On Multiple-Instance Learning of Halfspaces*

D. I. Diochnos[1], R. H. Sloan[1], and Gy. Turán[1,2]

[1]University of Illinois at Chicago ddioch2|sloan|gyt @uic.edu
[2]Hungarian Academy of Sciences & University of Szeged, Hungary

July 12, 2011

Abstract

In multiple-instance learning the learner receives bags, i.e., sets of instances. A bag is labelled 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, discuss the implications of recent active learning results for multi-instance learning 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. Instances in a bag are usually assumed to be independent and identically distributed. This setting, introduced by Dietterich *et al.* [7], is natural for several learning applications, for example, in drug design and image classification.

*This work waswork was supported by the National Science Foundation under Grant No. CCF-0916708. Theorems 2 and 4 of this paper have been presented as an impromptu talk at COLT 2011. After the talk Sivan Sabato informed us that a bound similar to Theorem 2 is included in a submitted article currently under review by a journal, and now available from arXiv [14].

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 [3] showed that every learning problem that is efficiently learnable with statistical queries is also efficiently learnable in the MIL model, and, more generally, the same holds for problems efficiently learnable with one-sided random classification noise. This implies the efficient multi-instance learnability of all known efficiently PAC-learnable classes. A detailed study of sample sizes in the MIL model was initiated Sabato and Tishby [13]. 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.* [11] 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 [13] from $\Omega(\log r)$ to $\Omega(d \log r)$, where d is the dimension and r is the bag size, which is optimal up to order of magnitude. 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 [11]. These two results, in view of the well-known relationship between PAC-learnability, VC-dimension and hypothesis finding, indicate differences between the PAC and MIL-PAC models.

Active learning is another variant of PAC learning. In this model the learner can decide whether to request the label of a random instance, and the complexity of an algorithm is measured by the number of label requests (see, e.g., Dasgupta [6]). Multi-instance active learning (MIAL) has been proposed by Settles *et al.* [15] and has been studied in several machine learning papers. We observe that the general active learning results of Hanneke [10] and Friedman [8] apply to the multi-instance setting as well.

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 $H = \{x \in \mathbf{R}^d \, : \, w \cdot x \geq t\}$, 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 $B = \{x_1, \ldots, x_r\}$ of $\mathbf{R}^d$.

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 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

Sabato and Tishby [13] 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 *[13] 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

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

Writing $x = s/\tilde{d}$ this becomes $2^x/x \leq er$. The function $2^x/x$ is monotone if $x \geq 1/\ln 2$. 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. □

Sabato and Tishby showed that the VC-dimension of r-bags of halfpaces 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 construction uses well-known properties of the d-dimensional *moment curve* (see, e.g., Matousek [12]), which are summarized here for completeness. 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 $t_1 < \ldots < t_n$, with $n \geq d + 1$, is called a *cyclic polytope*. For any

$I \subseteq [n], |I| \leq \lfloor \frac{d}{2} \rfloor$, the polynomial

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

is 0 at every $t_i, i \in I$ and positive at every $t_i, i \notin I$. Thus the halfspace $-\sum_{j=1}^{d} a_j t \geq a_0$ contains every point $x(t_i), i \in I$, and none of the points $x(t_i), i \notin I$. Thus every set of at most $\lfloor \frac{d}{2} \rfloor$ vertices forms a face of a cyclic polytope.

The facets (i.e., $(d-1)$-dimensional faces) of cyclic polytopes are described by *Gale's evenness condition*: for $t_{i_1} < \cdots < t_{i_d}$ the vertices $x(t_{i_1}), \cdots, x(t_{i_d})$ form a facet if and only if for any two other vertices $x(t_u)$ and $x(t_v)$ there are an even number of values t_{i_j} between t_u and t_v. This is proven by considering the hyperplane $\sum_{j=1}^{d} a_j t^j = -a_0$ defined by

$$\prod_{j=1}^{d} (t - t_{i_j}) = \sum_{j=0}^{d} a_j t^j.$$

The condition follows by counting the number of sign changes between t_u and t_v.

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,

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

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

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

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

$$B_k = \{x(t_j) \, : \, k \in f_i(x(t_j))\}$$

for every k such that $(i-1) \cdot (\ell+1) < k \leq i \cdot (\ell+1)\}$. 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 $S \subseteq [1, s]$ 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 $\{x(t_{j(i)}) \ : \ i = 1, \ldots, \lfloor d/2 \rfloor\}$ 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 are affinely independent, i.e., have no linear combination equal to 0, with coefficients adding up to 0. Halfspaces in $\mathbf{R}^d$ shatter *every* simplex, i.e., every set of $(d+1)$ points in general position. In analogy to this fact, we prove a VC-dimension lower bound similar to Theorem 2 for bags of halfspaces when the instances are restricted to *any* sufficiently large subset in general position. The proof uses another property of cyclic polytopes. The following lemma is referred to as "unpublished 'folklore' " and proven in an oriented matroid version by Cordovil and Duchet [5] [1]. It is also given as an exercise in Matousek [12]. Again, we give a simple proof for completeness.

