A Step towards a New Test for Learnability of Machine Learning*

Hiroshige INAZUMI† Kin-ichiro TOKIWA ‡ Robert C. HOLTE §

Abstract

Based on PAC learning, A new test for learnability is proposed from the viewpoint of rate distrotion theroy. The criterion depends on the potential property of concept classes, which shows the relationship between sample complexity and accuracy.

1 Introduction

Recent Machine Learning frameworks, not demanding that the hypothesis produced by learning algorithm will be exactly correct, have provided much interest in the research fields of Information Theory. The main problem of Information Theory is to analyze how to realize the reliable and effective communications assuming some noisy conditions, i.e. uncertain sources or channels. In the past, Information Theory has provided Machine Learning with some kinds of criteria (e.g. entropy) or biases (e.g. minimum description length; MDL) [1].

The purpose of this paper is not to get perfect analogy between framework of Machine Learning and that of Information Theory, but to provide reciprocal actions with each other by partial analogy between them, i.e. provide some effective information and suggestion with some Machine Learning strategies.

We consider Valiant's PAC(Probably Approximately Correct) learning framework [2], and FAC(Frequently Approximately Correct) learning framework introduced by Diettreigh [3], which is a little defferent from PAC learning framework. In learning a class C of concepts from examples, a single target concept is selected from C and we are given a finite sequence, each labeled "1" if it is in the target concept (a positive example) and "0" if it is not (a negative example). This set is a training set from instance space, which is also called a sample of the target concept. A learning function for C is a function that, given a large enough randomly drawn sample of any target concept in C, returns a region (a hypothesis) that is with high probability a good approximation to the target concept. In PAC learning model, a hypothesis must be guessed with arbitrarily small error with arbitrarily high probability for a large enough sample size, no matter which concept from C we are trying to learn. The bounds on the sample size must be independent of the underlying distribution P. Necessary and sufficient conditions on a class of concepts C for the existence of a learning function satisfying the above conditions are given by the simple combinatorial parameter called the Vapnik-Chervonenkis (VC) dimension of the class C of the concepts [4, 5, 6].

Then we consider the following arrangement of the previous framework.

- In learning a class C of concepts, let training set be the set of samples from the compressed instance space. Given compressed space, how much accuracy is guaranteed for guessing any target concept.

We treat trade-off ralationship between compression-rate and error-rate (=1-accuracy) in rate-distortion theory [7], by which the static and potential property of the model can be analyzed. If the source information is compressed under the source entropy through coding procedure, the source information will not be exactly represented after the decoding, i.e. with some errors. Such kind of errors is called distortion. The purpose of rate-distortion theory is to show the compression bounds, R, assuming an average distortion, D, and also show the distortion bounds assuming some fixed compression rate. Such kind of bound is shown as rate-distortion function, $R(D)$, which is monotone decreasing and downwards convex function.

From the above approach, training set from compressed instance space is regarded as codewords. In this case, compression-rate is regarded as sample size. If, however, compression process of instance space is allowed to be re-arranged e.g. any combination of examples or re-construction of attributes, this approach will be closest to the condition of rate-distortion theory. Basically, assuming the same instance space and

*This work was partially supported by the Ministry of Education under a Grant-in-Aid for Scientific Research No.06680364, and a grant from the Research Institute of Aoyama Gakuin University

†Collage of Science and Enginering, Aoyama Gakuin University, 6-16-1 Chitosedai, Setagaya-ku, Tokyo 157 Japan

‡Faculty of Engineering, Kobe University, 1-1 Rokkodai, Nada, Kobe 657 Japan

§Faculty of Science, University of Ottawa, 150 Louis Pasteur/Priv., Ottawa, Ontario Canada K1N 6N5

0-7803-2129-4/94 $3.00

the same accuracy, $1-D$, if compression-rate, $R_1(D)$ is lower than $R_2(D)$, the former is potentially better than the latter, which guarantees the existence of better learning algorithm in the former than the latter.

