Active Learning with Confidence

Mark Dredze and **Koby Crammer**
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
{mdredze,crammer}@cis.upenn.edu

Abstract

Active learning is a machine learning approach to achieving high-accuracy with a small amount of labels by letting the learning algorithm choose instances to be labeled. Most of previous approaches based on discriminative learning use the margin for choosing instances. We present a method for incorporating confidence into the margin by using a newly introduced online learning algorithm and show empirically that confidence improves active learning.

1 Introduction

Successful applications of supervised machine learning to natural language rely on quality labeled training data, but annotation can be costly, slow and difficult. One popular solution is Active Learning, which maximizes learning accuracy while minimizing labeling efforts. In active learning, the learning algorithm itself selects unlabeled examples for annotation. A variety of techniques have been proposed for selecting examples that maximize system performance as compared to selecting instances randomly.

Two learning methodologies dominate NLP applications: probabilistic methods — naive Bayes, logistic regression — and margin methods — support vector machines and passive-aggressive. Active learning for probabilistic methods often uses uncertainty sampling: label the example with the lowest probability prediction (the most "uncertain") (Lewis and Gale, 1994). The equivalent technique for margin learning associates the margin with prediction certainty: label the example with the lowest margin (Tong and Koller, 2001). Common intuition equates large margins with high prediction confidence.

However, confidence and margin are two distinct properties. For example, an instance may receive a large margin based on a single feature which has been updated only a small number of times. Another example may receive a small margin, but its features have been learned from a large number of examples. While the first example has a larger margin it has low confidence compared to the second. Both the margin value and confidence should be considered in choosing which example to label.

We present active learning with confidence using a recently introduced online learning algorithm called Confidence-Weighted linear classification. The classifier assigns labels according to a Gaussian distribution over margin values instead of a single value, which arises from parameter confidence (variance). The variance of this distribution represents the confidence in the mean (margin). We then employ this distribution for a new active learning criteria, which in turn could improve other margin based active learning techniques. Additionally, we favor the use of an online method since online methods have achieved good NLP performance and are fast to train — an important property for interactive learning. Experimental validation on a number of datasets shows that active learning with confidence can improve standard methods.

2 Confidence-Weighted Linear Classifiers

Common online learning algorithms, popular in many NLP tasks, are not designed to deal with the particularities of natural language data. Fea-

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 233–236,
Columbus, Ohio, USA, June 2008.

ture representations have very high dimension and most features are observed on a small fraction of instances. Confidence-weighted (CW) linear classification (Dredze et al., 2008), a new online algorithm, maintains a probabilistic measure of parameter confidence leading to a measure of prediction confidence, potentially useful for active learning. We summarize CW learning to familiarize the reader.

Parameter confidence is formalized with a distribution over weight vectors, specifically a Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^N$ and diagonal covariance $\Sigma \in \mathbb{R}^{N \times N}$. The values μ_j and $\Sigma_{j,j}$ represent knowledge of and confidence in the parameter for feature j. The smaller $\Sigma_{j,j}$, the more confidence we have in the mean parameter value μ_j.

A model predicts the label with the highest probability, $\max_{y \in \{\pm 1\}} \Pr_{\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)} [y(\boldsymbol{w} \cdot \boldsymbol{x}) \geq 0]$. The Gaussian distribution over parameter vectors $\boldsymbol{w}$ induces a univariate Gaussian distribution over the unsigned-margin $\mathrm{M} = \boldsymbol{w} \cdot \boldsymbol{x}$ parameterized by $\boldsymbol{\mu}$, Σ and the instance $\boldsymbol{x}$: $\mathrm{M} \sim \mathcal{N}(M, V)$, where the mean is $M = \boldsymbol{\mu} \cdot \boldsymbol{x}$ and the variance $V = \boldsymbol{x}^\top \Sigma \boldsymbol{x}$.

