CDL Relation Classification Using Co-Training Style Algorithm

Haibo Li Yutaka Matsuo Mitsuru Ishizuka

Department of Information and Communication Engineering
Graduate School of Information Science and Technology
University of Tokyo

In many machine learning applications, labeled data are insufficient; unlabeled data are easier to come by. Semi-supervised machine learning addresses this problem by combining the labeled data and a large amount of unlabeled data for learning. As described herein, we investigate co-training algorithm, a semi-supervised learning algorithm, for semantic relation classification task. Co-training algorithm splits all features into two views and trains classifiers by the labeled seeds in each view. Each classifier classifies the unlabeled data in the unlabeled data pool and provides the other classifier with a few unlabeled examples as training seeds that receive the highest confidence from the first classifier. We evaluate the co-training algorithm on *Concept Description Language for Natural Language* (CDL. nl) corpus for relation classification task. Experiment results show that co-training algorithm achieves better performance than Naive Bayes that treat all features as a single view, when only very few labeled data are available.

1. Introduction

Many tasks of machine learning have a feature that the data are naturally consist of several views—disjoint subsets of features. For instance, web pages can be described by their contents or hyperlinks pointing to these pages [Blum 1998]; the semantic role of phrases can be decided by the headwords and the paths in parsing tree [He 2004]. A popular paradigm of multi-view learning is the co-training algorithm, which splits all features into two subsets and trains classifiers by the labeled seeds in each view. Each classifier classifies the unlabeled data in the unlabeled data pool and provides the other classifier with a few unlabeled examples as training seeds that receive the highest confidence from the first classifier.

Semantic Relation classification is a basic problem in Natural Language Understanding and semantic processing. Moreover, Semantic Relation Classification is also a problem in which datasets can be naturally split into two views. This task can be represented as follows:

$$\mathrm{R} \rightarrow (C_{\mathrm{pre}}, n_1, C_{\mathrm{mid}}, n_2, C_{\mathrm{post}})$$

where n_1 and n_2 are nouns or base noun phrases and C_{pre}, C_{mid}, and C_{post} are the contexts before, between, and after the concept pairs. Usually, research set the mid-context window as all the words between n_1, n_2 and the pre-context and post-context window as up to two words before n_1 and after n_2 [Chen 2006].

In this paper, we evaluate the co-training algorithm for semantic relation classification task on the CDL. nl corpus. The experiments validate the effectiveness of co-training algorithm.

2. Related Work and Background

2.1 Co-Training

Many studies described in the literature of information extraction and text understanding show that properly combining information from different views can gain leverage from natural redundancy in data. Co-training outperforms EM-based algorithms using unlabeled data when the feature set is divisible into two independent and redundant sub-feature sets [Nigam 2000]. A named entity classification algorithm proposed in [Collins 1999], which is based on co-training framework, can reduce the need for supervision to a handful of seed rules. Ghani et al. developed a multi-class classification framework in the ECOC setup; the algorithm achieved both good accuracy and a good precision–recall tradeoff [Ghani 2002]. Figure 1 presents the co-training algorithm proposed in [Blum 1998].

Given:
-a set L of labeled training examples
-a set S of unlabeled data
Create a pool U' of examples by choosing u examples randomly from S
Loop for k iterations:
1. Use L to train a classifier h_1 using only the x_1 portion of x;
2. Use L to train a classifier h_2 using only the x_2 portion of x;
3. Allow h_1 to label p positive and n negative examples from U';
4. Allow h_2 to label p positive and n negative examples from U';
5. Add these self-labeled examples to L;
6. Search $2p + 2n$ examples using S to replenish U'

Figure 1. Co-Training Algorithm

2.2 CDL. nl Relation Set

Concept Description Language for Natural Language (CDL.nl) presented in [Yokoi 2005] is intended to describe the concept structure of text using a set of pre-defined semantic relations. Furthermore, CDL.nl defines a set of semantic relations to form the semantic structure of natural language sentences in a graphical representation.

CDL. nl relation set contains 44 semantic relationships which are used to add a layer of semantic annotation on natural language sentences. Different from PropBank which depends on verbs and usage of verbs, these predefined neural semantic relations cover different types of predicates.

contact: Haibo Li, University of Tokyo, 7-3-1 Hongo Bunkyo-ku Tokyo , 03-58416774 , Fax:03-58418570 , lihaibo@mi.ci.i.u-tokyo.ac.jp

3. Experiments and Results

In this section, we present our empirical study using CDL. nl corpus. This corpus consists of 1759 sentences and each sentence is marked with the CDL. nl relations. We extract 15697 relation instances from this corpus. Since some semantic relations do not frequently appear in the corpus, we only select 11 relation types that have more than 100 instances in the corpus. Table 1 presents the number of examples of each selected relation. We randomly sample 40% of the selected 11 types of relation as test set and other 60% of data are used as training set and unlabeled data. In this experiment, we randomly select different percentages of instances in the left 60% of data as seeds and other remained instances are treated as unlabeled data. The "All Instance" column of Table 1 presents all instances amount of each relation type. "Test Set" shows the amount of test set that is selected from each relation type.

