Agreement/Disagreement Classification: Exploiting Unlabeled Data using Contrast Classifiers

Sangyun Hahn **Richard Ladner**
Dept. of Computer Science and Engineering
University of Washington, Seattle, WA
{syhahn,ladner}@cs.washington.edu

Mari Ostendorf
Dept. of Electrical Engineering
University of Washington, Seattle, WA
mo@ee.washington.edu

Abstract

Several semi-supervised learning methods have been proposed to leverage unlabeled data, but imbalanced class distributions in the data set can hurt the performance of most algorithms. In this paper, we adapt the new approach of contrast classifiers for semi-supervised learning. This enables us to exploit large amounts of unlabeled data with a skewed distribution. In experiments on a speech act (agreement/disagreement) classification problem, we achieve better results than other semi-supervised methods. We also obtain performance comparable to the best results reported so far on this task and outperform systems with equivalent feature sets.

1 Introduction

In natural language understanding research with data-driven techniques, data labeling is an essential but time-consuming and costly process. To alleviate this effort, various semi-supervised learning algorithms such as self-training (Yarowsky, 1995), co-training (Blum and Mitchell, 1998; Goldman and Zhou, 2000), transductive SVM (Joachims, 1999) and many others have been proposed and successfully applied under different assumptions and settings. They all aim to improve classification accuracy by exploiting more readily available unlabeled data as well as labeled examples. However, these iterative training methods have shortcomings when trained on data with imbalanced class distributions. One reason is that most classifiers underlying these methods assume a balanced training set, and thus when one of the classes has a much larger number of examples than the other classes, the trained classifier will be biased toward the majority class. The imbalance will propagate through subsequent iterations, resulting in a more skewed data set upon which a further biased classifier will be trained. To exploit unlabeled data in learning an inherently skewed data distribution, we introduce a semi-supervised classification method using contrast classifiers, first proposed by Peng *et al.* (Peng et al., 2003). It approximates the posterior class probability given an observation using class-specific contrast classifiers that implicitly model the difference between the distribution of labeled data for that class and the unlabeled data.

In this paper, we will explore the applicability of contrast classifiers to the problem of semi-supervised learning for identifying agreements and disagreements in multi-party conversational speech. These labels represent a simple type of "speech act" that can be important for understanding the interaction between speakers, or for automatically summarizing or browsing the contents of a meeting. This problem was previously studied (Hillard et al., 2003; Galley et al., 2004), using a subset of ICSI meeting recording corpus (Janin et al., 2003). In semi-supervised learning, there is a challenge due to an imbalanced class distribution: over 60% of the data are associated with the default class and only 5% are with disagreements.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 53–56,
New York, June 2006. ©2006 Association for Computational Linguistics

2 Contrast Classifier

The contrast classifier approach was developed by Peng *et al* and successfully applied to the problem of identifying protein disorder in a protein structure database (outlier detection) and to finding articles about them (single-class detection) (Peng et al., 2003). A contrast classifier discriminates between the labeled and unlabeled data, and can be used to approximate the posterior class probability of a given data instance as follows. Taking a Bayesian approach, a contrast classifier for the j-th class is defined as:

$$cc_j(x) = \frac{r_j g(x)}{(1 - r_j)h_j(x) + r_j g(x)} \tag{1}$$

where $h_j(x)$ is the likelihood of x generated by class j in the labeled data, $g(x)$ is the distribution of unlabeled data, and r_j is the relative proportion of unlabeled data compared to the labeled data for class j. This discriminates the class j in the labeled data from the unlabeled data. Here, we constrain $r_j = 0.5$ for all j, using resampling to address class distribution skew, as described below. Rewriting equation 1, $h_j(x)$ can be expressed in terms of $cc_j(x)$ as:

$$h_j(x) = \frac{1 - cc_j(x)}{cc_j(x)} \cdot \frac{r}{1 - r} \cdot g(x). \tag{2}$$

Then, the posterior probability of an input x for class j, $p(j|x)$, can be approximated as:

$$p(j|x) = \frac{h_j(x)q_j}{\sum_i h_i(x)q_i} \tag{3}$$

where q_j is the prior class probability which can be approximated by the fraction of instances in the class j among the labeled data. By substituting eq. 2 into eq. 3, we obtain:

$$p(j|x) = \frac{q_j \cdot (1 - cc_j(x))/cc_j(x)}{\sum_i q_i \cdot (1 - cc_i(x))/cc_i(x)}. \tag{4}$$

Notice that we do not have to explicitly estimate $g(x)$. Eq. 4 can be used to construct the MAP classifier:

$$\hat{c} = \arg\max_j \frac{1 - cc_j(x)}{cc_j(x)} \cdot q_j \tag{5}$$

To approximate the class-specific contrast classifier, $cc_j(x)$, we can choose any classifier that outputs a probability, such as a neural net, logistic regression, or an SVM with outputs calibrated to produce a reasonable probability.