CW is an online algorithm inspired by the Passive Aggressive (PA) update (Crammer et al., 2006) — which ensures a positive margin while minimizing parameter change. CW replaces the Euclidean distance used in the PA update with the KL divergence over Gaussian distributions. It also replaces the minimal margin constraint with a minimal probability constraint: with some given probability $\eta \in (0.5, 1]$ a drawn classifier will assign the correct label. This strategy yields the following objective solved on each round of learning:

$$(\boldsymbol{\mu}_{i+1}, \Sigma_{i+1}) = \min \mathrm{D}_{\mathrm{KL}} \left(\mathcal{N}(\boldsymbol{\mu}, \Sigma) \parallel \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i) \right) \\ \text{s.t. } \Pr \left[y_i \left(\boldsymbol{\mu} \cdot \boldsymbol{x}_i \right) \geq 0 \right] \geq \eta \,,$$

where $(\boldsymbol{\mu}_i, \Sigma_i)$ are the parameters on round i and $\left(\boldsymbol{\mu}_{i+1}, \Sigma_{i+1}\right)$ are the new parameters after update. The constraint ensures that the resulting parameters will correctly classify $\boldsymbol{x}_i$ with probability at least η. For convenience we write $\phi = \Phi^{-1}(\eta)$, where Φ is the cumulative function of the normal distribution. The optimization problem above is not convex, but a closed form approximation of its solution has the following additive form: $\boldsymbol{\mu}_{i+1} = \boldsymbol{\mu}_i + \alpha_i y_i \Sigma_i \boldsymbol{x}_i$ and $\Sigma_{i+1}^{-1} = \Sigma_i^{-1} + 2\alpha_i \phi \boldsymbol{x}_i \boldsymbol{x}_i^\top$ for,

$$\alpha_i = \frac{-(1+2\phi M_i) + \sqrt{(1+2\phi M_i)^2 - 8\phi\,(M_i - \phi V_i)}}{4\phi V_i} \,.$$

Each update changes the feature weights $\boldsymbol{\mu}$, and increases confidence (variance Σ always decreases).

3 Active Learning with Confidence

We consider pool based active learning. An active learning algorithm is given a pool of unlabeled instances $\mathcal{U} = \{\boldsymbol{x}_i\}_{i=1}^n$, a learning algorithm $\mathcal{A}$ and a set of labeled examples initially set to be $\mathcal{L} = \emptyset$. On each round the active learner uses its selection criteria to return a single instance $\boldsymbol{x}_i$ to be labeled by an annotator with $y_i \in \{-1, +1\}$ (for binary classification). The instance and label are added to the labeled set $\mathcal{L} \leftarrow \mathcal{L} \cup \{(\boldsymbol{x}_i, y_i)\}$ and passed to the learning algorithm $\mathcal{A}$, which in turn generates a new model. At the end of labeling the algorithm returns a classifier trained on the final labeled set. Effective active learning minimizes prediction error and the number of labeled examples.

Most active learners for margin based algorithms rely on the magnitude of the margin. Tong and Koller (2001) motivate this approach by considering the half-space representation of the hypothesis space for learning. They suggest three margin based active learning methods: Simple margin, MaxMin margin, and Ratio margin. In Simple margin, the algorithm predicts an unsigned margin M for each instance in $\mathcal{U}$ and returns for labeling the instance with the smallest margin. The intuition is that instances for which the classifier is uncertain (small margin) provide the most information for learning. Active learning based on PA algorithms runs in a similar fashion but full SVM retraining on every round is replaced with a single PA update using the new labeled example, greatly increasing learning speed.

Maintaining a distribution over prediction functions makes the CW algorithm attractive for active learning. Instead of using a geometrical quantity ("margin"), it use a probabilistic quantity and picks the example whose label is predicted with the lowest probability. Formally, the margin criteria, $\boldsymbol{x} = \arg\min_{\boldsymbol{z} \in \mathcal{U}} (\boldsymbol{w} \cdot \boldsymbol{z})$, is replaced with a probabilistic criteria $\boldsymbol{x} = \arg\min_{\boldsymbol{z} \in \mathcal{U}} \left| \left(\Pr_{\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)} [\mathrm{sign}(\boldsymbol{w} \cdot \boldsymbol{z}) = 1] \right) - \frac{1}{2} \right|$.