What is the potential property of the concepts space? When the large scale problems are assumed, the system efficiency is defined by the normalized rate-distortion function $R(D,n)/R(0,n) < 1, 0 < D < D_{max}(n) < 1$, with the parameter of the problem size, n, e.g. the number of attributes in the case of Boolean concepts. The behaviour of the system efficiency will be evaluated as the system size becomes infinite. When the system efficiency becomes zero, such a system will be termed elastic or trivial elastic, which assures the existence of algorithms, assuming sufficiently large problem size, that a target concept can be guessed with a given accuracy from a highly compressed instance space. In elastic cnodition $R(D,n)/R(0,n)$ shows highly divergence speed for the problem size, n, and in trivial elastic condition $D_{max}(n)$ also shows highly divergence speed for the problem size [8, 9]. As a result, we evaluate the potential property of the class of concepts by $R(D,n)/R(0,n)$ and $D_{max}(n)$.

As examples,the theoretical bounds on approximate learning ofsome concept classes are proposed by using the following rate-distortion theoretical framework:

1. The problem is how to compress instance space in order to guess target concepts with a given accuracy.

2. The average compression-rate of instance space with a given accuracy is regarded as the minimum mutual information between the original instance space and the compressed one with a given error-rate.

3. The rate-distortion function, showing the trade-offs between compression-rate and error-rate, identify the potential property of the class of concept.

4. The behavior in the limit of the normalized rate-distortion functions shows either elastic, trivial elastic, or inelastic condition. In the case of elastic or trivial elastic condition, the divergence speed of $R(D,n)/R(0,n)$ or $D_{max}(n)$ is evaluated.

Considering the previous works, it is shown that the sample size function satisfying PAC learnability is derived from VC dimension of the class of the concepts. Although not refering strictly the sample size and learning algorithms, our criterion, $R(D,n)/R(0,n)$ and $D(n)$, also shows the the potential property of the concept class from the viewpoint of the relationship between sample complexity and accuracy. In the future, the classification of learnability for the class of concepts will be realized by using our criterion.

2 PAC Learnability

The following notions of learning functions and learnability is used in PAC learning framework [2, 6].

Definitions: A concept class is nonempty set $C \subseteq 2^X$ of concepts. It is assumed that X is a fixed set, either finite, countably infinite, $[0,1]^n$, or E^n (Euclidean n dimensional space) for some $n \geq 1$. In the latter cases, we assume that each $c \in C$ is a Borel set. X^m denotes the m-fold Cartesian product of X. For $\bar{x} = (x_1, x_2, ..., x_m) \in X^m, x_i \in X, 1 \leq i \leq m$, the m-sample of $c \in C$ generated by $\bar{x}$ is given by $smp_c(\bar{x}) = (\langle x_1, I_c(x_1)\rangle, \langle x_2, I_c(x_2)\rangle ..., \langle x_m, I_c(x_m)\rangle)$. $I_c(x)$ denotes the indicator function for c on X, that is, $I_c(x_i) = 1$, if $x_i \in c, I_c(x_i) = 0$, otherwise. The sample space of C, denoted S_C, is the set of all m-samples over all $c \in C$ and all $\bar{x} \in X^m$, for all $m \geq 1$.

$\mathbf{A}_{C,H}$ denotes the set of all functions $A : S_C \rightarrow H$, where H is a set of Borel sets on X. H is called the hypothesis space. Elements in H are called hypotheses. The hypothesis space is usually and also throughout this paper assumed to be C itself, although in some cases it is computationally advantageous to allow A to approximate concepts in C using hypotheses from a different class H. $A \in \mathbf{A}_{C,H}$ is consistent if its hypothesis always agrees with the sample, that is, whenever $h = A(\langle x_1, a_1\rangle, ..., \langle x_m, a_m\rangle)$ then for all i, $1 \leq i \leq m, a_i = I_c(x_i)$. For any learning function $A \in \mathbf{A}_{C,H}$, probability distribution P on X, $c \in C$, and $\bar{x} \in X^m$, let $c\Delta h$ denotes the symmetric difference of the target concept and the hypothesis, the error of A for concept c on $\bar{x}$ with respect to P is given by $error_{A,c,P}(\bar{x}) = P(c\Delta h)$, where $h = A(smp_c(\bar{x}))$. Thus, A's error is measured as the probability of the region that forms the symmetric difference between the target concept and A's hypothesis, which is just the probability that A's hypothesis will be inconsistent with the target concept on randomly drawn point with respect to P.

