Complexity of Computing Vapnik-Chervonenkis Dimension

Shinohara, Ayumi
Research Institute of Fundamental Information Science Kyushu University

出版情報：RIFIS Technical Report. 69, 1993-04-29. Research Institute of Fundamental Information Science, Kyushu University
バージョン：

RIFIS-TR-CS-69

RIFIS Technical Report

Complexity of Computing Vapnik-Chervonenkis Dimension

Ayumi Shinohara

April 29, 1993

Research Institute of Fundamental Information Science
Kyushu University 33
Fukuoka 812, Japan

E-mail: ayumi@rifis.sci.kyushu-u.ac.jp Phone: 092-641-1101 ex. 4458

Complexity of Computing Vapnik-Chervonenkis Dimension

Ayumi Shinohara

Research Institute of Fundamental Information Science,
Kyushu University 33, Fukuoka 812, Japan
ayumi@rifis.sci.kyushu-u.ac.jp

Abstract. The Vapnik-Chervonenkis (VC) dimension is known to be the crucial measure of the polynomial-sample learnability in the PAC-learning model. This paper investigates the complexity of computing VC-dimension of a concept class over a finite learning domain. We consider a decision problem called the discrete VC-dimension problem which is, for a given matrix representing a concept class $\mathcal{F}$ and an integer K, to determine whether the VC-dimension of $\mathcal{F}$ is greater than K or not. We prove that (1) the discrete VC-dimension problem is polynomial-time reducible to the satisfiability problem of length J with $O(\log^2 J)$ variables, and (2) for every constant C, the satisfiability problem in conjunctive normal form with m clauses and $C \log^2 m$ variables is polynomial-time reducible to the discrete VC-dimension problem. These results can be interpreted, in some sense, that the problem is "complete" for the class of $n^{O(\log n)}$ time computable sets.

1 Introduction

The PAC learnability due to Valiant [8] is to estimate the feasibility of learning a concept probably approximately correctly, from a reasonable amount of examples (polynomial-sample), within a reasonable amount of time (polynomial-time). It is well-known that the Vapnik-Chervonenkis Dimension (VC-dimension) which is a combinatorial parameter of a concept class plays the key role to determine whether the concept class is polynomial-sample learnable or not [2, 3, 5].

This paper settles a complexity issue on VC-dimension of a concept class over a finite learning domain. We remark that the complexity of computing VC-dimension is of independent interest from the polynomial-time learnability, since it is not directly related to the running time of learning algorithms.

Linial et al. [3] showed that the VC-dimension of a concept class over a finite learning domain can be computed in $n^{O(\log n)}$ time, where n is the size of a given matrix which represents the concept class. Nienhuys-Cheng and Polman [6] gave another $n^{O(\log n)}$-time algorithm, although they have not analyzed its running time. On the other hand, Linial et al. [3] pointed out that the decision version of the problem called the *discrete VC-dimension problem* may have some connection with the problem of finding a minimum dominating set in a tournament, which is shown by Megiddo and Vishkin [4] to be a kind of "complete" problem for the class of $n^{O(\log n)}$ time computable sets.

Along this line, we show that the discrete VC-dimension problem is also "complete" for the class of $n^{O(\log n)}$ time computable sets in the same sense. That is, we give the following two reductions: (1) The discrete VC-dimension problem is reducible in polynomial time to the satisfiability problem of a boolean formula of length J with $O(\log^2 J)$ variables. (2) On the other hand, for every constant C, the satisfiability problem in conjunctive normal form with m clauses and $C\log^2 m$ variables is polynomial-time reducible to the discrete VC-dimension problem. Therefore we can interpret that the discrete VC-dimension problem is one of the natural problems which seem to be neither NP-complete, nor in P.

2 Preliminaries

For a matrix M, let M_{ij} denote the element on row i and column j of M, and the size of M is the number of elements in M. The length of a boolean formula ψ, denoted by $|\psi|$, is the total number of variable occurrences in ψ. For a boolean formula ψ, we denote $[\psi, 1] = \psi$ and $[\psi, 0] = \neg\psi$. For any integers $i \geq 1$ and $t \geq 1$, let $b(i,t)$ denote the t-th binary digit of $(i-1)$, that is, $i = \sum_{t=1}^{\lceil \log i \rceil} 2^{t-1} \cdot b(i,t) + 1$. For example, $b(7,1) = 0$, $b(7,2) = 1$, and $b(7,3) = 1$.

