Original Paper — *Forma*, **14**, 303–307, 1999

Mathematical Study of Conceptualization in *N*-Dimensional Space

Masanobu OHTSUKI, Atsushi MINATO and Satoru OZAWA

Division of Applied Synergetics, Graduate School of Science and Engineering, Ibaraki University, 4-12-1 Nakanarusawa, Hitachi 316-8511, Japan

(Received November 19, 1999; Accepted January 13, 2000)

Keywords: Learning, *N*-Dimensional Rectangle, High Dimensional Space, Pattern Recognition

Abstract. A concept of a thing is characterized by a set of suitable parameters. It can be approximately represented as a rectangle in *n*-dimensional space. The process of obtaining the concept from a number of observations produces a series of "sample points" in *n*-dimensional space, where each sample point has a value of 1 or 0 depending upon whether the sample suits with the concept or not. The accuracy of guessing *n*-dimensional rectangle from sample points is related to the way of generating samples. The HIS (Half Interval Search) method is proposed to produce the sample points. It has been shown that it is an efficient method and the expected error in the determination of the concept is $O(1/m)$, where m is the number of sample points. This method is successfully used for identification of a "smiling face" in a cartoon figure.

1. Introduction

A concept of a thing is obtained through a number of observations. The observed thing is characterized by a set of properties. We assume that the properties can be expressed quantitatively by a set of parameters. The values of the parameters are normalized by their maximum. When we use these n parameters as n axes, the result of the observation can be represented in n-dimensional space, $[0, 1]^n$, where $[0, 1]$ means the set of real numbers between 0 and 1. We call this space as a "sample space" and points in the space as "sample points".

A sequence of observations produces a series of sample points in $[0, 1]^n$ and each sample point has the value of a unit or zero. The unit value means that the observed object suits with the concept and the zero value means that the object is not associated with the concept.

The concept of the thing is represented as a volume in which the sample points take a unit value. In this paper, we assume that the shape of the volume is an n-dimensional rectangle in $C_n \stackrel{def}{=} \{[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]; 0 \le a_i \le b_i \le 1, i = 1, 2, ..., n\}$. The process of obtaining the concept through the observation is thus modeled as a problem of guessing at the n-dimensional rectangle from a series of sample points.

Let us consider a problem how precisely we can guess an n-dimensional rectangle from the limited number, m, of sample points. For this purpose, we introduce the "error in the guess" which is defined as the volume of the symmetric difference between the guessed n-dimensional rectangle and the true n-dimensional rectangle. The error in the guess is mathematically expressed as c_n(guessed) $\triangle c_n$(true), where $\triangle$ is the symmetric difference operator, c_n(guessed) represents the concept guessed from observations, and c_n(true) represents the true concept.

Let us note that the error in the guess is closely related to the way of generating sample points (VALIANT, 1984; LAIRD, 1988; BLUMER *et al*., 1989; NATARAJAN, 1991). If the sample points are produced at random, the error in the guess is found to be O ($2n \ln m / m$) (OHTSUKI, 1998). It is expected that the error in the guess can be reduced when we use samples with correlation. In this paper, we propose a new method of generating correlated sample points from which a concept of a thing is defined efficiently. We call it the half interval search (HIS) method. We study the error in the guess included in this method. The HIS method is applied to defining the concept of a "face" and a "smiling face" in a cartoon figure.

2. Half Interval Search Method

As seen in the previous section, a concept of a thing is approximately represented by an n-dimensional rectangle in $[0, 1]^n$. Here we propose an efficient method named HIS to determine the n-dimensional rectangle.

The essence of this method is to examine samples which lie near the surface region of the n-dimensional rectangle. The algorithm of the generating sample points by the HIS method is expressed in Pascal language as the following. In the first stage of this method, we find out a sample point of a unit value (which suits with the concept). This process is carried out by a random searching. In the second stage, we examine repeatedly a sample at the "half interval position". The half interval position is defined as the central point between a sample point whose value is found to be a unit (a sample point inside the n-dimensional rectangle) and the sample point whose value is found to be zero (the sample point is at the outside of the rectangle).