Let $m(\epsilon,\delta)$ be an integer-valued function of ϵ and δ for $0 < \epsilon, \delta < 1$ and, P be a probability distribution on $X, A \in \mathbf{A}_{C,H}$ is a learning function for C with sample size $m(\epsilon,\delta)$ if for all $0 < \epsilon, \delta < 1$ and for all $c \in C, P^{m(\epsilon,\delta)}(W) \leq \delta$, where $W = \{\bar{x} \in X^{m(\epsilon,\delta)} : error_{A,c,P}(\bar{x}) > \epsilon\}$. It is insisted that using a randomly drawn sample of size $m(\epsilon,\delta)$ of any target concept in

C. A produces, with probability at least $1-\delta$, a hypothesis in H with error no more than ϵ. If such an A exists, it is said that C is uniformly learnable by H under the distribution P. The smallest sample size $m(\epsilon,\delta)$ is called the sample complexity of A.

Definitions: Given nonempty concept class $C \subseteq 2^X$ and a set of points $S \subseteq X$. $\Pi_C(S)$ denotes the set of all subsets of S that can be obtained by intersecting S with a concept in C, that is, $\Pi_C(S) = \{S \cap c : c \in C\}$. If $\Pi_C(S) = 2^S$, then it is said that S is shattered by C. The Vapnik-Chervonenkis (VC) dimension of C is the cardinality of the largest finite set of points $S \subseteq X$, that is shattered by C. If arbitrarily large finite sets are shattered, the VC dimension of C is infinite. For any integer $m \geq 0, \Pi_C(m) = \max(|\Pi_C(S)|)$ over all $S \subseteq X$ of cardinality m. That is, VC dimension of C can be defined as the largest integer d such that $\Pi_C(d) = 2^d$, or infinity.

Let C be any finite concept class. Then since it requires 2^d distinct concepts to shatter a set of d points, no set of cardinality larger than $\log|C|$ can be shattered. Hence, the VC dimension of C is at most $\log|C|$.

According to the above definitions, It is shown that a characterization of polynomial learnabiblity with respect to domain dimension. Let the concept classes $C_n \subseteq 2^{E^n}$ be all domains of Euclidean dimension $n \geq 1$ and for each n, and $C_n \subseteq 2^{\{0,1\}^n}$ be all domains of Boolean dimension $n \geq 1$ and for each n. It is shown that the concept classes $C_n, n \geq 1$, are polynomially learnable if and only if the VC dimension of C_n grows polynomially in n and there exists a polynomial time probabilistic algorithm for finding a consistent hypothesis in C_n for any sample of a target concept in C_n. Especially in the Boolean case, the concept classes $C_n \subseteq 2^{\{0,1\}^n}$, $n \geq 1$, are polynomially learnable if and only if $\log|C_n|$ grows polynomially in n and there exists a polynomial time probabilistic algorithm for finding a consistent hypothesis in C_n for any sample of a target concept in C_n. From the other aspects, a sufficient condition of polynomial learnability with respect to target complexity based on the principle of preferring the simpler hypothesis, usually called Ocam's Razor. This result shows that if we can efficiently produce a hypothesis that explains the sample data, and is sufficiently more compact than the sample data, then we can feasibly learn, which may be interpreted as showing a relationship between a kind of data compression and learning. Note that this information theoretic approach is based on Minimum Description Length (MDL) of source coding theorem without distortion.

3 Rate Distortion Theory

The problem is formalized using the following source model. Let $\mathbf{U} = \{u_1, u_2, \ldots, u_n\}$ be a discrete memoryless source $\mathbf{X}$, and $p_1, p_2, \ldots, p_n$ be their probabilities. We assume throughout this paper that n is finite and that $p_i > 0$ for each $i, i = 1.2, \ldots, n$. The source output is a sequence $x_1, x_2, \ldots$ of independent selections from the given alphabet with their given probabilities. The source sequence is to be represented at the destination by a sequence of letters $y_1, y_2, \ldots$, each selected from a destination alphabet, $\mathbf{V} = \{v_1, v_2, \ldots, v_m\}$, where m is finite. Let p_i be the a priori probability of the input alphabet u_i, and $P(j|i)$ be the transition probability of the output alphabet v_j. Let $\rho(i,j)$ be the distortion measure which is defined for $i = 1, 2, \ldots, n, j = 1, 2, \ldots, m$, assigning a numerical value to the distortion, if source alphabet u_i is represented at the destination by alphabet v_j. Then both an average mutual information and an average distortion are determined, and the rate distortion function, $R(D)$, of the source relative to the given distoriton measure is defined as

$$R(D) = \min_{P(j|i) \in \mathbf{P}_D} \sum_{i,j} p_i P(j|i) \log \frac{P(j|i)}{\sum_j p_i P(j|i)} \quad (1)$$

where

$$\mathbf{P}_D = \{P(j|i) \mid \sum_{i,j} p_i P(j|i) \rho(i,j) \leq D\}. \quad (2)$$

