Kybernetika

Jean Sallantin; Thierry Pyl

On the Bayesian inductive processes

Kybernetika, Vol. 19 (1983), No. 4, 319--334

Persistent URL: http://dml.cz/dmlcz/125095

KYBERNETIKA — VOLUME *19* (1983), NUMBER 4

ON THE BAYESIAN INDUCTIVE PROCESSES

JEAN SALLANTIN, THIERRY VAN DER PYL

The inductive Bayesian processes deal with the Bayes formula and with a concatenation rule of the tests. Generally the tests are supposed independent in probability to allow multiplication as the concatenation law. Then the average inverse H-theorem is true and the entropy decreases after a certain number of tests to get a limiting value. We introduce a one parameter family of semigroups instead of multiplication, to study the stability of Bayesian processes. We show that the decrease of the entropy is no more assured, but the convergence is still conserved. Thus, in the case when the independency of tests is not supposed we must exhibit a semigroup of the family. That is, we must adjust the theoretical results obtained for a semigroup to the experimental ones.

We see on an example that the principle of minimum entropy allows to select a solution among all the possible ones (selection of a new test, selection of a semi-group).

1. INTRODUCTION

The Bayesian inductive processes appear in pattern recognition every time the description of an object is not immediately available: the problem is then to construct an object representation sequentially, allowing to decide what concept is verified by the object [8]. At each step the questions arise 1) of the choice of the new test to be done to complete representation 2) of the use of the new information so obtained to decide on the concept verified by an object 3) to know when to stop.

Generally one suppose that the tests are statistically independent so that the probability $\mu(\mathbf{Y}_1 \wedge \mathbf{Y}_2 \mid A_i)$, called deductive similarity, for an object verifying concept A_i to give the issues $\mathbf{Y}_1$ to test $\mathbb{Y}_1$ and $\mathbf{Y}_2$ to test $\mathbb{Y}_2$, is equal to the product of probabilities $\mu(\mathbf{Y}_1 \mid A_i)$ and $\mu(\mathbf{Y}_2 \mid A_i)$.

With the Bayes formula one deduces $\mu(A_i \mid \mathbf{Y}_1 \wedge \mathbf{Y}_2)$ called inductive similarity between representation $\mathbf{Y}_1\mathbf{Y}_2$ and concept A_i.

The operating method is justified by the fact that on the average, when the representation becomes big enough, one can decide what concept is verified by the object:

this is expressed by the expected inverse H-theorem by Watanabe [10], Watanabe and al. [12] and the convergence theorems of the pseudoquestionnaires by Terrenoire [9].

After having defined the general principles of the inductive methods (Sections 2 and 3), we examine each step when we do not suppose the test independency any more then we introduce a concatenation law which is no longer the product (Section 4). We look at the asymptotic behaviour of the inductive similarities in Section 5, to conclude in Section 6 that the expected inverse H-theorem is no longer true and that to suppose the test independency cannot be a limiting case for the dependency.

Finally, in Section 7 we give an algorithm and we show how, on a biological application, and with the variation of the concatenation law, we can extract some production rules which could be used in a next step, to construct a new object representation.

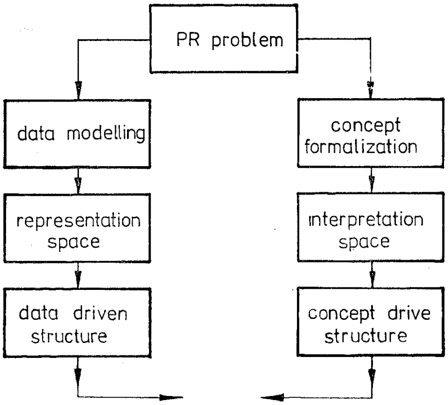

Fig. 1. Structure of a pattern recognition problem.

2. TRADITIONAL SCHEME

Taking into account the structural analysis of the pattern recognition algorithms (Simon, Backer, Sallantin [6]), we know that the principle of a lot of methods of pattern recognition is to establish a link between the representation space of the studied objects and the interpretation space of the concepts or observations characterizing the qualities of this objects [7].

