Information Processing Letters 112 (2012) 933–936

Contents lists available at SciVerse ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

On multiple-instance learning of halfspaces ☆

D.I. Diochnos [a], R.H. Sloan [a,*], Gy. Turán [a,b]

[a] *University of Illinois at Chicago, United States*

[b] *Research Group on AI, Hungarian Academy of Sciences & University of Szeged, Hungary*

ARTICLE INFO

Article history:
Received 20 August 2011
Received in revised form 2 June 2012
Accepted 28 August 2012
Available online 30 August 2012
Communicated by J. Torán

Keywords:
Machine learning
Halfspaces
Multiple-instance learning
Cyclic polytopes
PAC learning
VC-dimension
Combinatorial problems

ABSTRACT

In multiple-instance learning the learner receives bags, i.e., sets of instances. A bag is labeled positive if it contains a positive example of the target. An $\Omega(d \log r)$ lower bound is given for the VC-dimension of bags of size r for d-dimensional halfspaces and it is shown that the same lower bound holds for halfspaces over any large point set in general position. This lower bound improves an $\Omega(\log r)$ lower bound of Sabato and Tishby, and it is sharp in order of magnitude. We also show that the hypothesis finding problem is NP-complete and formulate several open problems.

1. Introduction

Multiple-instance or multi-instance learning (MIL) is a variant of the standard PAC model of concept learning where, instead of receiving labeled instances as examples, the learner receives labeled bags, i.e., labeled sets of instances. A bag is labeled positive if it contains at least one positive example, and it is labeled negative otherwise. There are different probability models for the distribution of bags; one possible model, which we will call the *independent* model, assumes that instances in a bag are independent and identically distributed. The multi-instance setting, introduced by Dietterich et al. [6], is natural for several learning applications, for example, in drug design and image classification. In drug design, a bag may consist of several shapes of a molecule and it is labeled positive if some shape binds to a specific binding site. In image classification, a bag may be a photo containing several objects and it is labeled positive if it contains some object of interest.

Blum and Kalai [2] showed that every learning problem that is efficiently learnable with statistical queries is also efficiently learnable in the independent MIL model, and, more generally, the same holds for problems efficiently learnable with one-sided random classification noise. Every problem known to be efficiently PAC-learnable is also known to be efficiently learnable with one-sided random classification noise, although no formal relationship is proven so far (see Simon [15] for further discussion of the one-sided random classification noise model). Thus [2] implies the efficient independent MIL–PAC-learnability of all known efficiently PAC-learnable classes.

A detailed study of sample sizes in the MIL model was initiated by Sabato and Tishby [12]. They proved a general upper bound for the VC-dimension of bags, and a lower bound for the concept class of halfspaces. Kundakcioglu et al. [10] considered margin maximization for bags of halfspaces and gave NP-completeness and experimental results.

In this note we continue the study of multi-instance learning of halfspaces. We improve the VC-dimension lower bound of [12] from $\Omega(\log r)$ to $\Omega(d \log r)$, where d

☆ This work was supported by the National Science Foundation under Grant No. CCF-0916708.

* Corresponding author.
E-mail addresses: ddioch2@uic.edu (D.I. Diochnos), sloan@uic.edu (R.H. Sloan), gyt@uic.edu (Gy. Turán).

0020-0190/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ipl.2012.08.017

is the dimension and r is the bag size, which is optimal up to order of magnitude. A similar result was given independently by Sabato and Tishby [13]. We also show that the same lower bound holds for bags over every sufficiently large point set in general position. Thus the situation is somewhat analogous to standard halfspaces, where every simplex forms a maximum shattered set. The proofs are based on cyclic polytopes. We also show that hypothesis finding for bags of halfspaces is NP-complete, using a variant of the construction of [10]. These two results, in view of the well-known relationship between PAC-learnability, VC-dimension and hypothesis finding, indicate differences between the PAC and the independent MIL–PAC models.

There are several open problems related to the multi-instance learning of halfspaces. Some of these are discussed in the concluding section of the paper.

2. Preliminaries

A halfspace in $\mathbf{R}^d$ is given as $H = \{x \in \mathbf{R}^d\colon\ w \cdot x \geqslant t\}$, for weight vector $w \in \mathbf{R}^d$ and threshold $t \in \mathbf{R}$. A bag of size r, or an r-bag, is an r-element multiset $B = \{x_1, \ldots, x_r\}$ in $\mathbf{R}^d$. An r-bag B is positive for H if $B \cap H \neq \emptyset$, and B is negative for H otherwise. A set of bags $\mathcal{B} = \{B_1, \ldots B_s\}$ is shattered by halfspaces if for every ± labeling of the bags there is a halfspace that assigns the same labels to the bags in $\mathcal{B}$. The VC-dimension of r-bags for d-dimensional halfspaces is the largest s such that there are s shattered bags. For $r = 1$ one gets the usual notion of VC-dimension of halfspaces and it is a basic fact that this equals $d + 1$.