Typically a lot more unlabeled data are available than labeled data, which causes class imbalance when training a contrast classifier. In a supervised setting, a resampling technique is often used to reduce the effect of imbalanced data. Here, we use a committee of classifiers, each of which is trained on a balanced training set sampled from each class. To compute the final output of the classifier, we implemented four different strategies.

- For each class, average the outputs of the contrast classifiers in the committee, and use the average as $cc_j(x)$ in eq. 5.

- Average only the outputs of contrast classifiers smaller than their corresponding threshold, and the fraction of the included classifiers is used as the strength of the probability output for the class.

- Use a meta classifier whose inputs are the outputs of the contrast classifiers in the committee for a class, and whose output is modeled by training it from a separate, randomly sampled data set. The output of the meta classifier is used as $cc_j(x)$.

- Classify an input as the majority class only when the outputs of the meta classifiers for the other classes are all larger than their corresponding thresholds.

Another benefit of the contrast classifier approach is that it is less affected by imbalanced data. When training the contrast classifier for each class, it uses the instances in only one class in the labeled data, and implicitly models the data distribution within that class independently of other classes. That is, given a data instance, the distribution within a class, $h_j(x)$, determines the output of the contrast classifier for the class (eq. 1), which in turn determines the posterior probability (eq. 4). Thus it will not be as highly biased toward the majority class as a classifier trained with a collection of data from imbalanced classes. Our experimental results presented in the next section confirm this benefit.

3 Experiments

We conducted experiments to answer the following questions. First, is the contrast classifier approach applicable to language processing problems, which often involve large amounts of unlabeled data? Second, does it outperform other semi-supervised learning methods on a skewed data set?

3.1 Features and data sets

The data set used consists of seven transcripts out of 75 meeting transcripts included in the ICSI meeting corpus (Janin et al., 2003). For the study, 7 meetings were segmented into spurts, defined as a chunk of speech of a speaker containing no longer than 0.5 second pause. The first 450 spurts in each of four meetings were hand-labeled as either *positive* (agreement, 9%), *negative* (disagreement, 6%), *backchannel* (23%) or *other* (62%).

To approximate $cc_j(x)$ we use a Support Vector Machine (SVM) that outputs the probability of the positive class given an instance (Lin et al., 2003). We use only word-based features similar to those used in (Hillard et al., 2003), which include the number of words in a spurt, the number of keywords associated with the *positive* and *negative* classes, and classification based on keywords. We also obtain word and class-based bigram language models for each class from the training data, and compute such language model features as the perplexity of a spurt, probability of the spurt, and the probability of the first two words in a spurt, using each language model. We also include the most likely class by the language models as features.

3.2 Results

First, we performed the same experiment as in (Hillard et al., 2003) and (Galley et al., 2004), using the contrast classifier (CC) method . Among the four meetings, the data from one meeting was set aside for testing. Table 1 compares the 3-class accuracy of the contrast classifier with previous results, merging *positive* and *backchannel* class together into one class as in the other work. When only lexical features are used (the first three entries), the SVM-based contrast classifier using meta-classifiers gives the best performance, outperforming the decision tree in (Hillard et al., 2003) and the maximum entropy model in (Galley et al., 2004). It also outperformed the SVM trained using the labeled data only. The contrast classifier is also competitive with the best case result in (Galley et al., 2004) (last entry), which adds speaker change, segment duration, and adjacency pair sequence dependency features using a dynamic Bayesian network.

Table 1: Comparison of 3-way classification accuracy on lexical (lex) vs. expanded (exp) features sets.