Let U be a finite set called a *learning domain.* We call a subset f of U a *concept.* A concept f can be regarded as a function $f : U \to \{0,1\}$, where $f(x) = 1$ if x is in the concept and $f(x) = 0$ otherwise. A *concept class* is a nonempty set $\mathcal{F} \subseteq 2^U$. We represent a concept class $\mathcal{F}$ over a finite learning domain U, by a $|U| \times |\mathcal{F}|$ matrix M with $M_{ij} = f_j(x_i)$. Each column represents a concept in $\mathcal{F}$. For a $\{0,1\}$-valued matrix M, let $\mathcal{F}_M$ denote the concept class which M represents.

Definition 1. We say that $\mathcal{F}$ *shatters* a set $S \subseteq U$ if for every subset $T \subseteq S$ there exists a concept $f \in \mathcal{F}$ which *cuts* T *out of* S, i.e., $T = S \cap f$. The *Vapnik-Chervonenkis dimension* of $\mathcal{F}$, denoted by VC-dim($\mathcal{F}$), is the maximum cardinality of a set which is shattered by $\mathcal{F}$.

Lemma 2. *[5] For any concept class $\mathcal{F}$, VC-dim$(\mathcal{F}) \leq \log|\mathcal{F}|$.*

By this lemma, Linial et al. [3] immediately claimed that a simple algorithm which enumerates all possible sets to be shattered shall terminate in $n^{O(\log n)}$ time, where n is the size of a given matrix.

Definition 3. [3] The *discrete VC-dimension problem* is, given a $\{0,1\}$-valued matrix M and integer $K \geq 1$, to determine whether VC-dim$\mathcal{F}_M \geq K$ or not.

Definition 4. [4] The classes $\mathrm{SAT}_{\log^k n}$ and $\mathrm{SAT}^{\mathrm{CNF}}_{\log^k n}$ for $k \geq 1$ are defined as follows:

(1) A set L is in $\mathrm{SAT}_{\log^k n}$ if there exists a Turing machine M, a polynomial $p(n)$, and a constant C, such that for every string I of length n, M converts I within $p(n)$ time into a boolean formula Ψ_I (whose length is necessarily less than $p(n)$) with at most $C\log^k n$ variables, so that $I \in L$ if and only if Ψ_I is satisfiable.
(2) The definition of $\mathrm{SAT}^{\mathrm{CNF}}_{\log^k n}$ is essentially the same as that of $\mathrm{SAT}_{\log^k n}$ except that the formula Ψ_I is in conjunctive normal form.

From the definitions, it is easy to see that for each $k \geq 1$,

$$P \subseteq \mathrm{SAT}^{\mathrm{CNF}}_{\log^k n} \subseteq \mathrm{SAT}_{\log^k n} \subseteq NP.$$

3 Discrete VC-dimension Problem is in $\mathrm{SAT}_{\log^2 n}$

In this section, we show that the discrete VC-dimension problem is polynomial-time reducible to the satisfiability problem of a boolean formula of length J with $O(\log^2 J)$ variables.

Theorem 5. *The discrete VC-dimension problem is in $SAT_{\log^2 n}$.*

Proof. Let M be an $m \times r$ matrix and K be an integer. By Lemma 2, we can assume that $K \le \log r$ without loss of generality. Moreover, we can also assume that $m = 2^l$ for some integer l; if $m < 2^l$ for $l = \lceil \log m \rceil$, then we enlarge M by duplicating the last row of M until the row size reaches 2^l. It is easy to see that the size of the enlarged matrix M' is less than twice as large as that of the original matrix M, and $\text{VC-dim}(\mathcal{F}_{M'}) = \text{VC-dim}(\mathcal{F}_M)$.

We now construct a boolean formula Ψ_M which contains $K \cdot l$ variables v_{kt} ($1 \le k \le K$, $1 \le t \le l$) as follows:

$$\begin{aligned}
\Psi_M &= \textstyle\bigwedge_{s=1}^{2^K} \bigvee_{j=1}^{r} \beta_{sj}, \\
\beta_{sj} &= \textstyle\bigwedge_{k=1}^{K} [\alpha_{kj}, b(s,k)] && (1 \le s \le 2^K,\ \ 1 \le j \le r), \\
\alpha_{kj} &= \textstyle\bigvee_{i \in \{i \mid M_{ij}=1\}} \psi_{ki} && (1 \le k \le K,\ 1 \le j \le r), \\
\psi_{ki} &= \textstyle\bigwedge_{t=1}^{l} [v_{kt}, b(i,t)] && (1 \le k \le K,\ 1 \le i \le m).
\end{aligned}$$

Note that the length of Ψ_M is

$$\begin{aligned}
|\Psi_M| &\le l \cdot m \cdot K \cdot r \cdot 2^K \\
&\le \log m \cdot m \cdot \log r \cdot r \cdot 2^{\log r} \\
&< n^2 \log^2 n,
\end{aligned}$$

where $n = m \cdot r$ is the size of the given matrix M. Also note that Ψ_M can be constructed in polynomial time with respect to n.

