Pattern Recognition Letters 51 (2015) 37–43

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Domain adaptation of weighted majority votes via perturbed variation-based self-labeling ☆,☆☆

Emilie Morvant*

Laboratoire Hubert Curien, UMR CNRS 5516, Université Jean Monnet de Saint-Etienne, 18 rue Benoit Lauras, Saint-Etienne 42000, France

ARTICLE INFO

Article history:
Received 29 January 2014
Available online 4 October 2014

Keywords:
Machine learning
Classification
Domain adaptation
Majority vote
PAC-Bayes

ABSTRACT

In machine learning, the domain adaptation problem arrives when the test (target) and the train (source) data are generated from different distributions. A key applied issue is thus the design of algorithms able to generalize on a new distribution, for which we have no label information. We focus on learning classification models defined as a weighted majority vote over a set of real-valued functions. In this context, Germain et al. [1] have shown that a measure of disagreement between these functions is crucial to control. The core of this measure is a theoretical bound—the C-bound [2]—which involves the disagreement and leads to a well performing majority vote learning algorithm in usual non-adaptative supervised setting: MinCq.

In this work, we propose a framework to extend MinCq to a domain adaptation scenario. This procedure takes advantage of the recent perturbed variation divergence between distributions proposed by Harel and Mannor [3]. Justified by a theoretical bound on the target risk of the vote, we provide to MinCq a target sample labeled thanks to a perturbed variation-based self-labeling focused on the regions where the source and target marginals appear similar. We also study the influence of our self-labeling, from which we deduce an original process for tuning the hyperparameters. Finally, our framework called PV-MinCq shows very promising results on a rotation and translation synthetic problem.

1. Introduction

Nowadays, due to the expansion of Internet a large amount of data is available. Then, many applications need to make use of supervised machine learning methods able to transfer knowledge from different information sources, which is known as transfer learning.[1] In such a situation, we cannot follow the strong standard assumption in machine learning that supposes the learning and test data drawn from the same unknown distribution. For instance, one of the tasks of the common spam filtering problem consists in adapting a model from one user to a new one who receives significantly different emails. This scenario, called domain adaptation, arises when we aim at learning from a source distribution a well performing model on a different target distribution, for which one considers an unlabeled sample (or few labels).[2] In this paper we design a new domain adaptation framework when we have no target label. This latter situation is known to be challenging [6].

To address this kind of issues, several approaches exist in the literature.[3] Among them, the instance weighting-based methods allow us to deal with the covariate-shift where the distributions differ only in their marginals (*e.g.* [8]). Another technique is to exploit self-labeling procedures. However, it often relies on iterative and heavy self-labeling. For example, one of the reference methods is DASVM [9]. Concretely at each iteration, DASVM learns a SVM classifier from the labeled source examples, then some of them are replaced by target data auto-labeled with this SVM classifier.[4] A third popular solution is to take advantage of a distance between distributions, with the intuition that we want to minimize this divergence while preserving good performance on the source data: If the distributions are close under this measure, then generalization ability may be "easier" to quantify. The most popular divergences, such as the $\mathcal{H}\Delta\mathcal{H}$-divergence of Ben-David et al. [10,11] and the discrepancy of Mansour et al. [12], involve the disagreement between classifiers. Although they lead to different analyses, they enhance to the same conclusion that is the disagreement/diversity between classifiers (from the set of

☆ This paper has been recommended for acceptance by M.A. Girolami.

☆☆ The work of this paper was carried out while E. Morvant was affiliated with Institute of Science and Technology (IST) Austria, Am Campus 1, Klosterneuburg 3400, Austria.

* Tel.: +33477915799.
E-mail address: emilie.morvant@univ-st-etienne.fr, milie.morvant@gmail.com

[1] See Refs. [4,5] for surveys on transfer learning.

[2] The task with few target labels is sometimes referred to as semi-supervised domain adaptation, and the one without target label as unsupervised domain adaptation.

[3] See Ref. [7] for a survey on domain adaptation.

