Theoretical Computer Science 519 (2014) 68–87

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Supervised learning and Co-training ☆

Malte Darnstädt [a], Hans Ulrich Simon [a,*], Balázs Szörényi [b,c]

[a] Fakultät für Mathematik, Ruhr-Universität Bochum, D-44780 Bochum, Germany
[b] Hungarian Academy of Sciences and University of Szeged, Research Group on Artificial Intelligence, H-6720 Szeged, Hungary
[c] INRIA Lille, SequeL project, F-59650 Villeneuve d'Ascq, France

ARTICLE INFO

Keywords:
PAC-learning
Co-training
Supervised learning

ABSTRACT

Co-training under the Conditional Independence Assumption is among the models which demonstrate how radically the need for labeled data can be reduced if a huge amount of unlabeled data is available. In this paper, we explore how much credit for this saving must be assigned solely to the extra assumptions underlying the Co-training Model. To this end, we compute general (almost tight) upper and lower bounds on the sample size needed to achieve the success criterion of PAC-learning in the realizable case within the model of Co-training under the Conditional Independence Assumption in a purely supervised setting. The upper bounds lie significantly below the lower bounds for PAC-learning without Co-training. Thus, Co-training saves labeled data even when not combined with unlabeled data. On the other hand, the saving is much less radical than the known savings in the semi-supervised setting.

1. Introduction

In the framework of semi-supervised learning, it is usually assumed that there is a kind of compatibility between the target concept and the domain distribution.[1] This intuition is supported by recent results indicating that, without extra assumptions, there exist purely supervised learning strategies which can compete fairly well against semi-supervised learners (or even against learners with full prior knowledge of the domain distribution) [3,9].

In this paper, we go one step further and consider the following general question: given a particular extra assumption which makes semi-supervised learning quite effective, how much credit must be given to the extra assumption alone? In other words, to which extent can labeled examples be saved by exploiting the extra assumption in a purely supervised setting? We provide a first answer to this question in a case study which is concerned with the model of Co-training under the Conditional Independence Assumption [5].

1.1. Related work

Supervised and semi-supervised learning. In the semi-supervised learning framework the learner is assumed to have access both to labeled and unlabeled data. The former is supposed to be expensive and the latter to be cheap, thus unlabeled data should be used to minimize the amount of labeled data required. Indeed, a large set of unlabeled data provides extra information about the underlying distribution.

☆ This work was supported by the bilateral Research Support Programme between Germany (DAAD 50751924) and Hungary (MÖB 14440).
* Corresponding author.
E-mail address: hans.simon@rub.de (H.U. Simon).
[1] See the introduction of [8] for a discussion of the most popular assumptions.

0304-3975/$ – see front matter
http://dx.doi.org/10.1016/j.tcs.2013.09.020

Already in 1991, Benedek and Itai [4] studied learning under a fixed distribution, which can be seen as an extreme case of semi-supervised learning, where the learner has full knowledge of the underlying distribution. They derive upper and lower bounds on the number of required labels based on ε-covers and -packings. Later in 2005, Kääriäinen [13] developed a semi-supervised learning strategy, which can save up to one half of the required labels. These results don't make use of extra assumptions that relate the target concept to the data distribution.

However, some recent results by Ben-David et al. in [3] and later by Darnstädt and Simon in [9] indicate that even knowing the data distribution perfectly does not help the learner for most distributions asymptotically, i.e. a reduction by a constant factor is the best possible. In fact, they conjecture a general negative result, which is nonetheless still absent. These results can be regarded as a justification of using extra assumptions in the semi-supervised framework in order to make real use of having access to unlabeled data.

Our work provides a similar analysis of these assumptions in the fashion of the above results: we investigate to what extent does such an assumption (Co-training with the Conditional Independence Assumption) alone help the learner, and how much is to be credited to having perfect knowledge about the underlying distribution.

Likewise, a study for the popular Cluster Assumption was done by Singh, Nowak and Zhu in [15]. They show that the value of unlabeled data under their formalized Cluster Assumption varies with the minimal margin between clusters.

Co-training and the Conditional Independence Assumption. The Co-training Model was introduced by Blum and Mitchell in [5], and has an extensive literature in the semi-supervised setting, especially from an empirical and practical point of view. (For the formal definition see Section 2.) A theoretical analysis of Co-training under the Conditional Independence Assumption [5], and the weaker α-expanding Assumption [2], was accomplished by Balcan and Blum in [1]. They work in Valiant's model of PAC-learning [16] and show that *one labeled example* is enough for achieving the success criterion of PAC-learning provided that there are sufficiently many unlabeled examples.[2]

Our paper complements their results: we also work in the PAC model and prove label complexity bounds, but in our case the learner has no access to unlabeled data. As far as we know, our work is the first that studies Co-training in a fully supervised setting. Assuming Conditional Independence, our label complexity bound is much smaller than the standard PAC bound (which must be solely awarded to Co-training itself), while it is still larger than Balcan and Blum's (which must also be awarded to the use of unlabeled data). See Section 1.2 for more details.

Agnostic active learning. We make extensive use of a suitably defined variant of Hanneke's disagreement coefficient, which was introduced in [11] to analyze agnostic active learning. (See Section 2.2 for a comparison of the two notions.) To our knowledge this is, besides a remark about classical PAC-learning in Hanneke's thesis [12], the first use of the disagreement coefficient outside of agnostic learning. Furthermore, our work doesn't depend on results from the active learning community, which makes the prominent appearance of the disagreement coefficient even more remarkable.

Learning from positive examples only. Another unsuspected connection that emerged from our analysis relates our work to the "learning from positive examples only" model from [10]. As already mentioned, we can upper bound the product of the VC-dimension and the disagreement coefficient by a combinatorial parameter that is strongly connected to Geréb-Graus' "unique negative dimension". Furthermore, we derive worst case lower bounds that make use of this parameter.

1.2. Our main result

Our paper is a continuation of the line of research started out by Ben-David et al. in [3] aiming at investigating the problem: how much can the learner benefit from knowing the underlying distribution. We investigate this problem focusing on a popular assumption in the semi-supervised literature. Our results are purely theoretical, which also stems from the nature of the problem.

As mentioned above, the model of Co-training under the Conditional Independence Assumption was introduced in [5] as a setting where semi-supervised can be superior to fully supervised learning. Indeed, in [1] it was shown that a single labeled example suffices for PAC-learning in the realizable case if unlabeled data is available. Recall that supervised, realizable PAC-learning without any extra assumption requires d/ε labeled samples (up to logarithmic factors) where d denotes the VC-dimension of the concept class and ε is the accuracy parameter [6]. The step from d/ε to just a single labeled example is a giant one. In this paper, we show however that part of the credit must be assigned to just the Co-training itself. More specifically, we show that the number of sample points needed to achieve the success criterion of PAC-learning in the purely supervised model of Co-training under the Conditional Independence Assumption has a linear growth in $\sqrt{d_1 d_2/\varepsilon}$ (up to some hidden logarithmic factors) as far as the dependence on ε and on the VC-dimensions of the two involved concept classes is concerned. Note that, as ε approaches 0, $\sqrt{d_1 d_2/\varepsilon}$ becomes much smaller than the well-known lower bound $\Omega(d/\varepsilon)$ on the number of examples needed by a traditional (not co-trained) PAC-learner.

[2] This is one of the results which impressively demonstrate the striking potential of properly designed semi-supervised learning strategies although the underlying compatibility assumptions are somewhat idealized and therefore not likely to be strictly satisfied in practice. See [2,17] for suggestions of relaxed assumptions.