Complexity of Computing Generalized VC-Dimensions

Ayumi Shinohara

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

Abstract. In the PAC-learning model, the Vapnik-Chervonenkis (VC) dimension plays the key role to estimate the polynomial-sample learnability of a class of binary functions. For a class of $\{0,\ldots,m\}$-valued functions, the notion has been generalized in various ways. This paper investigates the complexity of computing some of generalized VC-dimensions: VC*-dimension, Ψ_*-dimension, and Ψ_G-dimension. For each dimension, we consider a decision problem that is, for a given matrix representing a class $\mathcal{F}$ of functions and an integer K, to determine whether the dimension of $\mathcal{F}$ is greater than K or not. We prove that the VC*-dimension problem is polynomial-time reducible to the satisfiability problem of length J with $O(\log^2 J)$ variables, which includes the original VC-dimension problem as a special case. We also show that the Ψ_G-dimension problem is still reducible to the satisfiability problem of length J with $O(\log^2 J)$, while the Ψ_*-dimension problem becomes NP-complete.

1 Introduction

The PAC-learnability due to Valiant [13] is to estimate the feasibility of learning a *binary* function 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 class of binary functions plays the key role to determine whether the class is polynomial-sample learnable or not [3, 5, 8]. As a natural extension, the learnability of a class of $\{0,\ldots,m\}$-*valued* functions has been characterized by various generalized notions such as pseudo-dimension [4], graph dimension [7], and Natarajan dimension [7]. Ben-David et al. [2] unified them into a general scheme, by introducing a family Ψ of mappings which translate $\{0,\ldots,m\}$-valued functions into binary ones.

This paper deals with complexity issues on some of these dimensions of a class over a finite learning domain. We remark that the complexity of computing each dimension is of independent interest from the polynomial-time learnability, since it is not directly related to the running time of learning algorithms.

According to the complexity of finding VC-dimension of a class of binary functions over a finite learning domain, Linial et al. [5] showed that the VC-dimension can be computed in $n^{O(\log n)}$ time, where n is the size of a given

matrix which represents the class. Nienhuys-Cheng and Polman [9] gave another $n^{O(\log n)}$-time algorithm, although they have not analyzed its running time. The author [11] showed that the decision version of the problem is "complete" for the class of $n^{O(\log n)}$ time computable sets, in the same sense as the problem of finding a minimum dominating set in a tournament due to Megiddo and Vishkin [6]. That is, we showed that the VC-dimension problem is in the class $\mathrm{SAT}_{\log^2 n}$, and hard for the class $\mathrm{SAT}^{\mathrm{CNF}}_{\log^2 n}$, where these classes are defined as follows [6]:

- A set L is in $\mathrm{SAT}_{\log^2 n}$ 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 Φ_I (whose length is necessarily less than $p(n)$) with at most $C \log^2 n$ variables, so that $I \in L$ if and only if Φ_I is satisfiable.
- The definition of $\mathrm{SAT}^{\mathrm{CNF}}_{\log^2 n}$ is essentially the same as that of $\mathrm{SAT}_{\log^2 n}$ except that the formula Φ_I is in conjunctive normal form.

In this paper, we extend the above results in three ways. For the notions of generalized VC-dimensions, "VC*-dimension", "Ψ_*-dimension", and "Ψ_G-dimension", we settle the complexity issues (Theorem 3, 6, 8). These results give some connections between various dimension problems and the satisfiability problems of boolean formulae with restricted number of variables. Because of the space limitation, we just state our results briefly in this paper. The proofs are described in our technical report [12].

2 Complexity of VC*-Dimension Problem

In this section, we introduce a natural generalization of the VC-dimension for a class of $\{0, \ldots, m\}$-valued functions. Then we show that the generalized VC-dimension problem is still in $\mathrm{SAT}_{\log^2 n}$ and $\mathrm{SAT}^{\mathrm{CNF}}_{\log^2 n}$-hard, as well as the original VC-dimension problem.

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. Let U be a finite set called a *learning domain*, and $\mathcal{N}$ be the set of natural numbers. For a class $\mathcal{F}$ of functions from U to $\mathcal{N}$, we define $range(\mathcal{F}) = \bigcup_{f \in \mathcal{F}} \{f(x) \mid x \in U\}$. We represent $\mathcal{F}$ by a $|U| \times |\mathcal{F}|$ matrix M with $M_{ij} = f_j(x_i)$. Each column represents a function in $\mathcal{F}$. For an integer matrix M, let $\mathcal{F}_M$ denote the class of functions which M represents.

The following definition may be one of the most natural extensions of the VC-dimension for a class of $\{0, \ldots, m\}$-valued functions.

Definition 1. Let $\mathcal{F}$ be a class of functions over U. We say that $\mathcal{F}$ *shatters* a set $S \subseteq U$ if for every function g from S to $range(\mathcal{F})$, there exists a function $f \in \mathcal{F}$ such that $f(x) = g(x)$ for all $x \in S$. The *VC*-dimension* of $\mathcal{F}$, denoted by VC*-dim($\mathcal{F}$), is the maximum cardinality of a set which is shattered by $\mathcal{F}$.

We note that VC*-dim($\mathcal{F}$) coincides with the original VC-dimension for any class $\mathcal{F}$ of functions with $range(\mathcal{F}) = \{0, 1\}$.

Definition 2. The *VC*-dimension problem* is, given an integer matrix M and integer $K \geq 1$, to determine whether VC*-dim$(\mathcal{F}_M) \geq K$ or not.