The selection criteria naturally captures the notion that we should label the example with the highest uncertainty. Interestingly, we can show (omitted due to lack of space) that the probabilistic criteria can be translated into a *corrected* geometrical criteria. In practice, we can compute this normalized margin as $\bar{\mathrm{M}} = M/\sqrt{V}$. We call this selection criteria Active Confident Learning (ACL).

4 Evaluation

To evaluate our active learning methods we used a similar experimental setup to Tong and Koller (2001). Each active learning algorithm was given two labeled examples, one from each class, for initial training of a classifier, and remaining data as unlabeled examples. On each round the algorithm selected a single instance for which it was then given the correct label. The algorithm updated the online classifier and evaluated it on held out test data to measure learning progress.

We selected four binary NLP datasets for evaluation: 20 Newsgroups[1] and Reuters (Lewis et al., 2004) (used by Tong and Koller) and sentiment classification (Blitzer et al., 2007) and spam (Bickel, 2006). For each dataset we extracted binary unigram features and sentiment was prepared according to Blitzer et al. (2007). From 20 Newsgroups we created 3 binary decision tasks to differentiate between two similar labels from computers, science and talk. We created 3 similar problems from Reuters from insurance, business services and retail distribution. Sentiment used 4 Amazon domains (book, dvd, electronics, kitchen). Spam used the three users from task A data. Each problem had 2000 instances except for 20 Newsgroups, which used between 1850 and 1971 instances. This created 13 classification problems across four tasks.

Each active learning algorithm was evaluated using a PA (with slack variable $c = 1$) or CW classifier ($\phi = 1$) using 10-fold cross validation. We evaluated several methods in the Simple margin framework: PA Margin and CW Margin, which select examples with the smallest margin, and ACL. As a baseline we included selecting a random instance. We also evaluated CW and a PA classifier trained on all training instances. Each method was evaluated by labeling up to 500 labels, about 25% of the training data. The 10 runs on each dataset for each problem appear in the left and middle panel of Fig. 1, which show the test accuracy after each round of active learning. Horizontal lines indicate CW (solid) and PA (dashed) training on all instances. Legend numbers are accuracy after 500 labels. The left panel averages results over 20 Newsgroups, and the middle panel averages results over *all* 13 datasets.

To achieve 80% of the accuracy of training on all data, a realistic goal for less than 100 labels, PA Margin required 93% the number of labels of PA Random, while CW Margin needed only 73% of the labels of CW Random. By using fewer labels compared to random selection baselines, CW Margin learns faster in the active learning setting as compared with PA. Furthermore, adding confidence reduced labeling cost compared to margin alone. ACL improved over CW Margin on every task and after almost every round; it required 63% of the labels of CW Random to reach the 80% mark.

We computed the fraction of labels CW Margin and ACL required (compared to CW Random) to achieve the 80% accuracy mark of training with all data. The results are summarized in the right panel of Fig. 1, where we plot one point per dataset. Points above the diagonal-line demonstrate the superiority of ACL over CW Margin. ACL required fewer labels than CW margin twice as often as the opposite occurred (8 vs 4). Note that CW Margin used *more* labels than CW Random in three cases, while ACL only once, and this one time only about a dozen labels were needed. To conclude, not only does CW Margin outperforms PA Margin for active-learning, CW maintains additional valuable information (confidence), which further improves performance.

5 Related Work

Active learning has been widely used for NLP tasks such as part of speech tagging (Ringger et al., 2007), parsing (Tang et al., 2002) and word sense disambiguation (Chan and Ng, 2007). Many methods rely on entropy-based scores such as uncertainty sampling (Lewis and Gale, 1994). Others use margin based methods, such as Kim et al. (2006), who combined margin scores with corpus diversity, and Sassano (2002), who considered SVM active learning