We have shown that the expected error in the guess $E[e\mathcal{A}_n(m)]$ contained in this algorithm is estimated as $k_1/m^2 \leq E[e\mathcal{A}_n(m)] \leq k_2/m$, where k_1 and k_2 are constants. (OHTSUKI, 1999)

——— $\mathcal{A}_n$; Algorithm of HIS method ———

```
begin
read (m); {* the 1st stage: finding a sample point which suits with the concept * }
    q: = 0; {* q is the number of sample points investigated in the 1st stage * }
    repeat
        begin
            y: = I_c(x) ; {* I_c() is a sample generator, x is taken at random.*}
            q: = q + 1
        end;
    until (y = 1) or (q = m); {* end of the 1st stage, start of the 2nd stage * }
```

```
    if q = m
        then h: = φ
        else {* x = (x_1, ..., x_n) is the sample point obtained in the 1st stage * }
            begin
                p: = (m–q)/(2*n);
                for i: = 1 to n do
                    begin
                        examine the sample point (x_1, ..., x_{i–1}, a, x_{i+1}, ..., x_n)
                        at the half interval position, a
                    end;
                determine the smallest n-dimensional rectangle in which all of
                the sample points of a unit value are included
            end;
    write (the determined n-dimensional rectangle)
end.
```

3. Applications

The method developed in the previous sections has general character and is applied in many fields of sciences. In this section, we try to use it for recognition of a "smiling face" in cartoon figures.

Let us begin with an experiment of obtaining a concept of a "face" in a cartoon figure. We assume in our experiment that a face is composed of two eyes and a nose expressed by filled circles and a mouth expressed by a filled rectangle as shown in Fig. 1. The relative positions of the components are represented by the parameters $[x_1, x_2, ..., x_6]$. Here, the nose is placed at the center of the face. When we select suitable values for these parameters, the figure looks like a face. Then, we have a sample point $\boldsymbol{x} = [x_1, x_2, ..., x_6]$ of unit value (a sample point which suits with the concept of a "face"). If the parameters do not have good values, it does not look like a face and we have a sample point of zero value. The examples of these samples are shown in Fig. 2. We produced $m = 200$ sample figures by using the HIS method with the aid of computer graphics routine. From the 200 sample points, we guessed a 6-dimensional rectangle which represents the concept of a face in our experiment. The obtained rectangle is $[0.25, 0.50] \times [0.25, 0.50] \times [0.31, 0.75] \times [0.22, 0.50] \times [0.38, 0.50] \times [0.31, 0.75]$.

Now let us apply our method to recognize a "smiling face" in a cartoon figure. For this purpose, we must firstly obtain the concept of a "smiling face" in cartoon figures. This time, we assume that two eyes and the mouth are expressed by arcs, and the nose by a filled circle. The arc is a circle having a definite radius. The position and the shape of the arc are expressed by the center of the circle and two angle parameters which represent two ends of the arc. We introduced total 12 parameters to represent the figure. We carried out the HIS experiment. The number of sample points produced is $m = 400$. The samples are shown in Fig. 3. It has been found from the experiment that the 12-dimensional rectangle expressing a "smiling face" is $[0.07, 0.82] \times [0.26, 0.51] \times [0.00, 0.11] \times [0.33, 0.48] \times [0.16, 0.87] \times [0.10, 0.71] \times [0.00, 0.13] \times [0.37, 0.69] \times [0.30, 0.72] \times [0.41, 0.92] \times [0.33, 1.00] \times [0.00, 0.09]$.

Next, we used this rectangle for recognition of an emotional expression seen in a cartoon figure. When a cartoon figure is given, we measure the 12 parameters. If the obtained data produces a sample point inside the 12-dimensional rectangle, we can conclude that it is a figure of a smiling face. We carried out this type of the test for 100 cartoon figures. The examples of a cartoon figure used in the experiment are shown in Fig. 4. The sample is chosen at random from a library of digital articles. It has been found that the judgements are 95% successful. The 5% error is caused by various assumptions introduced in the method. We are now studying the possibility to express the concept using an *n*-dimensional polygon.