Lemma 3 *(See [5, 12].) There is a function $f(d, n)$ such that every set A of $m \geq f(d, n)$ points in general position in $\mathbf{R}^d$ contains n points such that their convex hull has the same structure as a d-dimensional cyclic polytope on n vertices.*

Proof The orientation of a d-dimensional ordered simplex $(a_0, \ldots, a_d)$ is the sign (+ or -) of the determinant with columns $a_1 - a_0, \ldots, a_d - a_0$, or, equivalently, with columns $a'_0, \ldots, a'_d$, where the primes denote an added first component of 1 to each vector.

Consider a d-dimensional cyclic polytope and let $x(t_{i_1}), \ldots, x(t_{i_{d+1}})$ be $d+1$ vertices of the polytope. The orientation of the simplex formed by these points using the increasing ordering of the parameters is +, as the corresponding determinant is a Vandermonde determinant.

According to Ramsey's theorem (see [9]), there is a function $R(u, v)$ such that if the u-subsets of a set of size at least $R(u, v)$ are two-colored then there is a subset of size v with all its u-subsets colored the same. Put $f(d, n) = R(d+1, n)$ and consider a set A of $m \geq f(d, n)$ points in general position. Fix an arbitrary ordering $<$ of the elements of A. Color each $(d+1)$-subset of A with the orientation (+ or -) of the corresponding simplex, ordered according to the fixed ordering. Then there is a subset $\{a_1, \ldots, a_n\}$ of A such that all ordered simplices from that subset have the same orientation.

[1] The paper is an updated version of an unpublished, but circulated, manuscript from 1986/87.

Consider an arbitrary ordered d-subset $v_1 < \ldots < v_d$ of A. Denote by H the hyperplane determined by these points. Then for any other point $v \in A$, the orientation of the ordered simplex $(v, v_1, \ldots, v_d)$ determines which side of H contains v. Thus vertices $v_1, \ldots, v_d$ form a facet if and only if the sign of the determinant $det(v', v'_1, \ldots, v'_d)$ is the same for every vertex v. This, however, is the same as Gale's evenness condition. Thus the face structure of the convex hull of $\{a_1, \ldots, a_n\}$ is the same as that of a cyclic polytope on n vertices. □

Theorem 4 *There is a function $g(d, r)$ such that for every set A of $m \geq g(d, r)$ points in general position in $\mathbf{R}^d$, halfspaces over bags of size r from A have VC-dimension at least $\lfloor d/2 \rfloor (\log r + 1)$.*

Proof The result follows by combining the construction of Theorem 2 with Lemma 3, setting $g(d, r) = f(d, dr)$. □

4 NP-completeness of hypothesis finding

The hypothesis-finding problem for r-bags for d-dimensional halfspaces is the following: given a set of labeled r-bags in $\mathbf{R}^d$, is there a halfspace that assigns these labels to the bags? The reduction below is a variant of a reduction in Kundakciouglu *et al.* [11].

Theorem 5 *The hypothesis finding problem for r-bags of d-dimensional halfspaces is NP-complete for every fixed $r \geq 3$.*

Proof We give a reduction from 3-SAT (containment in NP is trivial). Let $C_1, \ldots, C_m$ be an instance of 3-SAT over variables $x_1, \ldots, x_d$. Let e_i be the i'th unit vector in $\mathbf{R}^d$. For $j = 1, \ldots, m$ let B_j be a positive bag containing e_i if x_i is in C_j, and $-e_i$ if $\neg x_i$ is in C_j. For $i = 1, \ldots, d$ let B'_i be a positive bag containing e_i and $-e_i$. Finally, let B^* be a negative bag containing 0. We claim that the original formula is satisfiable iff the there is a consistent hypothesis for the set of bags described.

Let $(a_1, \ldots, a_d)$ be a satisfying truth assignment. Then the halfspace $w_1 u_1 + \ldots + w_d u_d \geq 1$ is consistent, where $w_i = 1$ if $a_i = 1$ and $w_i = -1$ otherwise, for $i = 1, \ldots, d$.

In the other direction, let $w_1 u_1 + \ldots + w_d u_d \geq t$ be a consistent hypothesis. Then $t > 0$ as B^* is negative. Also, $w_i \neq 0$, as B'_i is positive. It follows directly that the truth assignment defined by $a_i = sign(w_i)$ satisfies the formula. □

Note that this construction uses bags of size at most 3 (or r in the general case). Adding points to the bags sufficiently close to the given ones and slightly modifying the threshold one can get the same result for bags of the same size.

5 Further remarks and open problems

Active learning in the multi-instance model has received some attention in machine learning, but, as far as we know, has not been considered so far in learning theory. We point out the applicability of some recent active learning results in the context of multi-instance learning, without giving a detailed definition of the notions involved.