[1] http://people.csail.mit.edu/jrennie/20Newsgroups/

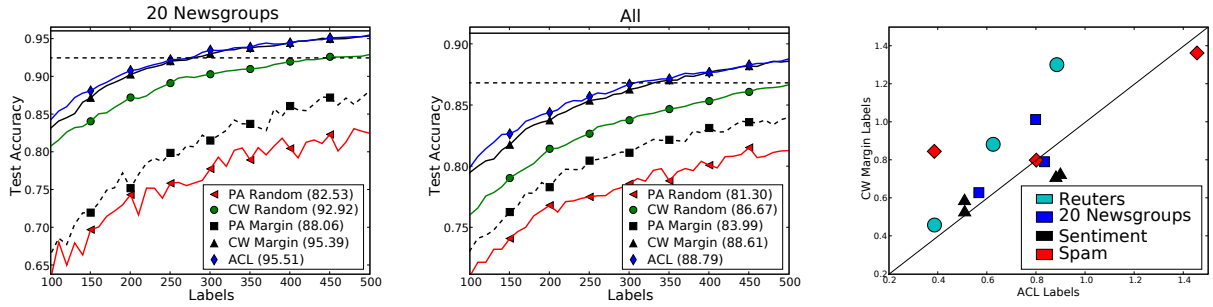

Figure 1: Results averaged over 20 Newsgroups (left) and all datasets (center) showing test accuracy over active learning rounds. The right panel shows the amount of labels needed by CW Margin and ACL to achieve 80% of the accuracy of training on all data - each points refers to a different dataset.

for Japanese word segmentation. Our confidence based approach can be used to improve these tasks. Furthermore, margin methods can outperform probabilistic methods; CW beats maximum entropy on many NLP tasks (Dredze et al., 2008).

A theoretical analysis of margin based methods selected labels that maximize the reduction of the version space, the hypothesis set consistent with the training data (Tong and Koller, 2001). Another approach selects instances that minimize the future error in probabilistic algorithms (Roy and McCallum, 2001). Since we consider an online learning algorithm our techniques can be easily extended to online active learning (Cesa-Bianchi et al., 2005; Dasgupta et al., 2005; Sculley, 2007).

6 Conclusion

We have presented techniques for incorporating confidence into the margin for active learning and have shown that CW selects better examples than PA, a popular online algorithm. This approach creates opportunities for new active learning frameworks that depend on margin confidence.

References

S. Bickel. 2006. Ecml-pkdd discovery challenge overview. In *The Discovery Challenge Workshop*.

J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*.

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stolt. 2005. Minimizing regret with label efficient prediction. *IEEE Tran. on Inf. Theory*, 51(6), June.

Y. S. Chan and H. T. Ng. 2007. Domain adaptation with active learning for word sense disambiguation. In *Association for Computational Linguistics (ACL)*.

K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. 2006. Online passive-aggressive algorithms. *JMLR*, 7:551–585.

S. Dasgupta, A.T. Kalai, and C. Monteleoni. 2005. Analysis of perceptron-based active learning. In *COLT*.

Mark Dredze, Koby Crammer, and Fernando Pereira. 2008. Confidence-weighted linear classification. In *ICML*.

S. Kim, Yu S., K. Kim, J-W Cha, and G.G. Lee. 2006. Mmr-based active machine learning for bio named entity recognition. In *NAACL/HLT*.

D. D. Lewis and W. A. Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR*.

D. D. Lewis, Y. Yand, T. Rose, and F. Li. 2004. Rcv1: A new benchmark collection for text categorization research. *JMLR*, 5:361–397.

E. Ringger, P. McClanahan, R. Haertel, G. Busby, M. Carmen, J. Carroll, K. Seppi, and D. Lonsdale. 2007. Active learning for part-of-speech tagging: Accelerating corpus annotation. In *ACL Linguistic Annotation Workshop*.

N. Roy and A. McCallum. 2001. Toward optimal active learning through sampling estimation of error reduction. In *ICML*.

Manabu Sassano. 2002. An empirical study of active learning with support vector machines for japanese word segmentation. In *ACL*.

D. Sculley. 2007. Online active learning methods for fast label-efficient spam filtering. In *CEAS*.

M. Tang, X. Luo, and S. Roukos. 2002. Active learning for statistical natural language parsing. In *ACL*.

S. Tong and D. Koller. 2001. Supprt vector machine active learning with applications to text classification. *JMLR*.