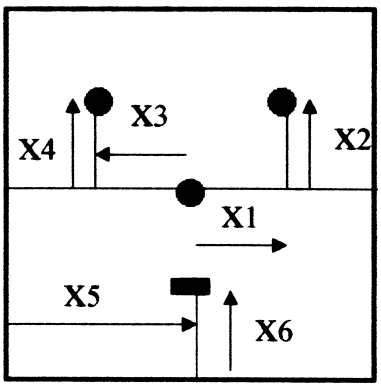

Fig. 1. Parameters of a face.

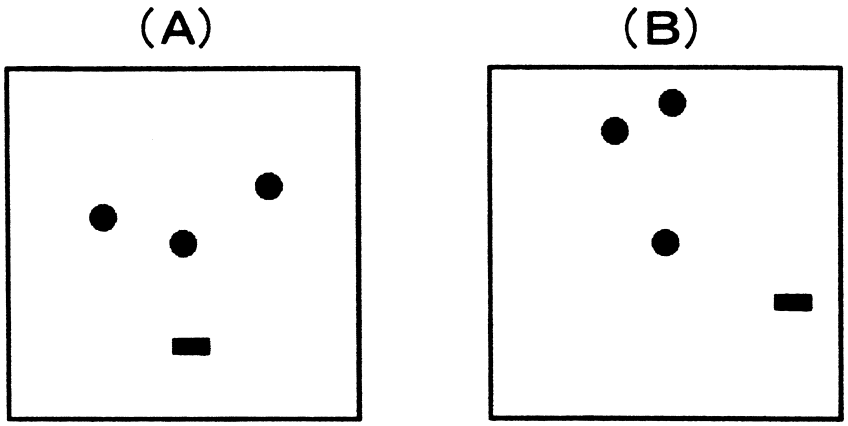

Fig. 2. (A) a sample that looks like a face, (B) a sample that does not look like a face.

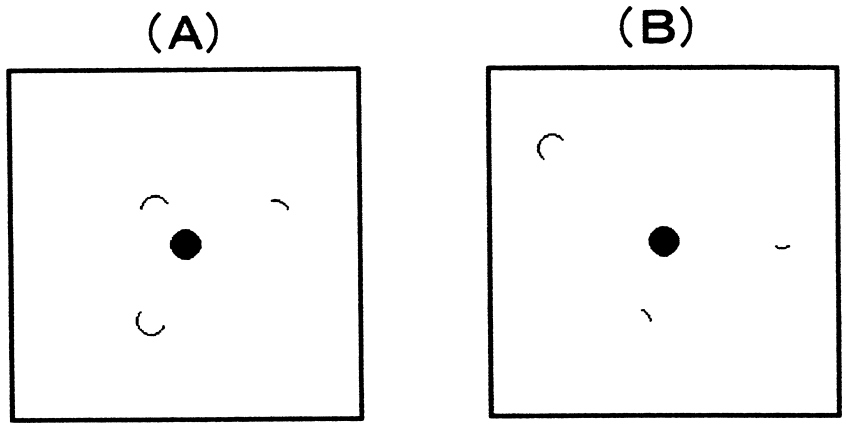

Fig. 3. (A) a sample that looks like a smiling face, (B) a sample that does not look like a smiling face.

Fig. 4. 5 examples of a cartoon figure.

REFERENCES

BLUMER, A., EHRENFEUCHT, A., HAUSSLER, D., MANFRED, K. and WARMUTH, K. (1989) Learnability and the Vapnik-Chervonenkis dimension, *Journal of the ACM*, **36**, No.4, 929–965.

LAIRD, P. D. (1988) *Learning from Good and Bad Data*, Kluwer Academic Publishers, Boston, pp. 1–186.

NATARAJAN, B. (1991) *MACHINE LEARNING*, Morgan Kaufmann Publishers, Inc. San Mateo, California, pp. 1–217.

OHTSUKI, M. (1998) An algorithm for learning *n*-dimensional rectangles with the uniform distribution, *Research Reports Fukushima National College of Technology No.38*, 7–19.

OHTSUKI, M. (1999) A half-interval-search algorithm for learning *n*-dimensional rectangles, *Research Reports Fukushima National College of Technology No.39*, 18–27.

VALIANT, L. G. (1984) A theory of the learnable, *Communications of the ACM*, **27**, No.11, 1134–1142.