	Accuracy
Hillard-lex	82
Galley-lex	85.0
SVM-lex	86.3
CC-lex	86.7
Galley-exp	86.9

Table 2: Comparison of the classification performance

Method	3-way Acc	A/D confusion	A/D recovery
unsupervised	79	8	83
cc	81.4	4	82.4
cc-threshold	76.7	6	85.2
cc-meta	86.7	5	81.3
cc-meta-thres	87.1	5	82.4

In table 2, we report the performance of the four classification strategies described in section 2. For comparison, we include a result from Hillard, obtained by training a decision tree on the labels produced by their unsupervised clustering technique. Meta classifiers usually obtained higher accuracy, but averaging often achieved higher recovery of agreement/disagreement (A/D) spurts. The use of thresholds increases A/D recovery, with a decrease in accuracy. We obtained the best accuracy using both meta classifiers and thresholds together here, but we more often obtained higher accuracy using meta classifiers only.

Next, we performed experiments on the entire ICSI meeting data. Only 1,318 spurts were labeled, and 62,944 spurts were unlabeled. Again, one of the labeled meeting transcripts was set aside as a test set. We compared the SVM trained only on labeled data

Table 3: Classification performance, training on the entire ICSI data set. F is defined as $\frac{2pr}{p+r}$ where p is macro precision and r is the macro recall.

Method	Acc	F	Neg recall
SVM	85.4	72.6	21.1
self-training	80.4	65.3	5.2
cotraining	85.1	73.8	47.4
cc	83.0	75.5	68.5

with three semi-supervised methods: self-training, co-training, and the contrast classifier with a meta-classifier. The self-training iteratively trained an SVM with additional data labeled with confidence by the previously trained SVM. For the co-training, each of an SVM and a multilayer backpropagation network was trained on the labeled data and the unlabeled data classified with high confidence (99%) by one classifier were used as labeled data for further training the other classifier. We used two different classifiers, instead of two independent view of the input features as in (Goldman and Zhou, 2000). Table 3 shows that the SVM obtained high accuracy, but the F measure and the recall of the smallest class, *negative*, is quite low. The bias toward the majority class propagates through each iteration in self-training, so that only 5% of the *negative* tokens were detected after 30 iterations. We observed the same pattern in co-training; its accuracy peaked after two iterations (85.1%) and then performance degraded drastically (68% after five iterations) due in part to an increase in mislabeled data in the training set (as previously observed in (Pierce and Cardie, 2001)) and in part because the data skew is not controlled for. The contrast classifier performs better than the others in both F measure and *negative* class recall, retaining reasonably good accuracy.

4 Conclusion

In summary, our experiments on agreement/disagreement detection show that semi-supervised learning using contrast classifiers is an effective method for taking advantage of a large unlabeled data set for a problem with imbalanced classes. The contrast classifier approach outperforms co-training and self-training in detecting the infrequent classes. We also obtain good performance relative to other methods using simple lexical features and performance comparable to the best result reported.

The experiments here kept the feature set fixed, but results of (Galley et al., 2004) suggest that further gains can be achieved by augmenting the feature set. In addition, it is important to assess the impact of semi-supervised training with recognizer output, where gains from using unlabeled data may be greater than with reference transcripts as in (Hillard et al., 2003).

References

A. Blum and T. Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proc. Conference on Computational Learning Theory (COLT-98)*, pages 92–100.

M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg. 2004. Identifying agreement and disagreement in conversational speech: use of Bayesian networks to model dependencies. In *Proc. ACL.*

S. Goldman and Y. Zhou. 2000. Enhancing supervised learning with unlabeled data. In *Proc. the 17th ICML*, pages 327–334.

D. Hillard, M. Ostendorf, and E. Shriberg. 2003. Detection of agreement vs. disagreement in meetings: training with unlabeled data. In *Proc. HLT-NAACL.*

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI meeting corpus. In *ICASSP-03.*

T. Joachims. 1999. Transductive inference for text classification using support vector machines. In *Proc. ICML*, pages 200–209.

H. T. Lin, C. J. Lin, and R. C. Weng. 2003. A note on platt's probabilistic outputs for support vector machines. Technical report, Dept. of Computer Science, National Taiwan University.

K. Peng, S. Vucetic, B. Han, H. Xie, and Z. Obradovic. 2003. Exploiting unlabeled data for improving accuracy of predictive data mining. In *ICDM*, pages 267–274.

D. Pierce and C. Cardie. 2001. Limitations of co-training for natural language learning from large datasets. In *Proc. EMNLP-2001).*

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. ACL*, pages 189–196.