The minimization in (1) is over all assignments of transition probabilities subject to the constraint that the average distortion is less than or equal to the average distortion, D. Regardless of what processing is done, the average distortion must exceed D, if a channel with capacity less than $R(D)$ nats per source symbol connects the source to the destination for given D. It is reasonable to interpret $R(D)$ as the rate of the source, in nats per symbol, relative to the fidelity criterion D.

Considering the above results, some general peoperties of this function have been summarized as follows. First of all, $R(D)$ is nonnegative, nonincreasing, and convex in D. The nonnegativeity is obvious, since the average mutual information is nonnegative. Observe next that the minimization in (1) is over a constraint set which is enlarged as D is increased. Thus the resulting minimum $R(D)$ is nonincreasing with D. Next, let D_{max} be the smallest D for which $R(D) = 0$, we can calculate D_{max} from

$$D_{max} = \min_i \sum_j p_i \rho(i,j). \quad (3)$$

It is noted that the smallest possible value for the average distortion is zero and is achieved by mapping each letter u_i of the source alphabet into an output letter v_j for which $\rho(i,j) = 0$. For $D < 0, R(D)$ is undefinded, since by definition result, $\rho(i,j) \geq 0$.

For a given source $\mathbf{U}$ and destination alphabet $\mathbf{V}$, a source code of M code words with block length L is defined as a mapping from the set of source sequences of length L into a set of M code words, where each code word, $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iL})$ is a sequence of L letters from the destination alphabet. The average distortion, d_L, per letter of a source code is given by

$$d_L = \frac{1}{L} \sum_{j} P_L(\mathbf{x}) \rho(\mathbf{x}, \mathbf{y}(\mathbf{x})) \tag{4}$$

where $P_L(\mathbf{x})$ is the probability of a source sequence $\mathbf{x} = (x_1, x_2, \ldots, x_L)$, and $\mathbf{y}(\mathbf{x})$ is the code word that $\mathbf{x}$ is mapped into such that

$$\rho(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{L} \rho(x_k, y_k). \tag{5}$$

At the point to find how small d_L can be made for a given L and M, and we shall attempt to analyze the behavior of a randomly chosen set of code words. Let $P(j|i)$ be a given set of transition probabilities between source and destination letters. Considering a discrete memoryless channel with these transition probabilities as a test channel, if $P(j|i)$ achieves $R(D)$ for a given D, then the associated test channel will achieve $R(D)$ for D. The output probabilities, $p(j)$, for given test channel, are

$$p(j) = \sum_{i} P(j|i), j = 1, 2, \ldots, m. \tag{6}$$

For any given test channel, we consider an ensemble of source codes in which each letter of each code word is chosen independently with the probability assignment $p(j)$. For a given set of code words $\mathbf{y}_i, i = 1, 2, \ldots, M$, in the ensemble, each source sequence $\mathbf{x}$ will be mapped into that code word $\mathbf{y}_i$ for which $\rho(\mathbf{x}, \mathbf{y}_i)$ is minimized over i.

Considering simultaneously two different probability measures on the input and the output sequences, one is the test channel ensemble and the other is the random coding ensemble. For the test channel ensemble, the probability measure on input sequences $\mathbf{x} = (x_1, x_2, \ldots, x_L)$ and output sequences $\mathbf{y} = (y_1, y_2, \ldots, y_L)$ is given by $P_L(\mathbf{x})P_L(\mathbf{y}|\mathbf{x})$ where

$$P_L(\mathbf{x}) \prod_{k=1}^{L} P(x_k), \tag{7}$$

$$P_L(\mathbf{y}|\mathbf{x}) = \prod_{k=1}^{L} P(y_k|x_k), \tag{8}$$

$$P_L(\mathbf{y}) = \sum_{\mathbf{x}} P_L(\mathbf{x}) P_L(\mathbf{y}|\mathbf{x}) = \prod_{k=1}^{L} P(y_k). \tag{9}$$