Let $U = \{x_1, x_2, \ldots, x_m\}$ be the learning domain and $\mathcal{F}_M = \{f_1, f_2, \cdots, f_r\} \subseteq 2^U$ be the concept class which M represents. We will show that the formula Ψ_M is satisfiable if and only if $\mathcal{F}_M$ shatters a set $S \subseteq U$ of cardinality K.

For a formula ψ and a truth assignment σ to the variables of ψ, let $\sigma(\psi)$ denote the truth value of ψ evaluated under σ. We denote truth values by 0 and 1. For each assignment σ, we define a set $S_\sigma \subseteq U$ as follows:

$$S_\sigma = \{x_{\langle \sigma, k \rangle} \mid 1 \le k \le K\}, \quad \text{where } \langle \sigma, k \rangle = \sum_{t=1}^{l} 2^{t-1} \cdot \sigma(v_{kt}) + 1.$$

It should be noticed that the cardinality of S_σ is not always equal to K, since there may be two distinct k_1 and k_2 with $\langle \sigma, k_1 \rangle = \langle \sigma, k_2 \rangle$ in general.

We now show through a sequence of equivalences that an assignment σ satisfies Ψ_M if and only if $|S_\sigma| = K$ and S_σ is shattered by $\mathcal{F}_M$.

First, for any $k \in \{1, \ldots, K\}$ and any $i \in \{1, \ldots, m\}$,

$$
\begin{aligned}
&\sigma(\psi_{ki}) = 1 \\
&\iff \sigma([v_{kt}, b(i,t)]) = 1 \text{ for each } t \in \{1, \ldots, l\} \\
&\iff \sigma(v_{kt}) = \begin{cases} 1 & \text{if } b(i,t) = 1 \\ 0 & \text{if } b(i,t) = 0 \end{cases} \quad \text{for each } t \in \{1, \ldots, l\} \\
&\iff b(i,t) = \sigma(v_{kt}) \text{ for each } t \in \{1, \ldots, l\} \\
&\iff \sum_{t=1}^{l} 2^{t-1} \cdot b(i,t) = \sum_{t=1}^{l} 2^{t-1} \cdot \sigma(v_{kt}) \\
&\iff i = \langle \sigma, k \rangle.
\end{aligned}
$$

Next, for any $k \in \{1, \ldots, K\}$ and any $j \in \{1, \ldots, r\}$,

$$
\begin{aligned}
&\sigma(\alpha_{kj}) = 1 \\
&\iff \sigma(\psi_{ki}) = 1 \text{ and } M_{ij} = 1 \text{ for some } i \in \{1, \ldots, m\} \\
&\iff i = \langle \sigma, k \rangle \text{ and } x_i \in f_j \\
&\iff x_{\langle \sigma, k \rangle} \in f_j.
\end{aligned}
$$

For an integer $s \in \{1, \ldots, 2^K\}$, the s-th subset $S_\sigma^{[s]}$ of S_σ is defined by $S_\sigma^{[s]} = \{x_{\langle \sigma, k \rangle} \mid b(s,k) = 1,\ 1 \leq k \leq K\}$. For example, $S_\sigma^{[1]} = \emptyset$, $S_\sigma^{[5]} = \{x_{\langle \sigma, 3 \rangle}\}$ and $S_\sigma^{[6]} = \{x_{\langle \sigma, 1 \rangle}, x_{\langle \sigma, 3 \rangle}\}$. Then, for any $s \in \{1, \ldots, 2^K\}$ and any $j \in \{1, \ldots, r\}$,