Table 1. CDL. nl Dataset Statistics

Relation Type	*All Instance*	*Test Set*
agt	1191	476
and	1283	513
aoj	2364	946
gol	446	178
man	912	365
mod	3694	1478
obj	3129	1252
plc	584	234
pur	350	140
qua	317	127
tim	384	154

3.1 Features

Following [Chen 2006], we use lexical and syntactic features of the contexts and concept pairs, which are extracted from CDL. nl corpus.

- **Words:** Surface tokens of the two concepts and words in the three contexts.
- **POS features:** Part-Of-Speech tags of all tokens in the two concepts and words in the three contexts.
- **Position features:**
  1) WBNULL: no words between the concept pair
  2) WBO1: the only word in between the concept pair
  3) WBF, WBL, WBO: the first word, the last word and other words between the concept pair, when there are at least two words between the concept pair
  4) WBF1, WBF2: the first word, the second word before n_1
  5) WAL1, WAL2: the first word, the second word after n_2

We split the feature into two views: *concept pair* and *context*. Two Naive Bayes classifiers are trained on each view respectively. During the training iteration, one classifier provides another classifier self-labeled data as the training set of next round. In our experiment, the training iteration is repeated 10 times. Particularly, in each iteration, the top 0.5% self-labeled data receiving the highest confidence are added into the training set for next round of training.

3.2 Results

Figure 2 shows the results of experiment. Co-Training algorithm (Combine A&B) is compared with the algorithms: 1) the classifier trained on *concept pair* view (A classifier); 2) the classifier trained on *context* view (B classifier); 3) the Naive Bayes classifier treating all features as one view;

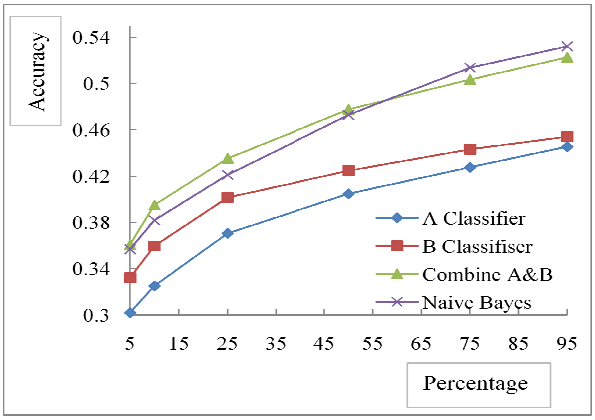

Figure 2. Accuracy on Different Percent of Labeled Seeds

We can observe from Figure 2 that, when the labeled seed less than 50% the co-training outperform all three classifiers. When we randomly label more than 50 percentages of data as seeds, the co-training algorithm cannot beat the Naive Bayes classifier.

4. Conclusion

This paper approaches the problem of semi-supervised relation classification using the co-training algorithm. Experiment results show that when only very few labeled examples are available, co-training algorithm can achieve better performance than Naive Bayes which regards all feature as one view. And also outperforms the two classifiers trained on each view.

References

[Blum 1998] Blum, A, Mitchell, T., Combining Labeled and Unlabeled Data with Co-Training, COLT'98, 1998

[He 2004] He, S., Gildea, D., Self-training and Co-Training for Semantic Role Labeling, University of Rochester technical report 891, 2004

[Chen 2006] Chen, J., Ji, D., Tan, C., Niu, Z, Semi-supervised Relation Extraction with Label Propagation, HLT-NAACL, 2006

[Nigam 2000] Nigam, K., & Ghani, R., Analyzing the Effectiveness and Applicability of Co-training, CIKM-2000, 2000

[Colins 1999] Collins, M., Singer, Y., Unsupervised Models for Named Entity Classification, EMNLP/VLC-99, 1999

[Ghani 2002] Ghani, R., Combining Labeled and Unlabeled Data for MultiClass Text Categorization, ICML-02, 2002

[Dasgupta 2001] Dasgupta, S., Littman, M. L., McAllester, D., PAC Generalization Bounds for Co-training, NIPS, 2001

[Balcan 2005] Balcan, M.-F., Blum, A., Yang, K., Co-training and Expansion: Towards Bridging Theory and Practice, NIPS, 2005

[Yokoi 2005] Yokoi. T., H. Yasuhara, H. Uchida, CDL (Concept Description Language): A Common Language for Semantic Computing , WWW2005(SeC2005), 2005