[4] In DASVM, the self-labeled points correspond to those with the lowest confidence, and the deleted source points are those with the highest confidence.

http://dx.doi.org/10.1016/j.patrec.2014.08.013
0167-8655/

possible classifiers) must be controlled while keeping a good source performance. However, these analysis only focus on domain adaptation algorithms that return a single classifier.

In this work, we tackle the issue of learning a majority vote over a set of classifiers or functions in a domain adaptation scenario. A majority vote is an ensemble method[5] where each function is assigned a specific weight. From a theoretical standpoint it is well-known that considering a set of functions with a high diversity is a desirable property [13]. One non-domain adaptation illustration is given by the algorithm AdaBoost of Freund and Schapire [15] that weights weak classifiers according to different distributions of the training data, introducing some diversity. From the theoretical side, the PAC-Bayesian theory [16] offers a nice framework to study majority votes and has been recently extend to domain adaptation [1], which is the first analysis of domain adaptation done for learning target majority votes over a set of functions (or voters).

This analysis stands in the class of approaches based on a divergence between distributions. This latter, called the domain disagreement, has been justified by a tight bound over the risk of the majority vote—the C-bound [2]—and has the advantage to take into account the expectation of the disagreement between pairs of voters. Although their theoretical analysis is elegant and well-founded, the algorithm derived is restricted to linear classifiers. We then intend to design a learning framework able to deal with weighted majority votes over real-valued voters in this PAC-Bayesian domain adaptation scenario. With this aim in mind and knowing the C-bound has lead to a simple and well performing algorithm for supervised classification, called MinCq [17], we extend it to domain adaptation thanks to a non-iterative self-labeling. Firstly, we propose a new formulation of the C-bound suitable for every self-labeling function (which associates a label to an example). Then, we design such a function with the help of a divergence between the marginal distributions called the perturbed variation (PV) [3] and based on the following principle: Two samples are similar if each instance of one sample is close to an instance of the other sample.

Concretely, our PV-based self-labeling focuses on the regions where the source and target marginals are closer, then it labels the (unlabeled) target sample only in these regions (see Fig. 1, in Section 3.2). This self-labeled sample is then provided to MinCq. Afterward, we highlight the influence of our self-labeling, and deduce an original validation procedure. Finally, our framework, named PV-MinCq, implies good and promising results, better than a nearest neighborhood-based self-labeling, and than other domain adaptation methods.

The rest of the paper is organized as follows. Section 2 recalls the PAC-Bayesian domain adaptation setting of Germain et al. [1], and then MinCq and its theoretical basis in the supervised setting [17]. In Section 3 we present PV-MinCq, our adaptive MinCq based on a PV-based self-labeling procedure. Before conclude, we experiment our framework on a synthetic problem in Section 4.

2. Notations and background

In this section, we first review the PAC-Bayesian setting in a non-adaptive setting, and then the results of Germain et al. [1] and Laviolette et al. [17].

2.1. PAC-Bayesian setting in supervised learning

We recall the usual setting of the PAC-Bayesian theory—introduced by McAllester [16]—which offers generalization bounds (and algorithms) for weighted majority votes over a set of real-valued functions, called voters.

[5] See Refs. [13,14] for survey on ensemble method in a non-domain adaptation scenario.

Let $X \subseteq \mathbb{R}^d$ be the input space of dimension d and $Y = \{-1, +1\}$ be the output space, *i.e.* the set of possible labels. P_S is an unknown distribution over $X \times Y$, that we called a domain. $(P_S)^{m_s} = \bigotimes_{s=1}^{m_s} P_S$ stands for the distribution of a m_s-sample. The marginal distribution of P_S over X is denoted by D_S. We consider $S = \{(\mathbf{x}_s, y_s)\}_{s=1}^{m_s}$ a m_s-sample independent and identically distributed (*i.i.d.*) according to $(P_S)^{m_s}$, commonly called the learning sample. Let $\mathcal{H}$ be a set of n (bounded) real-valued voters such that: $\forall h \in \mathcal{H},\ h : X \to \mathbb{R}$. Given $\mathcal{H}$, the ingredients of the PAC-Bayesian approaches are a prior distribution π over $\mathcal{H}$, a learning sample S and a posterior distribution ρ over $\mathcal{H}$. Prior distribution π models an *a priori* belief on what are the best voters from $\mathcal{H}$, before observing the learning sample S. Then, given the information provided by S, the learner aims at finding a posterior distribution ρ leading to a ρ-weighted majority vote B_ρ over $\mathcal{H}$ with nice generalization guarantees. B_ρ and its true and empirical risks are defined as follows.

