also is able to achieve improvements with real-valued TF-IDF features. XR and self-training do not work well in these experiments. In most cases, they obtain inferior results compared even to the conventional l_2 MaxEnt (supervised learning only).

Table 2. 20 newsgroups topic classification results

Downsampling	l_2	XR	self	FAR
1%	15.3%	15.2%	13.8%	**15.4%**
2%	34.5%	32.4%	32.4%	**34.6%**
5%	55.6%	54.6%	54.5%	**57.4%**
10%	66.9%	67.8%	66.7%	**68.2%**
20%	73.9%	73.8%	74.0%	**75.6%**
50%	80.2%	79.9%	79.9%	**81.7%**

In our experiments, the number of features N is on the order of tens of thousands. In order to run FAR, we need to store $\mathbf{W}$ which is $N \times N$ and in most cases sparse. It is manageable for the movie review data even if we do not do any pruning ($\theta = 0$), but we had to set $\theta = 10$ for the newsgroups data to reduce storage and computation burden. We will face a bigger problem when we use millions of features, which is not uncommon for some tasks. Setting a greater θ increases the sparsity of $\mathbf{W}$, but we lose more feature affinity information. We are working on a more compressed representation of $\mathbf{W}$ to be able to use more features. In this paper, we use universal feature co-occurrence to compute feature affinity, but it may not be optimal for all tasks. Users may tailor the method such that $\mathbf{W}$ captures the most important feature affinity information with regard to their task at hand.

4. CONCLUSIONS

We have introduced a novel semi-supervised learning method named FAR for text classification. A feature affinity matrix is learned from unlabeled data, typically from feature co-occurrence statistics. Then a regularization term is added to the supervised MaxEnt training objective, measuring the sum of distances between pairs of feature weights weighted by feature affinity values. The overall effect of FAR on MaxEnt is that it favors models that assign similar weights to correlated features, therefore smooths the weights over all features. We have shown that FAR can extend the model to features that are unseen in the training data, and can also improve the generalizability of the model. Experiments on text classification have been carried out, and we have observed consistent improvement from FAR over other semi-supervised learning approaches for MaxEnt models.

There are several avenues for future work. Improving the efficiency of FAR would facilitate use with more feature, such as n-grams. There are other possibilities for computing the affinity matrix, e.g. using general quantities such as mutual information. It would also be of interest to compare FAR to other baselines, including weight tying determined by automatic clustering, other variations of network regularization in [11], and label posterior regularization. Although we have only tested FAR on text classification, the same principle can be applied to other classification problems. For problems in domains with real-valued (vs. symbolic) observations, computing feature affinity from feature correlation may be a viable approach.

5. REFERENCES

[1] Xiaojin Zhu, "Semi-supervised learning literature survey," Tech. Rep. 1530, "Computer Sciences, University of Wisconsin-Madison, 2005.

[2] Kamal Nigam, Andrew Mccallum, and Tom Mitchell, "Semi-supervised text classification using EM," in *Semi-Supervised Learning*, pp. 33–56. MIT Press, 2006.

[3] Thorsten Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *LNCS, Proc. ECML*, 1998, vol. 1398, pp. 137–142.

[4] Kamal Nigam, John Lafferty, and Andrew McCallum, "Using maximum entropy for text classification," in *Proc. IJCAI-99 Workshop on Machine Learning for Information Filtering*, 1999, pp. 61–67.

[5] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. EMNLP*, 2002, pp. 79–86.

[6] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguist.*, vol. 22, no. 1, pp. 39–71, 1996.

[7] Yves Grandvalet and Yoshua Bengio, "Semi-supervised learning by entropy minimization," in *Proc. NIPS*, 2004.

[8] Gideon S. Mann and Andrew McCallum, "Simple, robust, scalable semi-supervised learning via expectation regularization," in *Proc. ICML*, 2007, pp. 593–600.

[9] Amarnag Subramanya, Slav Petrov, and Fernando Pereira, "Efficient graph-based semi-supervised learning of structured tagging models," in *Proc. EMNLP*, 2010, pp. 167–176.

[10] Gideon S. Mann and Andrew McCallum, "Generalized expectation criteria for semi-supervised learning with weakly labeled data," *J. Mach. Learn. Res.*, vol. 11, pp. 955–984, 2010.

[11] Ted Sandler, Partha Pratim Talukdar, Lyle H. Ungar, and John Blitzer, "Regularized learning with networks of features," in *Proc. NIPS*, 2008.

[12] Sam T. Roweis and Lawrence K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[13] John Blitzer, Ryan McDonald, and Fernando Pereira, "Domain adaptation with structural correspondence learning," in *Proc. EMNLP*, 2006, pp. 120–128.

[14] Bo Pang and Lillian Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proc. ACL*, 2004, pp. 271–278.

[15] Jason Rennie, "The 20 newsgroups data set," `http://people.csail.mit.edu/jrennie/20Newsgroups`.