Theorem 3. *The VC*-dimension problem is in* $SAT_{\log^2 n}$, *and* $SAT^{\mathrm{CNF}}_{\log^2 n}$*-hard.*

3 Complexity of Ψ-Dimension Problems

The VC*-dimension introduced in the previous section seems to be one of the most natural extension of the VC-dimension. However, it has not been used actually in the literatures. The reason is that the cardinality of the largest class $\mathcal{F}$ of functions over U of a given dimension grows exponentially in $|U|$ for all $|range(\mathcal{F})| > 2$ [1, 2], whereas polynomial growth is desirable for the PAC-learning model. As alternative definitions, a variety of notions of dimension to classes of $\{0, \ldots, m\}$-valued functions had been proposed [4, 7], and Ben-David et al. gave a general scheme [2] which unified them. They introduced Ψ-dimension, where Ψ is a family of mappings which translate $\{0, \ldots, m\}$-valued functions into $\{0, 1\}$-valued ones. In this section, we investigate the complexity of computing Ψ-dimension for two special families Ψ_* and Ψ_G. We show that the Ψ_*-dimension problem is NP-complete, while the Ψ_G-dimension is still in $\mathrm{SAT}_{\log^2 n}$.

Definition 4. Let Ψ be a family of the mappings ψ from $\mathcal{N}$ to $\{0, 1, *\}$, where $*$ will be thought of as a null element. Let $\mathcal{F}$ be a class of functions over U. We say that $\mathcal{F}$ *Ψ-shatters* [1] a set $S \subseteq U$ if there exists a mapping $\psi \in \Psi$ which satisfies the following condition: for every subset $T \subseteq S$, there exists a function $f \in \mathcal{F}$ with $\psi(f(x)) = 1$ for any $x \in T$ and $\psi(f(x)) = 0$ for any $x \in S - T$. That is, Ψ-shattering requires that under some mapping $\psi \in \Psi$, $\mathcal{F}$ contains all functions from U to $\{0, 1\}$. The *Ψ-dimension* of $\mathcal{F}$, denoted by Ψ-dim$(\mathcal{F})$, is the maximum cardinality of a set which is Ψ-shattered by $\mathcal{F}$.

Definition 5. For a family Ψ of mappings from $\mathcal{N}$ to $\{0, 1, *\}$, we define the *Ψ-dimension problem* as the decision problem to determine whether Ψ-dim$(\mathcal{F}_M) \geq K$ or not for given integer matrix M and an integer $K \geq 1$.

Let Ψ_* be the family of all mappings from $\mathcal{N}$ to $\{0, 1, *\}$. Therefore, the Ψ_*-dimension problem is the most general one in the family of Ψ-dimension problems.

Theorem 6. *The Ψ_*-dimension problem is NP-complete.*

Natarajan [7] introduced the *graph dimension* in order to characterize the learnability of a class of $\{0, \ldots, m\}$-valued functions.

Definition 7. The *graph dimension* is the Ψ_G-dimension with $\Psi_G = \{\psi_{G,\tau} \mid \tau \in \mathcal{N}\}$, where $\psi_{G,\tau}(a)$ is 1 if $a = \tau$, and 0 otherwise.

[1] In [2], they introduced more general notions of Ψ-shatter and Ψ-dimension. Our definition of the Ψ-dimension corresponds to the *uniform Ψ-dimension* they call.

From the definition, Ψ_G is a subset of Ψ_*. The following theorem gives an interesting contrast with the Theorem 6.

Theorem 8. *The Ψ_G-dimension problem is in $SAT_{\log^2 n}$, and $SAT^{\mathrm{CNF}}_{\log^2 n}$-hard.*

4 Concluding Remarks

As another crucial characterization of the complexity of computing VC-dimension, Papadimitriou and Yannakakis [10] defined a new complexity class LOGNP, for which the (original) VC-dimension problem becomes complete. We will analyze the complexity of some other dimensions, such as pseudo-dimension [4] and Natarajan dimension [7], together with the relations to the class LOGNP in future works.

References

1. N. Alon. On the density of sets of vectors. *Discrete Mathematics*, 46:199–202, 1983.
2. S. Ben-David, N. Cesa-Bianchi, and P. M. Long. Characterizations of learnability for classes of $\{0,\ldots,n\}$-valued functions. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pages 333–340, 1992.
3. A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *JACM*, 36(4):929–965, 1989.
4. D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.
5. N. Linial, Y. Mansour, and R. L. Rivest. Results on learnability and the Vapnik-Chervonenkis dimension. *Information and Computation*, 90:33–49, 1991.
6. N. Megiddo and U. Vishkin. On finding a minimum dominating set in a tournament. *Theoretical Computer Science*, 61:307–316, 1988.
7. B. Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.
8. B. Natarajan. *Machine Learning — A Theoretical Approach.* Morgan Kaufmann Publishers, 1991.
9. S. 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.
10. C.H. Papadimitriou and M. Yannakakis. On limited nondeterminism and the complexity of the V-C dimension. In *Proc. 8th Annual Conference on Structure in Complexity Theory*, pages 12–18, 1993.
11. A. Shinohara. Complexity of computing Vapnik-Chervonenkis dimension. In *Proc. 4th Workshop on Algorithmic Learning Theory*, pages 279–287, 1993.
12. A. Shinohara. Complexity of computing generalized VC-dimensions. RIFIS Technical Report, RIFIS-TR-CS 78, Research Institute of Fundamental Information Science, Kyushu University, 1993.
13. L. Valiant. A theory of the learnable. *CACM*, 27(11):1134–1142, 1984.