$$
\begin{aligned}
&\sigma(\beta_{sj}) = 1 \\
&\iff \sigma([\alpha_{kj}, b(s,k)]) = 1 \text{ for each } k \in \{1, \ldots, K\} \\
&\iff \sigma(\alpha_{kj}) = \begin{cases} 1 & \text{if } b(s,k) = 1 \\ 0 & \text{if } b(s,k) = 0 \end{cases} \quad \text{for each } k \in \{1, \ldots, K\} \\
&\iff \begin{cases} x_{\langle \sigma, k \rangle} \in f_j & \text{if } b(s,k) = 1 \\ x_{\langle \sigma, k \rangle} \notin f_j & \text{if } b(s,k) = 0 \end{cases} \quad \text{for each } k \in \{1, \ldots, K\} \\
&\iff \{x_{\langle \sigma, k \rangle} \mid b(s,k) = 1,\ 1 \leq k \leq K\} \subseteq f_j \text{ and} \\
&\qquad \{x_{\langle \sigma, k \rangle} \mid b(s,k) = 0,\ 1 \leq k \leq K\} \subseteq U - f_j \\
&\iff S_\sigma \cap f_j = S_\sigma^{[s]} \text{ and} \\
&\qquad b(s,k_1) \neq b(s,k_2) \text{ implies } \langle \sigma, k_1 \rangle \neq \langle \sigma, k_2 \rangle \text{ for any } k_1, k_2 \in \{1, \ldots, K\}.
\end{aligned}
$$

Finally, we get the following equivalence:

$$
\begin{aligned}
&\sigma(\Psi_M) = 1 \\
&\iff \sigma\Big(\bigvee_{j=1}^{r} \beta_{sj}\Big) = 1 \text{ for any } s \in \{1, \ldots, 2^K\} \\
&\iff \text{for each } s \in \{1, \ldots, 2^K\}, \\
&\qquad \text{there exists } f_j \in \mathcal{F}_M \text{ with } S_\sigma \cap f_j = S_\sigma^{[s]} \text{ and} \\
&\qquad b(s,k_1) \neq b(s,k_2) \text{ implies } \langle \sigma, k_1 \rangle \neq \langle \sigma, k_2 \rangle \\
&\iff k_1 \neq k_2 \text{ implies } \langle \sigma, k_1 \rangle \neq \langle \sigma, k_2 \rangle, \text{ and} \\
&\qquad \text{for each } s \in \{1, \ldots, 2^K\} \text{ there exists } f_j \in \mathcal{F}_M \text{ with } S_\sigma \cap f_j = S_\sigma^{[s]} \\
&\iff |S_\sigma| = K \text{ and } S_\sigma \text{ is shattered by } \mathcal{F}_M
\end{aligned}
$$

Thus the formula Ψ_M is satisfiable if and only if VC-dim$\mathcal{F}_M \geq K$. □

4 Discrete VC-dimension Problem is $\mathrm{SAT}^{\mathrm{CNF}}_{\log^2 n}$-hard

This section shows that every set in $\mathrm{SAT}^{\mathrm{CNF}}_{\log^2 n}$ is reducible to the discrete VC-dimension problem in polynomial time, i.e., the problem is $\mathrm{SAT}^{\mathrm{CNF}}_{\log^2 n}$-hard.

Theorem 6. *Every $L \in SAT^{\mathrm{CNF}}_{\log^2 n}$ is polynomial-time reducible to the discrete VC-dimension problem.*

We use the following lemma in the proof of Theorem 6.

Lemma 7. *Let $\mathcal{F}$ be a concept class over a learning domain U, and S be a subset of U with $|S| = d \geq 2$. If S is shattered by $\mathcal{F}$, then for any two distinct x and y in S, the number of concepts which contain exactly one of either x or y is at least 2^{d-1}, i.e.,*

$$\left|\{f \in \mathcal{F} \mid f(x) \neq f(y)\}\right| \geq 2^{d-1}.$$

Proof. Let $\mathcal{F}_{\bar{x}y} = \{f \in \mathcal{F} \mid f(x) = 0,\ f(y) = 1\}$, and $\mathcal{F}_{x\bar{y}} = \{f \in \mathcal{F} \mid f(x) = 1,\ f(y) = 0\}$. Then $\{f \in \mathcal{F} \mid f(x) \neq f(y)\} = \mathcal{F}_{\bar{x}y} \cup \mathcal{F}_{x\bar{y}}$, and $\mathcal{F}_{\bar{x}y} \cap \mathcal{F}_{x\bar{y}} = \emptyset$. It is easy to see that if S is shattered by $\mathcal{F}$ then the set $S - \{x, y\}$ is shattered by both $\mathcal{F}_{\bar{x}y}$ and $\mathcal{F}_{x\bar{y}}$. By Lemma 2, $|S - \{x,y\}| \leq \log|\mathcal{F}_{\bar{x}y}|$ and $|S - \{x,y\}| \leq \log|\mathcal{F}_{x\bar{y}}|$. Thus $|\mathcal{F}_{\bar{x}y}| \geq 2^{d-2}$ and $|\mathcal{F}_{x\bar{y}}| \geq 2^{d-2}$, which yield $\left|\{f \in \mathcal{F} \mid f(x) \neq f(y)\}\right| = |\mathcal{F}_{\bar{x}y}| + |\mathcal{F}_{x\bar{y}}| \geq 2^{d-2} + 2^{d-2} = 2^{d-1}$. □

