Abstract. Semi-supervised learning has attracted much attention over the past decade because it provides the advantage of combining unlabeled data with labeled data to improve the learning capability of models. Co-training is a representative paradigm of semi-supervised learning methods. Typically, some co-training style algorithms, such as co-training and co-EM, learn two classifiers based on two views of the instance space. But they have to satisfy the assumptions that these two views are sufficient and conditionally independent given the class labels. Other co-training style algorithms, such as multiple-learner, use two different underlying classifiers based on only a single view of the instance space. However, they could not utilize the labeled data effectively, and suffer from the early convergence. After analyzing various co-training style algorithms, we have found that all of these algorithms have symmetrical framework structures that are related to their constraints. In this paper, we propose a novel unsymmetrical-style method, which we call the unsymmetrical co-training algorithm. The unsymmetrical co-training algorithm combines the advantages of other co-training style algorithms and overcomes their disadvantages. Within our unsymmetrical structure, we apply two unsymmetrical classifiers, namely, the self-training classifier and the EM classifier, and then train these two classifiers in an unsymmetrical way. The unsymmetrical co-training algorithm not only avoids the constraint of the conditional independence assumption, but also overcomes the flaws of the early convergence and the ineffective utilization of labeled data. We conduct experiments to compare the performances of these co-training style algorithms. From the experimental results, we can see that the unsymmetrical co-training algorithm outperforms other co-training algorithms.

1 Introduction

Over the course of the past decade, researchers have developed various types of semi-supervised learning methods. Co-training is a representative paradigm of semi-supervised learning methods that are based on the multiple representations from different views. Co-training was inspired by the observation discovered in the Web pages classification, in which a Web page has two different representations (views): the words occurring on the page itself; and the words...
contained in the anchor text of hyperlinks pointing to the page. The initial form of co-training is to train two classifiers separately on two sufficient and redundant views of data, and let these two classifiers label some unlabeled instances for each other. Like other semi-supervised learning methods, co-training requires its own assumptions to guarantee its success. Blum and Mitchell [1] proved that co-training can be successful if the two sufficient and redundant views are conditionally independent given the class label. Many researchers have supported the observation that co-training is sensitive to this theoretical assumption [18] [2]. However, the sufficient and redundant views are rarely found in most real-world application scenarios.

In order to tease out the effect of view-splitting from the effect of labeling, Nigam and Ghani [2] proposed a hybrid algorithm of expectation-maximization (EM) and co-training, called co-EM. Like co-training, co-EM tries to divide the instance space into two conditional independent views, and to train two EM classifiers based on these two views, respectively. But unlike co-training, co-EM uses all the unlabeled data every time, instead of incrementally selecting some confident predictions to update the training set. Nigam and Ghani [2] also provided a method for ideally splitting the view of instance space based on the conditional mutual information criteria between two subsets of attributes. However, this method is NP-hard and difficult to apply in practice.

Since both co-training and co-EM suffer from the conditional independence assumption, variants of co-training have been developed based on only a single view (without splitting the attribute set). For example, Goldman and Zhou [3] used two different learning algorithms in the paradigm of co-training without splitting the attribute set. Steedman et al. [4] developed a similar co-training algorithm that applies two diverse statistical parsers. Wang and Zhou [5] proved that if the two classifiers are largely diverse, co-training style algorithms are able to succeed. Because these variants substitute multiple views by multiple classifiers, these algorithms are referred to as multiple-learner algorithms. Since the multiple-learner algorithms are trained on the same attribute set, it is important to keep the two classifiers different during the process in order to prevent early convergence. Maintaining separated training sets is one approach for this purpose. However, assigning labeled instances to two different initial training sets will cause the ineffective utilization of labeled data sets in a semi-supervised learning scenario.

Considering the framework structures of co-training, co-EM, and multiple-learner algorithms, we can see that each structure is symmetrical. The co-training algorithm splits the instance space into two symmetrical views, trains two classifiers symmetrically, and lets two classifiers teach each other in a symmetrical way. Similarly, the co-EM algorithm sets up two symmetrical EM classifiers based on their related views. And likewise, the multiple-learner algorithm also has a symmetrical structure, where two classifiers are trained in parallel and combined together to score the unlabeled instances. Therefore, we define these algorithms as the symmetrical-style co-training algorithms.