This link is established with the help of a family of similarity measures $f(\mathbf{X}; A)$

explaining the similarity between the object which has **X** as a representation, and the concept A. Thus the structure of a pattern recognition problem can be expressed as in Fig. 1. Then there are two ways of establishing the object-concept similarities:

- the deductive similarities ("concept driven" similarities) which show what an imposed structure on the concept implies on the object-concept link.
- the inductive similarities which show what a representation induces on the concepts ("data driven" similarities).

Remarks

- For Watanabe [10], the notions of deduction and induction are expressed as follows:
  let H be an hypothesis,
  let D be a particular fact (experimental data)
  let A be an auxiliary fact

Induction means to induce H (resp. A, $H \wedge A$) from $A \wedge D$ (resp. $H \wedge D$, resp. D). Deduction means to deduce D from $H \wedge A$.

In this paper, we call A a concept, D a representation, H a structural hypothesis on the concepts. The deductive similarities correspond to the deduction notion expressed by Watanabe: the influence of the concept structure on the representations.

The inductive similarities are of another kind: they consist in inducing concept A from knowledge of the structure hypothesis H and of representation D.

Thus presented, the pattern recognition procedures are characterized not only by the *structure hypothesis* but also by a *sequence of inductions and deductions*.

- the best formalized case of inductive and deductive similarity measures is the one when the representations and the concepts are elements of a probability space of measure μ.

Thus they correspond to the conditional probabilities $\mu(\mathbf{X} \mid A)$ (deduction) and $\mu(A \mid \mathbf{X})$ (induction). But this is not the only frame in which we can conceive induction and deduction.

3. INDUCTIVE PROCESS

Let us suppose the objects representation is a real sequence x_i, $i = 1, \ldots, n$. We call $\mathbf{X}_n$ this representation. A new representation is obtained by adding a new variable x_{n+1} to $\mathbf{X}_n$. We call $\mathbf{X}_{n+1}$ this new representation.

We denote by $f(\mathbf{X}; A)$ the object-concept similarity measure; more precisely: $f(A \mid \mathbf{X})$ is the inductive similarity measure and $f(\mathbf{X} \mid A)$ the deductive one.

Let us consider the similarity measures $f(\mathbf{X}_n; A)$ and $f(\mathbf{X}_{n+1}; A)$.

The inductive process establishes a link between the similarity measures $f(\mathbf{X}_n; A)$ and $f(\mathbf{X}_{n+1}; A)$ taking into account the information given by the new representation.

Of course such inductive processes are made to select, among all the available tests which complete a representation the ones that will increase the discriminating power of the inductive similarity measures $f(A_i \mid \mathbf{X}_{n+1})$ for a set of concepts A_i, $i = 1, \ldots, m$.

When the similarities are valued in $[0, 1]$, the criterion to determine the new test is to maximize the transmitted information. That is what we are going to use in what follows.

We define an inductive automaton as an inductive process which decides on the choice of the new representation to increase the discriminating power of the inductive similarity measures.

Remarks

- We request, from an inductive automaton, to decide early, at least to estimate the decision risk.
- the truth of the decision can be estimated by confrontation with the experiment; this is a problem of another kind.
- the inductive automata may depend or not on the order of the tests which allow the construction of the representation. In the works of Terrenoire [9] and in Watanabe [11], and more generally, the order of the tests does not interfere.

Considering that the inductive and deductive similarity measures act like conditional probabilities we define an operating method. If they are supposed to be conditional probabilities then we can develop a mathematical formalism which is much more than an operating method. In this work we distinguish the operative hypothesis from the justification hypothesis; this allows us to show the singular behaviour of the automata which suppose the independence of the successive tests.

4. BAYESIAN INDUCTIVE PROCESS

Let us define the formalism:

Ω is the set of objects

$\mathscr{L}$ is the set of concepts

$\mathbb{Y}_1, \ldots, \mathbb{Y}_n$ are the elementary representation spaces obtained respectively as the set of issues of the objects to the first test, ..., to the nth test (these tests are also denoted by $\mathbb{Y}_1, \ldots, \mathbb{Y}_n$).

We can construct 2^n representation spaces corresponding to all the possible sequences of tests taken among $\mathbb{Y}_1, \ldots, \mathbb{Y}_n$ (if we do not consider their order, and if we exclude the repetition of the same test).