Proof of Theorem 6. Let $L \in \mathrm{SAT}^{\mathrm{CNF}}_{\log^2 n}$. Then there is a constant C_L and a polynomial $p_L(n)$ such that every string I of length n can be reduced in $p_L(n)$ time to a boolean formula in conjunctive normal form with at most $C_L \log^2 n$ variables, whose satisfiability coincides with the membership $I \in L$. Therefore we have only to show that, for any C, there is a polynomial-time reduction from the satisfiability problem in conjunctive normal form with at most $C \log^2 n$ variables to the discrete VC-dimension problem. Let $\Psi = E_1 \wedge \cdots \wedge E_m$ $(m \geq 2)$ be a boolean formula where each E_i is a disjunction and the total number of distinct variables occurring in Ψ is not greater than $C \log^2 m$. Without loss of generality, we can assume that m is a power of 2. We can also assume that the number of variables is exactly $C \log^2 m$, and let us rename them, for convenience, with double indices v_{st} $(1 \leq s \leq \log m,\ 1 \leq t \leq C \log m)$. We first construct a matrix M_Ψ which has $(m^C + 1)\log m$ rows and $m^2 + m(\log m - 1)$ columns, and then prove that VC-dim$(\mathcal{F}_{M_\Psi}) = 2\log m$ if and only if Ψ is satisfiable.

The learning domain U corresponding to Ψ is defined as $U = X \cup Y$ with $X \cap Y = \emptyset$, where $Y = \{y_u \mid 1 \leq u \leq \log m\}$ and $X = \{x_{sl} \mid 1 \leq s \leq \log m,\ 1 \leq l \leq m^C\}$. Let $X_s = \{x_{sl} \in X \mid 1 \leq l \leq m^C\}$ for each $s \in \{1, \ldots, \log m\}$, and let $X^{[k]} = \bigcup_{s \in \{s \mid b(k,s)=1\}} X_s$ for each $k \in \{1, \ldots, m\}$. The i-th subset $Y^{[i]}$ of Y is defined by $Y^{[i]} = \{y_u \in Y \mid b(i,u) = 1\}$ for each $i \in \{1, \ldots, m\}$.

The concept class $\mathcal{F} \subseteq 2^U$ is defined as the union of distinct subclasses $F_1, \ldots, F_m$, and G. Here, the structure of G depends only on the number m:

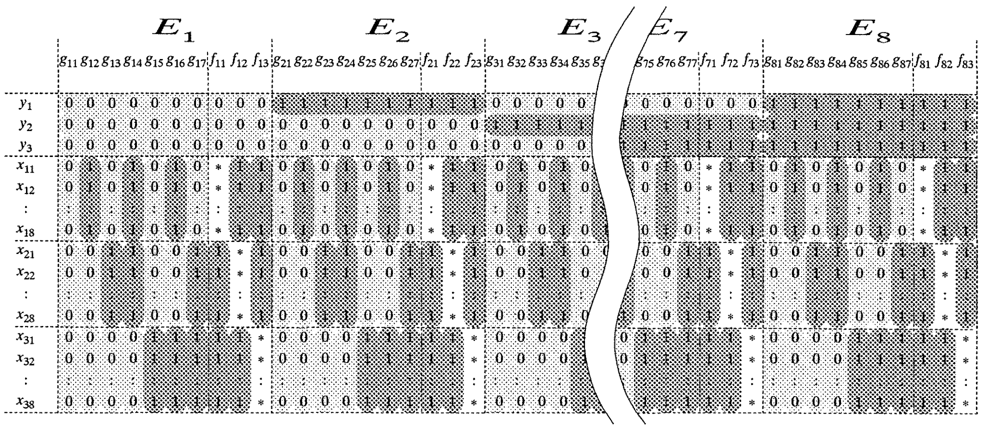


Fig. 1. Structure of the matrix M_Ψ reduced from a boolean formula $\Psi = E_1 \wedge E_2 \wedge \cdots \wedge E_8$ with $C = 1$. In this case, $K = 2\log 8 = 6$. The only elements marked '$*$' depend on the structure of each clause E_i in Ψ.