In these equations, $P(x_k)$ and $P(y_k|x_k)$ are the source and test channel probabilities respectively. The other ensemble is the ensemble of codes in which M code words are independently chosen with the probability assignment $P_L(\mathbf{y})$, the source sequence is chosen with the same assignment $P_L(\mathbf{x})$ as above, and for each code in the ensemble, $\mathbf{x}$ is mapped into that $\mathbf{y}_i$, denoted $\mathbf{y}(\mathbf{x})$, that minimizes $\rho(\mathbf{x}, \mathbf{y}_i)$ over $1 \leq i \leq M$.

Following the above condition, we state both the negative and positive parts of source coding theorem.

Theorem 1 [7]: Let $R(D)$ be the rate distortion function of a discrete memoryless source with a finite distortion measure. For any $D \geq 0$, any $\epsilon \geq 0$, and any sufficiently large block length L, there exists no source code with $M \leq \exp\{LR(D)\}$ code words for which the average distortion per letter satisfies $d_L < D + \epsilon$. From the other aspect, there exists a source code with $M \leq \exp\{L(R(D) + \epsilon)\}$ code words for which the average distortion per letter satisfies $d_L \leq D + \epsilon$.

4 A New Test for Learnability

The interests of PAC learnability, e.g. sample complexity and learning algorithms, focuses not guessing of each target concept but the class of concepts. We also analyze learnability of concept classes from the viewpoint of their potential property.

Considering instance space, the sample space S_C, the set of all m-samples over all $c \in C$ for all $m \geq 1$ can be regarded as the compressed instance space. In order to evaluate the compressed instance space using rate distortion theory, coding procedure must be defined to learning process.

Difinition: Let X be the instance space, $|X|(= N)$ be the cardinality of X, and $|C|(= K)$ be the cadinality of C. The m-sample of $x_i \in X$ generated by C is given by $smp_{x_i}(C) = (\langle c_1, I_{x_i}(c_1)\rangle, \langle c_2, I_{x_i}(c_2)\rangle, \ldots, \langle c_K, I_{x_i}(c_K)\rangle), c_j \in C$, where $I_{x_i}(c_j) = 1$, if $x_i \in c_j, I_{x_i}(c_j) = 0$, otherwise. We define source and destination alphabet of the class C of concepts to be $I_{x_i}(C), i = 1, 2, \ldots, N$, and distortion measure of the class C of concepts, $\rho(i,j)$, to be $\sum_{k=1}^{K} |I_{x_i}(c_k) - I_{x_j}(c_k)|/K$. Then coding procedure is as follows.

$$C \overset{source-encoding}{\longrightarrow} S_C \overset{source-decoding}{\longrightarrow} H$$

$$\{I_{x_i}(C)_{i=1}^N\} \rightarrow \{I_{x_j}(C)_{j=1}^m\} \rightarrow A(\{I_{x_j}(C)_{j=1}^m\})$$

where A is the guessing function, that is, $A : S_C \rightarrow H$.

For example, assuming that instance space X consists of 4 points, x_1, x_2, x_3, x_4, the sample space S_C is the set of all 2-samples, and block code length L is 16, the following mapping is one of the coding procedure with distortion. That is, practically, source sequences of length 16 means 4-times guessing processes to target condcepts without distortion, and source code of length 16 means 8-times guessing processes to target concepts with some distortion. Note that the average distortion can be counted after decoding process by the guessing function A.

$$\{x_1x_2x_3x_4x_1x_2x_3x_4x_1x_2x_3x_4x_1x_2x_3x_4\}$$
$$\downarrow$$
$$\{x_1x_2\underline{x_1x_2}x_3x_4\underline{x_3x_4}x_1x_3\underline{x_1x_3}x_2x_4\underline{x_2x_4}\}$$

In this case, block code length means the frequency of learning, and sufficiently large block code length would realize the estimation of the average performance of learnability of a given concept class for any sample size. Therefore, rate distortion function would clarify the potential property of the concept class regardless learnig algorithms.