3. The VC-dimension of *r*-bags for *d*-dimensional halfspaces

We first formulate a general upper bound of Sabato and Tishby [12], and then we give the matching lower bound for halfspaces. The lower bound is based on properties of cyclic polytopes. The discussion is essentially self-contained as we include a brief overview of the background material (details not given here can be found in Matoušek [11]).

3.1. A general upper bound

Sabato and Tishby [12] showed that the VC-dimension of r-bags for any concept class is essentially at most a $\log r$ factor larger than the VC-dimension of the concept class. We formulate their result in a slightly different form.

Theorem 1. *(See [12].) For any concept class of VC-dimension $\tilde{d}$, the VC-dimension of r-bags is $O(\tilde{d} \log r)$.*

Proof. Let $\mathcal{B} = \{B_1, \ldots, B_s\}$ be a shattered set of r-bags. Then $\mathcal{B}$ contains at most rs instances, and by Sauer's lemma, those can be classified by concepts in the class in at most $((ers)/\tilde{d})^{\tilde{d}}$ many ways. The classification of the instances in the bag determines the classification of the bags. Thus

$$2^s \leqslant \left(\frac{ers}{\tilde{d}}\right)^{\tilde{d}}.$$

Writing $x = s/\tilde{d}$ this becomes $2^x/x \leqslant er$. The function $2^x/x$ is monotone if $x \geqslant 1/\ln 2$. Thus it is sufficient to show that $2^x/x > er$ for $x = \log r + 2 \log\log r$, if r is sufficiently large, which follows directly. □

3.2. Lower bound for halfspaces

Sabato and Tishby showed that the VC-dimension of r-bags of halfspaces in the plane is at least $\lfloor \log r \rfloor + 1$, which implies the same bound for higher dimensions. We now prove a lower bound by adding the 'missing' factor d, which is optimal in order of magnitude by Theorem 1.

The d-dimensional moment curve is given parametrically as $x(t) = (t, t^2, \ldots, t^d)$. The convex hull of points $x(t_1), \ldots, x(t_n)$ on the moment curve, for $t_1 < \cdots < t_n$, with $n \geqslant d + 1$, is called a *cyclic polytope*. For any $I \subseteq [n]$, $|I| \leqslant \lfloor d/2 \rfloor$, the polynomial

$$\prod_{i \in I} (t - t_i)^2 = \sum_{j=0}^{d} w_j t^j$$

is 0 at every t_i, $i \in I$ and positive at every t_i, $i \notin I$. Thus the halfspace $-\sum_{j=1}^{d} w_j x_j \geqslant w_0$ contains every point $x(t_i)$, $i \in I$, and none of the points $x(t_i)$, $i \notin I$. Hence every set of at most $\lfloor d/2 \rfloor$ vertices forms a face of a cyclic polytope.

Theorem 2. *The VC-dimension of d-dimensional halfspaces over bags of size r is at least $\lfloor d/2 \rfloor (\lfloor \log r \rfloor + 1)$.*

Proof. Let ℓ be an integer,

$$s = \left\lfloor \frac{d}{2} \right\rfloor (\ell + 1), \qquad r = 2^\ell, \qquad n = \left\lfloor \frac{d}{2} \right\rfloor \cdot 2^{\ell+1}.$$

Let $t_1 < \cdots < t_n$ be arbitrary and consider the set of n instances $X = \{x(t_1), \ldots, x(t_n)\}$. Divide X into $\lfloor d/2 \rfloor$ blocks of size $2^{\ell+1}$ each, i.e., let

$$X_i = \left\{x(t_j)\colon\ (i-1) \cdot 2^{\ell+1} < j \leqslant i \cdot 2^{\ell+1}\right\},$$
$$i = 1, \ldots, \lfloor d/2 \rfloor.$$

Let f_i be a bijection between X_i and the subsets of integers in the interval $[(i-1) \cdot (\ell+1) + 1, i \cdot (\ell+1)]$ and let

$$B_k = \left\{x(t_j)\colon\ k \in f_i\left(x(t_j)\right)\right\}$$

for every k such that $(i-1) \cdot (\ell+1) < k \leqslant i \cdot (\ell+1)$. We claim that $\{B_1, \ldots, B_s\}$ is a family of bags of size r shattered by d-dimensional halfspaces. Each bag is of size r as it contains a half of a block. For any subset $S \subseteq [s]$ let $S_i = S \cap [(i-1) \cdot (\ell+1) + 1, i \cdot (\ell+1)]$ and let $x(t_{j(i)})$ be the point such that $f_i(x(t_{j(i)})) = S_i$, for $i = 1, \ldots, \lfloor d/2 \rfloor$. Then the set $\{x(t_{j(i)})\colon\ i = 1, \ldots, \lfloor d/2 \rfloor\}$ can be separated from the rest of X by a halfspace, and that halfspace classifies precisely those bags B_k as positive for which $k \in S$. Thus the family of bags is indeed shattered by halfspaces. The VC-dimension bound follows directly from the definition of s and r. □

Now we prove a strengthening of Theorem 2. A finite subset of $\mathbf{R}^d$ is in *general position* if all its $(d+1)$-subsets