$$G = \{g_{ik} \mid 1 \le i \le m,\ 1 \le k \le m-1\}, \quad \text{where } g_{ik} = Y^{[i]} \cup X^{[k]}.$$

On the other hand, each concept in F_i reflects the structure of the clause E_i in Ψ:

$$F_i = \{f_{ij} \mid 1 \le j \le \log m\}, \quad \text{where } f_{ij} = Y^{[i]} \cup (X - X_j) \cup X_j^*(E_i) \quad \text{with}$$

$$X_j^*(E_i) = \left\{ x_{jl} \in X_j \;\middle|\; \begin{array}{l} E_i \text{ contains a positive literal } v_{jt} \text{ with } b(l,t) = 1, \text{ or} \\ E_i \text{ contains a negative literal } \neg v_{jt} \text{ with } b(l,t) = 0 \\ \text{for some } t \in \{1, \ldots, C\log m\} \end{array} \right\}.$$

Figure 1 illustrates the structure of the matrix M_Ψ.

Clearly the cardinality of learning domain, i.e., the row size of the matrix M representing $\mathcal{F}$ is

$$|U| = |X| + |Y| = m^C \cdot \log m + \log m = (m^C + 1)\log m,$$

and the cardinality of the concept class $\mathcal{F}$, i.e., the column size of M is

$$|\mathcal{F}| = |G| + |F_1| + \cdots + |F_m| = m(m-1) + m \cdot \log m.$$

Moreover, it is easy to see that M_Ψ can be constructed in polynomial time with respect to the length of given formula Ψ.

Now we prove that if the formula Ψ is satisfiable then VC-dim$(\mathcal{F}) = 2\log m$. For an assignment σ which satisfies Ψ, we consider the set $S_\sigma = Y \cup X_\sigma$ with

$$X_\sigma = \{x_{s,\langle\sigma,s\rangle} \in X \mid 1 \le s \le \log m\}, \quad \text{where } \langle\sigma,s\rangle = \sum_{t=1}^{C\log m} 2^{t-1} \cdot \sigma(v_{st}) + 1.$$

It is clear that $|S_\sigma| = |Y| + |X_\sigma| = 2\log m$. We will show that S_σ is shattered by $\mathcal{F}$, i.e., for every $T \subseteq S_\sigma$, there exists an $f \in \mathcal{F}$ with $S_\sigma \cap f = T$. Let $i_T = \sum_{y_u \in T\cap Y} 2^{u-1} + 1$. It is easy to see that $i_T \in \{1, \ldots, m\}$ and $T \cap Y = Y^{[i_T]}$. According to $T \cap X_\sigma = X_\sigma$ or not, we have the following two cases.

(1) In case of $T \cap X_\sigma \subsetneq X_\sigma$: Let $k_T = \sum_{x_{s,\langle\sigma,s\rangle} \in T \cap X_\sigma} 2^{s-1} + 1$. Then we can see that $k_T \in \{1, \ldots, m-1\}$ and $T \cap X_\sigma = X^{[k_T]} \cap X_\sigma$. Therefore the concept $g_{i_T,k_T} \in G \subseteq \mathcal{F}$ cuts T out of S_σ as follows:

$$\begin{aligned} g_{i_T,k_T} \cap S_\sigma &= (Y^{[i_T]} \cup X^{[k_T]}) \cap (Y \cup X_\sigma) = (Y^{[i_T]} \cap Y) \cup (X^{[k_T]} \cap X_\sigma) \\ &= (T \cap Y) \cup (T \cap X_\sigma) = T. \end{aligned}$$

(2) In case of $T \cap X_\sigma = X_\sigma$: Since σ satisfies Ψ, the disjunction E_{i_T} in Ψ is also satisfied by σ. That means E_{i_T} contains either positive literal v_{st} with $\sigma(v_{st}) = 1$, or negative literal $\neg v_{st}$ with $\sigma(v_{st}) = 0$, for some s and t. Let us take such an s (not necessarily unique), and let $j_T = s$. Then by the definition of $\langle\sigma, j_T\rangle$, we see $b(\langle\sigma, j_T\rangle, t) = \sigma(v_{j_T,t})$ for each t. Thus $x_{j_T,\langle\sigma,j_T\rangle}$ is included in $X^*_{j_T}(E_{i_T})$, and moreover, $X^*_{j_T}(E_{i_T}) \cap X_\sigma = \{x_{j_T,\langle\sigma,j_T\rangle}\}$. Therefore the concept $f_{i_T,j_T} \in F_{i_T} \subseteq \mathcal{F}$ cuts T out of S_σ as follows:

