JOURNAL OF COMBINATORIAL THEORY, Series A 73, 376–380 (1996)
ARTICLE NO. 0031

Note

Witness Sets for Families of Binary Vectors

EYAL KUSHILEVITZ*

Department of Computer Science, Technion, Haifa, Israel

NATHAN LINIAL†

Department of Computer Science, Hebrew University, Jerusalem, Israel

YURI RABINOVICH

Department of Computer Science, University of Toronto, Toronto, Canada

AND

MICHAEL SAKS‡

Department of Mathematics, Rutgers University, New Brunswick, New Jersey

Communicated by the Managing Editors

Received November 22, 1994

Given a family $\mathscr{R} \subseteq \{0, 1\}^m$ of binary vectors of length m, a set $W \subseteq \{1, ..., m\}$ is called a *witness set* for $r \in \mathscr{R}$, if for all other $r' \in \mathscr{R}$ there exists a coordinate $c \in W$ such that $r_c \neq r'_c$. The smallest cardinality of a witness set for $r \in \mathscr{R}$ is denoted $w(r) = w_{\mathscr{R}}(r)$. In this note we show that $\sum_{r \in \mathscr{R}} w(r) = O(|\mathscr{R}|^{3/2})$ and constructions are given to show that this bound is tight. Further information is derived on the distribution of values of $\{w(r) \mid r \in \mathscr{R}\}$. © 1996 Academic Press, Inc.

1. INTRODUCTION

Let $\mathscr{R} \subseteq \{0, 1\}^m$ be a family of distinct binary vectors of length m. A set $W \subseteq [m]$ of coordinates is a *witness set* for a vector r in $\mathscr{R}$, if for every

[*] E-mail: eyalk@cs.technion.ac.il.

[†] Work supported in part by a grant from the Israeli Academy of Sciences. E-mail: nati@huji.ac.il.

[‡] Supported in part by NSF Contract CCR-9215293 and by DIMACS, which is partially funded by NSF Grant STC-91-19999, and the NJ Comission of Science and Technology.

0097-3165/96 $12.00

other $r' \in \mathscr{R}$ there exists a coordinate c in W such that r_c differs from r'_c. In other words, W is a set of entries of r that *distinguish* it from every other vector in $\mathscr{R}$. We also say that *exposing* the entries of r corresponding to W uniquely determines r among vectors in $\mathscr{R}$. Let $w(r) = w_{\mathscr{R}}(r)$ denote the size of the smallest witness set for $r \in \mathscr{R}$. The notion of a witness set seems to be quite natural and it arises, in particular, in the area of computational learning theory under various names: it is called a *discriminant* in [7], and a *specifying set* in [2]; also $\max_{r \in \mathscr{R}} w(r)$ is referred to as the "teaching dimension" (see [6]).

A natural question is what can be said on the average value of $w(r)$ over $r \in \mathscr{R}$. This question goes back to 1965, when Cover [5] proved that for a certain type of families that arise, for example, in the context of pattern recognition the average is $O(1)$. Other related bounds (also for particular families) are given in [1]. We had initiallly conjectured that if $|\mathscr{R}| = n$, then the average value of $w(r)$ over $r \in \mathscr{R}$ is at most logarithmic in n, but as we shall soon see, this is not true and the average can be as large as $\Omega(n^{1/2})$. This result, together with a matching $O(n^{1/2})$ upper bound that holds for *every* such $\mathscr{R}$, constitutes the main contribution of the present note. A similar proof for the lower bound was independently found by Cherniavsky and Statman [4]. Weaker bounds for the average witness size are considered in [7, Exercise 2.12.d]. Finally, we obtain some bounds on the distribution of values of $\{w(r) \mid r \in \mathscr{R}\}$.

2. The Average Size of the Witness Set

Obviously, $w(r) \leqslant |\mathscr{R}| - 1$ for any r and $\mathscr{R}$, and simple examples show that this is tight. If $\mathscr{R}$ consists of the all-0 vector $\mathbf{0}$ and the m unit vectors $e_1, ..., e_m$, then $w(\mathbf{0}) = m$. Since the worst-case witness set may have to be large, we turn to study the average witness size. Let $\mathscr{R}$ be a family of n distinct binary vectors of length m. Define

$$\bar{w}(\mathscr{R}) \equiv \frac{1}{n} \sum_{r \in \mathscr{R}} w(r),$$

the average size of smallest witness sets over the members of $\mathscr{R}$.