We denote by $\mathbb{X}_i$ $(i = 1, \ldots, n)$ the set of issues of the objects of Ω, to a sequence of i tests we suppose to be $\mathbb{Y}_1, \ldots, \mathbb{Y}_i$, without loss of generality.

Let us suppose we have defined a deductive similarity measure between the elementary spaces $\mathbb{Y}_1, \ldots, \mathbb{Y}_n$ and the set of concepts $\mathscr{L}$:

$$f : \mathbb{Y}_i \times \mathscr{L} \to [0, 1]$$

$(\mathbf{Y}_i, A) \to f(\mathbf{Y}_i \mid A)$ where $\mathbf{Y}_i$ is one of the possible issue to the test $\mathbb{Y}_i$.
Let us suppose we have the normalization:

$$\sum_{\mathbf{Y}_i} f(\mathbf{Y}_i \mid A) = 1, \quad \text{where } \mathbf{Y}_i \in \mathbb{Y}_i.$$

From these similarity measures between $\mathbb{Y}_i$ and $\mathscr{L}$, we would like to construct a similarity measure be.tween $\mathbb{X}_n$ and $\mathscr{L}$; for this purpose we shall use an associative composition law also called a semigroup from $[0, 1] \times [0, 1]$ in $[0, 1]$,
with $0 * x = 0 = x * 0$
$1 * x = x = x * 1$

Thus we define $f(\mathbf{X}_i \mid A)$, where $\mathbf{X}_i$ is a sequence of issues $\mathbf{Y}_1, \ldots, \mathbf{Y}_i$ to the tests $\mathbb{Y}_1, \ldots, \mathbb{Y}_i$ i.e. $\mathbf{X}_i \in \mathbb{X}_i$

$$f(\mathbf{X}_i \mid A) = f(\mathbf{Y}_1 \mid A) * f(\mathbf{Y}_2 \mid A) * \ldots * f(\mathbf{Y}_i \mid A) \tag{1}$$

Remark. A fundamental result from Ling [5] and Kampe de Feriet [3] determines all the topological semigroups which take their values on a closed interval $[a, b]$ from $\bar{\mathbb{R}}$ and satisfy:

- $*$ is continuous
- $*$ is nondecreasing with respect to the right and to the left
- for every $x \in [a, b] : b * x = x$ ("contracting") semigroup or $a * x = x$ ("expanding" semigroup)

All these semigroups are necessarily commutative ones, and are characterized by their idempotent i.e. the set of $x \in [a, b]$ such as $x * x = x$.

A lot of authors use particular semigroups to determine deductive or inductive similarity measures (e.g. Kayser [4]).

4.1. Law of Concatenation

In this paper we shall consider the commutative semigroups defined on $[0, 1]$ by:

$$x * y = \frac{(x - \varepsilon)(y - \varepsilon)}{1 - 2\varepsilon} + \varepsilon \quad \text{if} \quad (x, y) \in [\varepsilon, 1 - \varepsilon]^2$$

$$x * y = \inf(x, y) \quad \text{if} \quad (x, y) \notin [\varepsilon, 1 - \varepsilon]^2$$

where $\varepsilon \in [0, \frac{1}{2}]$.

From Kampe de Feriet [3], we can show that there exists no nilpotent element in $]\varepsilon, 1 - \varepsilon[$, i.e. element x such that:

$$\exists n \in \mathbb{N} \quad x * x * \ldots * x = x^n = \varepsilon \quad \text{and} \quad x^{n-1} \neq \varepsilon$$

but

$$\forall x \in [\varepsilon, 1 - \varepsilon[, \quad \lim_{n \to \infty} x^n = \varepsilon$$

ε and $1 - \varepsilon$ play in $[\varepsilon, 1 - \varepsilon]$ respectively the part of 0 and 1 in $[0, 1]$.

It is easy to show that:

$$\forall (x, y) \in [0, 1]^2 , \quad x * y \geqq xy \quad \text{and} \quad \lim_{\varepsilon \to 0} x * y = xy$$

with the two following particular cases:

$$\text{if } \varepsilon = 0, \; x * y = xy$$
$$\text{if } \varepsilon = \tfrac{1}{2}, \; x * y = \inf(x, y).$$

Thus is ε varies from 0 to $\frac{1}{2}$, we uniformly pass from multiplicative law to infimum law.