$$\begin{aligned} f_{i_T,j_T} \cap S_\sigma &= (Y^{[i_T]} \cup (X - X_{j_T}) \cup X^*_{j_T}(E_{i_T})) \cap (Y \cup X_\sigma) \\ &= (Y^{[i_T]} \cap Y) \cup ((X - X_{j_T}) \cap X_\sigma) \cup (X^*_{j_T}(E_{i_T}) \cap X_\sigma) \\ &= (T \cap Y) \cup (X_\sigma - \{x_{j_T,\langle\sigma,j_T\rangle}\}) \cup \{x_{j_T,\langle\sigma,j_T\rangle}\} \\ &= (T \cap Y) \cup X_\sigma = (T \cap Y) \cup (T \cap X_\sigma) = T. \end{aligned}$$

In each case, T is shown to be cut out of S_σ by some concept in $\mathcal{F}$. Therefore S_σ is shuttered by $\mathcal{F}$.

Now we show the converse. Suppose that VC-dim$(\mathcal{F}) = 2\log m$. Then there is a set $S \subseteq U$ of cardinality $2\log m$ which is shattered by $\mathcal{F}$.

Claim 1 *S contains exactly one element from each X_s $(1 \le s \le \log m)$, and all elements from Y.*

Proof of Claim 1. Case $m = 2$: The learning domain is $U = \{y_1\} \cup X_1$ and the concept class is $\mathcal{F} = \{f_{11}, g_{11} f_{21}, g_{21}\}$. Since $|\mathcal{F}| = 4$ and $g_{11}(x) = g_{21}(x) = 0$ for any $x \in X_1$, no two elements from X_1 can be included in S which is to be shattered by $\mathcal{F}$. Moreover, since $|Y| = |\{y_1\}| = 1$, the claim holds.

Case $m \ge 3$: Let $s \in \{1, \ldots, \log m\}$ be fixed arbitrarily, and x_1, x_2 be distinct elements in X_s. Suppose that S contains both x_1 and x_2. Then by Lemma 7,

$$\left|\{h \in \mathcal{F} \mid h(x_1) \neq h(x_2)\}\right| \ge 2^{2\log m - 1} = \frac{m^2}{2}.$$

On the other hand, let us consider a concept $h \in \mathcal{F}$ with $h(x_1) \neq h(x_2)$. Since $g_{ik}(x_1) = g_{ik}(x_2) = b(k, s)$ for any $g_{ik} \in G$, the concept h is not in G. Moreover, since $f_{ij}(x_1) = f_{ij}(x_2) = 1$ for any $f_{ij} \in F_1 \cup \cdots \cup F_m$ with $j \neq s$, thus h must be one of the concepts from $\{f_{1s}, f_{2s}, \ldots, f_{ms}\}$. Therefore

$$\left|\{h \in \mathcal{F} \mid h(x_1) \neq h(x_2)\}\right| \le \left|\{f_{1s}, f_{2s}, \ldots, f_{ms}\}\right| = m,$$

which yields a contradiction since $\frac{m^2}{2} > m$ for any $m \ge 3$. Thus S can contain at most one element from X_s for each $s \in \{1, \ldots, \log m\}$. Since $|S| = 2\log m$ and $|Y| = \log m$, the set S must contain exactly one element from each X_s and all elements from Y. □

Therefore for each $s \in \{1, \dots, \log m\}$, there is a unique $l = l(s) \in \{1, \dots, m^C\}$ such that $x_{s,l(s)} \in S$, and we can assume that $S = Y \cup X_{(l)}$, where $X_{(l)} = \{x_{s,l(s)} \mid 1 \leq s \leq \log m\}$. Let σ_S be an assignment corresponding to S with

$$\sigma_S(v_{st}) = b(l(s), t) \quad (1 \leq s \leq \log m,\ 1 \leq t \leq C \log m).$$

Now we show that σ_S satisfies all disjunctions E_i in Ψ. Let $i \in \{1, \dots, m\}$ be fixed arbitrarily. Since S is shattered by $\mathcal{F}$, for the subset $T_i = Y^{[i]} \cup X_{(l)}$ of S there is a concept $h_i \in \mathcal{F}$ with $S \cap h_i = T_i$. Since $S \cap h_i = (Y \cap h_i) \cup (X_{(l)} \cap h_i)$, the concept h_i must satisfy the following two conditions:

(1) $Y \cap h_i = Y^{[i]}$.
(2) $X_{(l)} \cap h_i = X_{(l)}$.