Theorem 1. *For any $\mathscr{R}$ as above, $\bar{w}(\mathscr{R}) = O(n^{1/2})$. The bound is tight; i.e., there exist $\mathscr{R}'$ with $\bar{w}(\mathscr{R}') = \Omega(n^{1/2})$.*

Proof. We start with an explicit construction of $\mathscr{R}'$ which achieves the lower bound. Let p be a prime, and let P be the projective plane of order p. The plane P contains $m = p^2 + p + 1$ points, and m lines. We consider

m-dimensional vectors where the coordinates correspond to P's points. $\mathscr{R}'$ is a family of $n=2m$ binary m-vectors, of which m are the characteristic vectors of lines of P, and another m are the m unit vectors.

For $r \in \mathscr{R}'$ corresponding to a line l, $w(r)=2$, since it suffices to expose the coordinates corresponding to any two points on l. Such a pair distinguishes r from all singletons, and since distinct lines share exactly one point, this pair of coordinates distinguishes r also from the characteristic vectors of other lines.

On the other hand, $w(r)=p+2$ if r corresponds to the singleton point q. To distinguish r from the characteristic vector of a line l containing q, a zero in r should be exposed in a coordinate that corresponds to a point on l other than q. There are $p+1$ such lines l, whose pairwise intersection is $\{q\}$, so to distinguish r from all of them, at least $p+1$ distinct 0-entries should be exposed. To distinguish r from other singletons, the 1-entry should be exposed as well, the alternative being to expose *all* (p^2+p) 0-entries.

$$\bar{w}(\mathscr{R}')=\frac{1}{n}\sum_{r\in\mathscr{R}'} w(r)=\frac{1}{2m}\,(m\cdot 2+m\cdot(p+2))=\frac{p+4}{2}\geqslant\frac{1}{2\sqrt{2}}\,n^{1/2}.$$

We turn now to prove the upper bound. For $v\in\{0,1\}^m$ and $W\subseteq[m]$, let $v|_W$ be the restriction of v to the coordinates in W. Bondy (see [3, p. 4]) had shown the following.

LEMMA 2. *For every set $\mathscr{R}\subseteq\{0,1\}^m$ there exists a set W of $\leqslant|\mathscr{R}|-1$ coordinates such that all vectors $\{v|_W : v\in\mathscr{R}\}$ are distinct.*

Order the vectors $r_1, r_2, ..., r_n$ of $\mathscr{R}$ by decreasing value of w. Consider the sum of the k largest values $\sum_{i=1}^{k} w(r_i)$ for a value k soon to be set. Find a set T of at most $k-1$ coordinates as guaranteed by the lemma applied to the family $\{r_1, ..., r_k\}$ and expose the T-coordinates in *all* vectors of $\mathscr{R}$. By the property of T, $r_1, ..., r_k$ are already mutually distinguished. The T-coordinates of every other vector $r\in\mathscr{R}$ distinguish r from all $r_1, ..., r_k$, except, perhaps, *one* r_i. It is possible to expose a single additional bit in r_i to distinguish r_i from r. Apply this step to every r_j, $j>k$. Consequently, each of $r_1, ..., r_k$ is distinguished from every other vector in $\mathscr{R}$. No more than $n-k$ bits get exposed in this process, so

$$\sum_{i=1}^{k} w(r_i)\leqslant k(k-1)+n-k=k^2-2k+n.$$

In particular, it follows that $w(r_k)\leqslant k-2+n/k$.

Putting the two observations together we get

$$\sum_{i=1}^{n} w(r_i) = \sum_{i=1}^{k} w(r_i) + \sum_{i=k+1}^{n} w(r_i) \leqslant (k^2 - 2k + n) + (n-k)\left(k - 2 + \frac{n}{k}\right).$$

Pick $k = \sqrt{n}$; the above inequality yields $\sum_{i=1}^{n} w(r_i) \leqslant 2n^{3/2}$, or $\bar{w}(\mathscr{R}) \leqslant 2\sqrt{n}$. ▮

We cannot resist presenting an alternative proof of the upper bound that is algorithmic and yields a slightly better constant.