Practically, rate distortion function can be derived for any alphabet size, N, as follows. If u_s and v_s are functions of s such that

$$u_s \geq \max_{1 \leq j \leq N} \sum_{i=1}^{N} e^{s\rho(i,j)}, s \leq 0, \quad (10)$$

$$v_s \leq \min_{1 \leq j \leq N} \sum_{i=1}^{N} e^{s\rho(i,j)}, s \leq 0, \quad (11)$$

and both of them are differentiable and log-convex, then a lower bound and a upper bound to $R(D)$, $R_L(D)$ and $R_U(D)$, are derived as the following set of parametric equations.

$$R_L(D) = \log N + sD - \log u_s, D = \frac{d}{ds} \log u_s, \quad (12)$$

$$R_U(D) = \log N + sD - \log v_s, D = \frac{d}{ds} \log v_s. \quad (13)$$

Example 1: Let X be the real line and let C be the set of all intervals(open or closed) on X. Then given any set S consisting of two points $x_1, x_2 \in X$, we can find concepts $c_1, c_2, c_3, c_4 \in C$ such that $c_1 \cap S = \{x_1\}, c_2 \cap S = \{x_2\}, c_3 \cap S = \emptyset$, and $c_4 \cap S$. If S consists of three points $x_1 \leq x_2 \leq x_3$, then there is no concept $c \in C$ that contains x_1 and x_3 but not x_2. Thus the above problem is a typical case that VC dimension of C is 2. If, however, N-points $x_1, x_2, \ldots, x_N \in X$, are considered, we can find $|C| = 2N$ concepts. When, for example, $N = 4$ we show the input alphabet as follows.

$$I_{x_1}(C) = (0\,0\,0\,0\,1\,1\,1\,1)$$
$$I_{x_2}(C) = (0\,0\,0\,1\,0\,1\,1\,1)$$
$$I_{x_3}(C) = (0\,0\,1\,1\,0\,0\,1\,1)$$
$$I_{x_4}(C) = (0\,1\,1\,1\,0\,0\,0\,1)$$

To calculate the average distortion, the distortion matrix, $[\rho(i,j)]$, must be derived. The j-th column of the distoriton matrix is

$$\begin{pmatrix} (j-1)/N \\ \vdots \\ 1 \\ 0 \\ 1 \\ \vdots \\ (N-j)/N \end{pmatrix}$$

and taking $z = e^{s/N}$,

$$v_s^j = z^{j-1} + \cdots + z + 1 + z + \cdots + z^{N-j}$$

A lower bound to v_s^j independent on j can be derived, such that

$$v_s^j \geq 1 + z + \cdots + z^{N-1} = \frac{1 - z^N}{1 - z} = v_s,$$

since $z^i \geq z^{N-j+i}$, for $i, j \geq N$, and $z + \cdots + z^{j-1}$ can be replaced by $z^{N-j+1} + \cdots + z^{N-1}$. Therefore, assuming sufficiently large alphabet size N, asymptotic behaviour of rate distortion function satisfies

$$R(D) \approx \log \frac{N}{N+t} - tD + \log \frac{t}{1 - e^{-t}}, t > 0, \quad (14)$$

where

$$D = \frac{1}{t} - \frac{1}{e^t - 1}, t > 0.$$

When the large scale problems are assumed, the system efficiency is defined by the normalized rate distortion function $R(D,N)/R(0,N)$. If for any $D > 0$, $R(D,N)/R(0,N) \rightarrow 0$ as $N \rightarrow \infty$, such a condition assure that for given compression-rate, the reduction of error-rate ($= 1 - accuracy$) is feasible, assuming large scale problems. Note that the compression-rate is approximately equivalent to the sample size. Thus in accordance with the behaviour of the system efficiency, the potential property can be classified by elastic if for any $D > 0$ $R(D,N)/R(0,N) \rightarrow 0$ as $N \rightarrow \infty$, and

inelastic, if $R(D,N)/R(0,N) > 0$ as $N \to \infty$. Note that trivial case , $D_{max}(N) \to 0$ as $N \to \infty$, is also included in elastic, denoted trivial elastic. The elasticity test was first proposed by J.Pearl in Question-Answering Systems, which has evaluated the problems of memory versus error trade-offs.

In Example 1, the system efficiency, such that

$$\frac{R(D,N)}{R(0,N)} \approx \frac{\log \frac{N}{N+t} - tD + \log \frac{t}{1-e^{-t}}}{\log N}, t > 0 \quad (15)$$

shows elastic condition, i.e. the fastest possible rate of convergence.

