RC22968 (W0311-032) November 5, 2003
Computer Science

IBM Research Report

An Approach to the Unlabeled-Labeled Data Problem

Rie Kubota Ando
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

An Approach to the Unlabeled-Labeled Data Problem

Rie Kubota Ando
IBM T.J. Watson Research Center
19 Skyline Dr., Hawthorne, NY 10532
rie1@us.ibm.com

Supervised learning techniques have been successfully applied to a wide range of natural language processing (NLP) problems – e.g., part-of-speech tagging, syntactic phrase chunking, named entity recognition, and so forth. However, the need for large labeled training data raises pragmatic issues when new target classes or texts of new domains need be explored.

While generation of labeled training data involves expensive manual effort, *unlabeled data* can be easily obtained in large amounts. This fact motivates supervised learning with unlabeled data, such as *co-training* (e.g., Blum and Mitchell (1998)) and *active learning* (e.g.,(Krogh and Vedelsby, 1994; McCallum and Nigam, 1998)). Active learning seeks to minimize the amount of labeled training data by actively choosing the most informative examples from unlabeled data. In practice, chosen examples need be labeled manually. Co-training involves two classifiers employing two different 'views' (e.g., 'textual content' and 'hyperlink' of web documents as in Blum and Mitchell (1998)). Each of the classifiers is repeatedly trained, in turn, using the original labeled data plus the data selected from the data newly labeled by the other. Blum and Mitchell (1998)'s experiments show that the final classifier (trained by automatically augmented labeled data) outperforms the initial classifier on the text classification task. Nevertheless, co-training seems not so easy to apply. It has the potential danger of degrading the labeled data (Pierce and Cardie, 2001). In theory, conditions for co-training to work well have been relaxed from conditional independence of two views (Blum and Mitchell, 1998) to *weak rule dependence* (Abney, 2002). Nevertheless, two "*redundantly sufficient* (Mitchell, 1999)" views may not be easy to find.

To exploit unlabeled data in supervised learning, we take a somewhat different approach. Rather than augmenting labeled data, we consider the problem of *feature vector* generation for linear classifiers. In the training phase, a *linear classifier* is given some number of data points and their class labels, from which it learns a weight vector and a threshold typically by minimizing *empirical risk*[1]. The label of any new data point $\vec{\phi}$ is predicted by thresholding its inner product with the weight vector. The data points are also called *feature vectors* – vectors that represent the features (or attributes) of real-world objects (e.g., words, documents, images, etc.) to be classified.

What sort of feature vectors would require a smaller amount of labeled training data? Since a linear classifier predicts labels by thresholding inner products, feature vectors close to each other (i.e., having larger inner products) tend to produce the same prediction. Thus, it is intuitive that if vectors for objects with the same label (with different labels) are close to (far from) each other (respectively), then relatively small training data can ensure low generalization error. Moreover, consider an extreme case where feature vectors were exactly the same if their labels were the same, and they were orthogonal otherwise. Then, we would need labeled examples only one for each class to achieve perfect classification[2] for any objects.

[1] The *empirical risk* (also called *training error*) is the label prediction error on the training data.
[2] Such vector representation is sufficient for zero expected error, but it is not necessary.

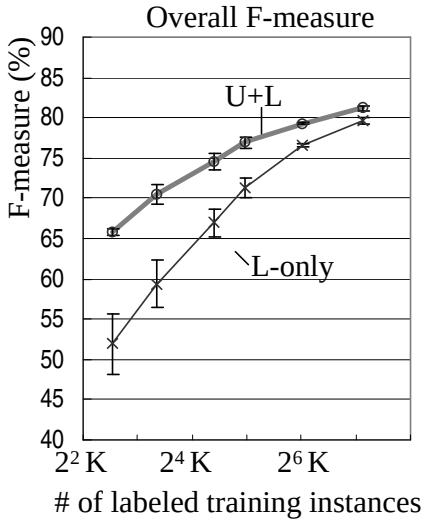

Figure 1: EMR results. Overall F-measure.
L-only: Baseline, trained by labeled data only, U+L: Feature vectors generated by our method. Vertical bars show the ranges of one standard deviation over three runs.

We start with examining the notion of how well one could guess from an arbitrary *view* (viewpoint) ψ whether two objects have the same label or not, *if* ψ's *unknown* distributional relation to labels were precisely known. Based on this analysis, we propose a novel method for feature vector generation using unlabeled data.

The proposed method achieves significant performance improvement over a baseline (trained with labeled data only) on two NLP tasks: part-of-speech tagging (POS) and entity mention recognition (EMR). For instance, when training data is miniscule (less than 4K words; $< 0.5\%$ of typical size) our method still achieves 85% accuracy on the POS task, whereas the baseline degrades to 73%. In our experimental setting, our method requires less than one fourth (POS) or one half (EMR) of the labeled data used by the baseline method to exceed its performance.

The labeled data for EMR experiments consists of 190K words, manually annotated following the ACE Entity Detection Annotation Guideline. There are 15 classes – obtained by combining five entity types (Person, Organization, Facility, GPE, Location) and three mention types (Name, Nominal, Pronoun). For instance, "Mike Smith saw his friend." contains the following three mentions: "Mike Smith" (Person-Name), "his" (Person-Pronoun), and "friend" (Person-Nominal). When EMR is recast to a token classification problem, the number of classes becomes $31 (= 15 \times 2 + 1)$ – multiplied by 2 for encoding the beginning and continuation of the entity mention, and one is added for the token outside of any mention chunk (e.g., "saw" in the example sentence above). We use three-months of AP newswire articles as our unlabeled data. Figure 1 plots the overall F-measure of the baseline ('L-only') and our method ('U+L') in relation to the number of labeled training examples. L-only and U+L use exactly the same linear classifier and the same sequential labeling algorithm. The only difference between them is that L-only was trained with labeled data only, and the feature vectors for U+L were generated by our method using unlabeled data and labeled data. We confirm that significant performance improvement is gained by our method.

Our approach is non-parametric, thus, there is no need for making task-specific model assumptions. This is in contrast to prevailing generative approaches such as *Fisher kernel* (Jaakkola and Haussler, 1998) and approaches employing *Expectation Maximization (EM)* (e.g., Nigam et al. (2000)), where unlabeled data could be rather harmful if it violated model assumptions. The method we propose is generally applicable as long as weakly dependent views of features can be designed. As we will see from our experiments, such views are

intrinsic in many NLP tasks. Moreover, unlike co-training, each view is not required to be sufficient by itself. Hence, we presume that a wide range of NLP tasks may potentially benefit from this method.

We will describe our analysis and method in a forthcoming full paper.

References

Steven Abney. 2002. Bootstrapping. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unalbeled data with co-training. In *Proceedings of the Conference on Computational Learning Theory (COLT-98)*.

Tommi S. Jaakkola and David Haussler. 1998. Exploiting generative models in discriminative classifiers. *Advances in Neural Information Processing Systems*, 11.

J. Krogh and A. Vedelsby. 1994. Neural network ensembles, cross validation, and active learning. In *NIPS'94*, pages 231–238.

Andrew McCallum and Kamal Nigam. 1998. Employing EM in pool-based active learning for text classification. In *Proceedings of the Fifteenth International Conference of Machine Learning (ICML 98)*, pages 350–358.

Tom Mitchell. 1999. The role of unlabeled data in supervised learning. In *Proceedings of the Sixth International Colloquium on Cognitive Science*.

Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134.

David Pierce and Claire Cardie. 2001. Limitations of co-training for natural language learning from large datasets. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*.