There are several possibilities for formulating a model of active learning in the multi-instance model, such as querying bag labels, or various ways of querying instance labels within bags, and these variants may be relevant in different learning scenarios (see Settles *et al.* [15]). Here we assume that the learner gets unlabeled r-bags and then is charged for querying the label of a bag.

The *mellow active learning algorithm* of Cohn *et al.* [4] works as follows: query the label of a bag iff its label is not determined by the labels of the previously queried bags. Hanneke [10] shows that with high probability the error of hypotheses returned by the mellow algorithm decreases exponentially in the number of labels queried. Besides the usual parameters, his bound contains the *disagreement coefficient*, which depends on the concept class, the underlying distribution, and also on the target concept. Friedman [8] proved a general bound for the disagreement coefficient. In particular, his results, and therefore Hanneke's bounds, apply to the learning of hyperplanes over *smooth* distributions. Friedman assumes a smoothness condition for the combined parametrized representation of instances and concepts, but he also gives several extensions to cases where such assumptions do not hold.

Multi-instance learning of r-bags of d-dimensional halfspaces corresponds to learning concepts in (rd)-dimensional space of the form

$$\{(x_1^1, \ldots, x_d^1, \ldots, x_1^r, \ldots, x_d^r) \,:\, w_1 x_1^i + \ldots + w_d x_d^i \geq t \text{ for some } i,\ 1 \leq i \leq i\}.$$

Among the extensions discussed by Friedman, this class is covered by, for example, the result of Balcan *et al.* [2] on the union of exponential rate classes. Thus we conclude that halfspaces are *actively learnable from bags at an exponential rate*.

We showed that the VC-dimension of r-bags of d-dimensional halfspaces is $\Theta(d \log r)$ over every sufficiently large point set in general position and hypothesis finding for r-bags of d-dimensional halfspaces is NP-complete. This means that, unlike the case of learning halfspaces, one does not get an efficient PAC learning algorithm by drawing $O(d \log r)$ random bags and finding a consistent hypothesis. On the other hand, the result of Blum and Kalai [3] *does* provide an efficient algorithm with sample size polynomial in r and d.

This raises two open questions. What is the minimal sample size of r-bags sufficient for efficiently learning d-dimensional halfspaces? What is the minimal sample size of r-bags for PAC learning d-dimensional halfspaces without taking computational complexity into account? For the second question note that distributions over bags generated from arbitrary distributions over instances from a subclass of all possible distributions over bags[2], thus the VC-dimension only provides an upper bound. Multi-instance learning under more general settings has been discussed by Auer *et al.* [1] and by Sabato and Tishby [13].

The mellow algorithm for active learning has an efficient implementation whenever hypothesis finding can be done efficiently. A new instance has to be queried iff the previously queried labels are consistent with both labels for the new instance. This, again, does not work for bags of halfspaces. Thus it seems to be an open problem whether there is an efficient active learning algorithm with exponential error rate.

Acknowledgement We would like to thank Robert Langlois for several interesting discussions.

References

[1] P. Auer, P. M. Long, and A. Srinivasan. Approximating hyper-rectangles: Learning and pseudorandom sets. *J. Comput. Syst. Sci.*, 57:376–388, 1998.

[2] M.-F. Balcan, S. Hanneke, and J. W. Vaughan. The true sample complexity of active learning. *Machine Learning*, 80:111–139, 2010.

[3] A. Blum and A. Kalai. A note on learning from multiple-instance examples. *Machine Learning*, 30(1):23–29, 1998.

[4] D. A. Cohn, L. A. Atlas, and R. A. Ladner. Improving generalization with active learning. *Machine Learning*, 15:201–221, 1994.

[2] This explains why, unlike the standard setting, the efficient PAC learning algorithm of Blum and Kalai [3] does not lead to an efficient hypothesis finding algorithm for bags.

[5] R. Cordovil and P. Duchet. Cyclic polytopes and oriented matroids. *Eur. J. Comb.*, 21:49–64, 2000.

[6] S. Dasgupta. Two faces of active learning. *Theor. Comput. Sci.*, 412(19):1767–1781, 2011.

[7] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89:31–71, 1997.

[8] E. Friedman. Active learning for smooth problems. In *COLT*, 2009.

[9] R. Graham, B. Rothschild, and J. H. Spencer. *Ramsey Theory*. Wiley, second edition, 1990.

[10] S. Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Carnegie Mellon University, 2009.

[11] O. E. Kundakcioglu, O. Seref, and P. M. Pardalos. Multiple instance learning via margin maximixation. *Applied Numerical Mathematics*, 60:358–369, 2010.

[12] J. Matousek. *Lectures on Discrete Geometry*, volume 212 of *Graduate Texts in Mathematics*. Springer, 2002.

[13] S. Sabato and N. Tishby. Homogeneous multi-instance learning with arbitrary dependence. In *COLT*, 2009.

[14] S. Sabato and N. Tishby. Multi-instance learning with any hypothesis class. arXiv:1107.2021v1 [cs.LG], July 2011.

[15] B. Settles, M. Craven, and S. Ray. Multiple-instance active learning. In *NIPS*, 2007.