Note that no concept in G satisfies the condition (2), and no concept in $F_{i'}$ with $i' \neq i$ satisfies the condition (1). Therefore such an $h_i \in \mathcal{F}$ is in F_i, and thus we can assume $h_i = f_{ij}$ for some $j \in \{1, \dots, \log m\}$. The above condition (2) requires that f_{ij} contains all elements from $X_{(l)}$. Especially, remark that $x_{j,l(j)} \in X_{(l)}$ is included in f_{ij} for the above j. By the definition of f_{ij}, the element $x_{j,l(j)}$ is in $X_j^*(E_i)$. Thus the clause E_i satisfies either (a) or (b):

(a) E_i contains a positive literal v_{jt} with $b(l(j), t) = 1$.
(b) E_i contains a negative literal $\neg v_{jt}$ with $b(l(j), t) = 0$.

By the definition of σ_S, we see $\sigma_S(v_{jt}) = 1$ in case of (a), and $\sigma_S(v_{jt}) = 0$ in case of (b). In each case, $\sigma_S(E_i) = 1$. Therefore σ_S satisfies every disjunction E_i in Ψ. Thus Ψ is satisfiable. □

5 Conclusion

We showed that the discrete VC-dimension problem is in $\mathrm{SAT}_{\log^2 n}$ and $\mathrm{SAT}^{\mathrm{CNF}}_{\log^2 n}$-hard. Therefore we may interpret that the discrete VC-dimension problem is, in some sense, "complete" for the class of $n^{O(\log n)}$ time computable sets. It remains open that the discrete VC-dimension problem is in $\mathrm{SAT}^{\mathrm{CNF}}_{\log^2 n}$, or $\mathrm{SAT}_{\log^2 n}$-hard.

As a dual to the VC-dimension, Romanik [7] defined the *testing dimension* of a concept class $\mathcal{F}$ as the *minimum* cardinality of a set $S \subseteq U$ which is *not* shattered by $\mathcal{F}$. We can see that testing dimension problem is also in $\mathrm{SAT}_{\log^2 n}$, by a similar reduction in the proof of Theorem 5. It is open whether the problem is $\mathrm{SAT}^{\mathrm{CNF}}_{\log^2 n}$-hard or not. It is also interesting to evaluate the complexity of computing another various dimensions of the class of *multi-valued functions* introduced in [1].

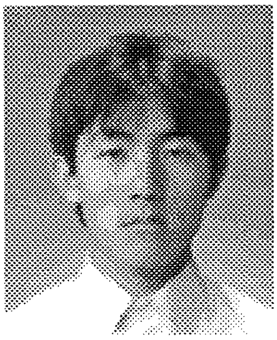
Acknowledgments

The author would like to thank Satoru Miyano and Hiroki Arimura for helpful discussions.

References

1. S. Ben-David, N. Cesa-Bianchi, and P.M. Long. Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions. In *Proc. 5th Annual Workshop on Computational Learning Theory*, pages 333–340, 1992.
2. A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *JACM*, 36(4):929–965, 1989.
3. N. Linial, Y. Mansour, and R.L. Rivest. Results on learnability and the Vapnik-Chervonenkis dimension. *Information and Computation*, 90:33–49, 1991.
4. N. Megiddo and U. Vishkin. On finding a minimum dominating set in a tournament. *Theoretical Computer Science*, 61:307–316, 1988.
5. B.K. Natarajan. *Machine Learning — A Theoretical Approach.* Morgan Kaufmann Publishers, 1991.
6. S.H. Nienhuys-Cheng and M. Polman. Complexity dimensions and learnability. In *Proc. European Conference on Machine Learning, (Lecture Notes in Artificial Intelligence 667)*, pages 348–353, 1993.
7. K. Romanik. Approximate testing and learnability. In *Proc. 5th Annual Workshop on Computational Learning Theory*, pages 327–332, 1992.
8. L.G. Valiant. A theory of the learnable. *CACM*, 27(11):1134–1142, 1984.

About the Author



Ayumi Shinohara (篠原　歩) was born in Fukuoka on July 18, 1965. He received the B.S. degree in 1988 in Mathematics and the M.S degree in 1990 in Information Systems from Kyushu University. Presently, he is an Assistant of Research Institute of Fundamental Information Science, Kyushu University. His research interests are computational learning theory and algorithms.

Research Institute of Fundamental Information Scinece, Kyushu University 33, Fukuoka 812, Japan.