(*Algorithmic*) *Proof of Theorem* 1. We seek upper bounds for $f(n) \triangleq \max_{\mathscr{R} : |\mathscr{R}| = n} \sum_{r \in \mathscr{R}} w(r)$. Let $\mathscr{R}$ be a family that achieves the maximum $f(n)$. We first limit our attention to a set T of $n-1$ coordinates as guaranteed by Lemma 2. Fix an $i \in T$ and consider the sets $\mathscr{R}_0$, $\mathscr{R}_1$ of vectors in $\mathscr{R}$ whose ith coordinate is 0 (resp. 1). If both $|\mathscr{R}_0|$, $|\mathscr{R}_1| \geqslant \theta$ (a parameter that we soon set), then add i to the witness set of every vector in $\mathscr{R}$ and proceed recursively with the sets $\mathscr{R}_0$, $\mathscr{R}_1$. Since the ith coordinate distinguishes every vector in $\mathscr{R}_0$ from any vector in $\mathscr{R}_1$, it follows that in this case $f(n) \leqslant n + f(\theta) + f(n-\theta)$. On the other hand, if for every $i \in T$ one of the values appears fewer than θ times we do the following: If there is a vector $u \in \mathscr{R}$ each of whose coordinates equals the majority value (there can be at most one such vector) let all of T be u's witness set. For every other $v \in \mathscr{R}$ there is at least one coordinate where v is in the minority. Letting this index be in v's witness set distinguishes v from $\geqslant n - \theta$ vectors in $\mathscr{R}$ and an additional $\theta - 1$ indices suffice to distinguish it from all the rest. The two considerations together yield that

$$f(n) \leqslant \max(n - 1 + (n-1)\theta,\ n + f(\theta) + f(n-\theta))$$

holds always. Let $\theta = \sqrt{n}$ and solve the recursion to conclude that $f(n) \leqslant n^{3/2}$. ▮

3. Remarks and Open Problems

Given a family $\mathscr{R}$ as before and an integer t, $1 \leqslant t \leqslant n$, define

$$U(\mathscr{R}, t) = |\{r \in \mathscr{R} \mid w(r) \geqslant t\}|, \qquad L(\mathscr{R}, t) = |\{r \in \mathscr{R} \mid w(r) \leqslant t\}|.$$

Our proof of Theorem 1 in fact shows the following.

Lemma 3. *For every $\mathscr{R}$ (of size n) and $t \leqslant n$, $U(\mathscr{R}, t + n/t - 1) < t$.*

It is not difficult to obtain also the following lower bound.

LEMMA 4. *For every $\mathscr{R}$ (of size n) and $t \leqslant \sqrt{n}$, $L(\mathscr{R}, 2t + \log_2 n) \geqslant t^2 - t$.*

Proof. We reorder the vectors in $\mathscr{R}$ as follows: Let those vectors whose first bit is in the minority precede those with the majority bit. Expose the first coordinate in vectors of the minority group. Proceed recursively in the same manner on each group separately, and so on, until each group reduces to a single vector. Observe that (1) each vector is distinguished from all those following it (but not necessarily from those preceding it); (2) no vector has more than $\log_2 n$ bits exposed.

Let $\mathscr{K}$ be the set of the first t^2 vectors. Recall that every vector in $\mathscr{K}$ is distinguished from all those that follow it and has at most $\log_2 n$ bits exposed. Apply Lemma 3 to $\mathscr{K}$ to conclude that $U(\mathscr{K}, 2t) < t$. Therefore, at least $t^2 - t$ of the vectors in $\mathscr{K}$ can be distinguished from other members of $\mathscr{K}$ at the cost of exposing $\leqslant 2t$ additional bits each. Each such vector r satisfies, therefore, $w(r) \leqslant 2t + \log_2 n$ and the desired bound follows. ▌

The most interesting problem left open at this note is, in our opinion, to estimate the minimum (over $\mathscr{R}$) of $L(\mathscr{R}, \log_2 n)$.

REFERENCES

1. M. ANTHONY, G. BRIGHTWELL, AND J. SHAWE-TAYLOR, On specifying Boolean functions by labeled examples, *Discrete Appl. Math.* **61** (1995).
2. M. ANTHONY, G. BRIGHTWELL, D. COHEN, AND J. SHAWE-TAYLOR, On exact specification by examples, *in* "Proceedings, 5th Workshop on Computational Learning Theory, 1992," pp. 311–318.
3. B. BOLLOBAS, "Combinatorics," Cambridge Univ. Press, Cambridge, UK, 1986.
4. J. CHERNIAVSKY AND R. STATMAN, Testing: An abstract approach, *in* "Proceedings, 2nd Workshop on Software Testing, 1988."
5. T. M. COVER, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, *IEEE Trans. Electron. Comput.* **14** (1965), 326–334.
6. S. A. GOLDMAN AND M. J. KEARNS, On the complexity of teaching, *in* "Proceedings, 4th Workshop on Computational Learning Theory, 1991," pp. 303–315.
7. B. K. NATARAJAN, "Machine Learning: A Theoretical Approach," Morgan Kaufmann, San Mateo, CA, 1991.