We consider several classes of Boolian concepts.

Example 2: Boolian Perfect Concept is the full class of disjunctive normal form (DNF) consisting of any arbitrary Boolean expression. Over n Boolian variables, the size of the instance space N is 2^n, and the size of the concept class is 2^{2^n}. When, for example, $n = 2$, we show the input alphabet as follows.

$$I_{x_1}(C_{BP}) = (0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,1\,1\,1\,1\,1\,1)$$
$$I_{x_2}(C_{BP}) = (0\,0\,0\,0\,1\,1\,1\,1\,0\,0\,0\,0\,1\,1\,1\,1)$$
$$I_{x_3}(C_{BP}) = (0\,0\,1\,1\,0\,0\,1\,1\,0\,0\,1\,1\,0\,0\,1\,1)$$
$$I_{x_4}(C_{BP}) = (0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1)$$

The average distortion is very simple, such that

$$v_s = 1 + \sum_{i \neq j} e^{s/2} = 1 + (2^n - 1)e^{s/2}.$$

Thus the system efficiency is derived as follows.

$$\frac{R(D,n)}{R(0,n)} \approx 1 - 2D, 0 \leq D < \frac{1}{2}, \quad (16)$$

which shows inelastic.

Example 3: Boolian conjuctions is also typical Boolian concept. Over n Boolian variables, the size of the instance space N is 2^n, and the size of the concept class is 3^n, When, for example, $n = 2$, we show the input alphabet as follows

$$I_{x_1}(C_{BC}) = (1\,0\,0\,0\,0\,0\,1\,1\,1\,1\,)$$
$$I_{x_2}(C_{BC}) = (1\,0\,1\,0\,0\,1\,0\,1\,0\,0\,0\,)$$
$$I_{x_3}(C_{BC}) = (1\,1\,0\,0\,1\,0\,0\,0\,1\,0\,1\,)$$
$$I_{x_4}(C_{BC}) = (1\,1\,1\,1\,0\,0\,0\,0\,0\,1\,0\,)$$

The average distortion, and the system efficiency is derived as follows.

$$u_s = \sum_{i=0}^{n} \binom{n}{i} e^{2^{n+1-i}(2^i-1)s/3^n},$$

$$\frac{R(D,n)}{R(0,n)} \approx 1 - \frac{1}{2}(\frac{3}{2})^n D, \; 0 \leq D < 2(\frac{2}{3})^n. \quad (17)$$

which shows trivial elastic.

From some examples, the concept classes with finite VC dimension often show elastic. In the future, the relationship between VC dimension and elasticity, and convergence speed in the case of elastic condition will be strictly analysed.

Acknowledgement

One of the authors wish to thank Prof.Stan Matwin for providing nice conditions at Machine Learning Group, University of Ottawa, and for his helpful comments, and also to thank Prof.S.Hirasawa for providing a chance to start a study of some applications of rate-distortion theory and for his helpful comments.

References

[1] J.R.Quinlan,and R.L.Rivest, "Inferring decision trees using the minimum description length principle," *Informatin and computation,* , vol.80, pp.227-248, 1989.

[2] L.G.Variant, "A theory of the learnable," *Communications of ACM* vol.27, no.11, pp.1134-1142, 1984.

[3] H.Almuallim and T.G.Dietterich, "Learning with many irrelevant features," *Proceedings of AAAI-91*, pp.547-552, 1991.

[4] V.N.Vapnik and A.Ya.Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and its Applications*, vol.16, no.2, pp.264-280, 1971.

[5] V.N.Vapnik, *Estimation of Dependences Based on Empirical Data.* Springer Verlag, New York, 1982.

[6] A.Blumer, A.Ehrenfeucht, D.Haussler and M.Warmuth, "Learnability and the Vapnik- Chervonenkis dimension," *Journal of ACM* , vol.36, no.4, pp.929-965, 1989.

[7] T. Berger, *Rate Distortion Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1971.

[8] A.Crolotte and J.Pearl, "Elasticity conditions for storage versus error exchange in question-answering systems," *IEEE Trans. on Information Theory*, vol.IT-25, no.6, pp.653-664, 1979.

[9] H.Inazumi, "Studies on the evaluations for information systems based on rate distortion theory," Dr.Eng. dissertation, Waseda University, 1989.