Definition 1. Let $\mathcal{H}$ be a set of real-valued voters. Let ρ be a distribution over $\mathcal{H}$. The ρ-weighted majority vote B_ρ (sometimes called the Bayes classifier) is:

$$\forall \mathbf{x} \in X,\ B_\rho(\mathbf{x}) = \operatorname{sign}\left[\mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x})\right].$$

The true risk of B_ρ on a domain P_S and its empirical risk[6] on a m_s-sample S are respectively:

$$\mathbf{R}_{P_S}(B_\rho) = \frac{1}{2}\left(1 - \mathop{\mathbf{E}}_{(\mathbf{x}_s,y_s)\sim P} y_s B_\rho(\mathbf{x}_s)\right),$$

$$\mathbf{R}_S(B_\rho) = \frac{1}{2}\left(1 - \frac{1}{m_s}\sum_{s=1}^{m_s} y_s B_\rho(\mathbf{x}_s)\right).$$

Usual PAC-Bayesian analyses[7] do not directly focus on the risk of B_ρ, but bound the risk of the closely related stochastic Gibbs classifier G_ρ. It predicts the label of an example $\mathbf{x}$ by first drawing a classifier h from $\mathcal{H}$ according to ρ, and then it returns $h(\mathbf{x})$. The risk of G_ρ corresponds thus to the expectation of the risks over $\mathcal{H}$ according to ρ:

$$\mathbf{R}_P(G_\rho) = \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{R}_P(h) = \frac{1}{2}\left(1 - \mathop{\mathbf{E}}_{(\mathbf{x}_s,y_s)\sim P}\ \mathop{\mathbf{E}}_{h\sim\rho} y_s h(\mathbf{x}_s)\right). \quad (1)$$

Note that it is well-known in the PAC-Bayesian literature that the deterministic B_ρ and the stochastic G_ρ are related by:

$$\mathbf{R}_P(B_\rho) \le 2\mathbf{R}_P(G_\rho). \quad (2)$$

2.2. PAC-Bayesian domain adaptation of the Gibbs classifier

Throughout the rest of this paper, we consider the PAC-Bayesian domain adaptation setting introduced by Germain et al. [1]. The main difference between supervised learning and domain adaptation is that we have two different domains over $X \times Y$: The source domain P_S and the target domain P_T (D_S and D_T are the respective marginals over X). The aim is then to learn a good model on the target domain P_T knowing that we only have label information from the source domain P_S. Concretely, in the setting described in Ref. [1], we have a labeled source m_s-sample $S = \{(\mathbf{x}_s, y_s)\}_{t=1}^{m_s}$ *i.i.d.* from $(P_S)^{m_s}$ and a target unlabeled m_t-sample $T = \{\mathbf{x}_t\}_{t=1}^{m_t}$ *i.i.d.* from $(D_T)^{m_t}$. One thus desires to learn from S and T a weighted majority vote with lowest possible expected risk on the target domain $\mathbf{R}_{P_T}(B_\rho)$, *i.e.* with good generalization guarantees on P_T. Recalling that usual PAC-Bayesian generalization bound study the risk of the Gibbs classifier, Germain et al. [1] have done an analysis of its target risk $\mathbf{R}_{P_T}(G_\rho)$. Their main result is the following theorem.

[6] We express the risk with the linear loss since we deal with real-valued voters, but in the special case of B_ρ the linear loss is equivalent to the 0–1-loss.

[7] Usual PAC-Bayesian analyses can be found in Refs. [18–22].