This family of semigroups is sufficient for our study because:

- the existence of other idempotents that 0 and 1 is the only one to play a part.
- these semigroups uniformly approach the multiplicative law used in probability to express independence.

From the experiment (issues to the tests) it is possible to deduce a deductive similarity measure between the elements $\mathbf{X}_i$ of the representation spaces $\mathbb{X}_i$ and the concepts of $\mathscr{L}$.

Now using the Bayes inversion formula we shall define an inductive similarity measure.

4.2. Bayesian Model of Induction

In probability, one considers a probability space $(\Omega, \mathscr{A}, \mu)$, where Ω is the set of objects, $\mathscr{A}$ a σ-field and μ a probability measure; the concepts A_i of $\mathscr{L}$ are considered as a measurable countable partit'on of Ω, and the objects of Ω having the same description X belonging to the representation space $\mathscr{X}$ are considered as a measurable set of $\mathscr{A}$. The deductive similarity measure is then defined by:

$$X \in \mathscr{X} , \quad A_i \in \mathscr{L} , \quad f(X \mid A_i) = \mu(X \mid A_i) \quad \text{(conditional probability)} .$$

The inductive similarity measure is defined by:

$$f(A_i \mid X) = \mu(A_i \mid X) .$$

Bayes formula can be written as:

$$\mu(A_j \mid X) = \frac{\mu(X \mid A_j)\, \mu(A_j)}{\sum_i \mu(X \mid A_i)\, \mu(A_i)}$$

and thus we can write in a probability frame

(2)
$$f(A_j \mid X) = \frac{f(X \mid A_j)\, \mu(A_j)}{\sum_i f(X \mid A_i)\, \mu(A_i)}$$

In a general frame, when we use similarity measures which are not built as conditional probabilities, we shall keep this last formula where the $\mu(A_j)$'s are the a priori weights (>0) on the concepts A_j of $\mathscr{L}$ corresponding to their plausibility (likelihood, credibility, preference ...). The denominator of (2) is nothing but a normalization.

A main property of the Bayesian inversion is kept : for any weight on the concept, if we know by deduction that $f(X \mid A_j) = 0$, then $f(A_j \mid X) = 0$. This means the concept A_j is logically refutable.

Moreover:

$$\forall A_i,\ A_j \in \mathscr{L} \qquad \frac{f(A_i \mid X)}{f(A_j \mid X)} = \frac{f(X \mid A_i)\,\mu(A_i)}{f(X \mid A_j)\,\mu(A_j)}$$

4.3. Choice of the New Test

The new test $\mathbb{Y}$ is chosen among all the possible ones maximizing the transmitted information. This quantity is defined by:

$$\text{(3)} \qquad \mathrm{U}(f(A_j \mid \mathbf{X})) - \sum_{\mathbf{Y}} \frac{m(\mathbf{XY})}{m(\mathbf{X})}\,\mathrm{U}(f(A_j \mid \mathbf{XY}))$$

where

$\mathbf{X}$ is the object description before experimenting the test $\mathbb{Y}$.

$\mathbf{XY}$ is the new object description after experimenting the test $\mathbb{Y}$ and when the issue to the test $\mathbb{Y}$ is $\mathbf{Y}$.

$m(Z) = \sum_{A_i \in \mathscr{L}} f(Z \mid A_i)\,\mu(A_i)$

$\mathrm{U}(f(A_j \mid \mathbf{X}))$ (resp. $\mathrm{U}(f(A_j \mid \mathbf{XY}))$ is Shannon's entropy of the distribution:
$(f(A_1 \mid \mathbf{X}), \ldots, f(A_j \mid \mathbf{X}), \ldots)$ where $A_i \in \mathscr{L}$
(resp. $(f(A_1 \mid \mathbf{XY}), \ldots, f(A_j \mid \mathbf{XY}) \ldots)$

Shannon's entropy $\mathrm{U}(p_j)$ of any incomplete distribution $(p_1, \ldots, p_N)$ with $\sum_{j=1}^{N} p_j \leqq 1, 0 \leqq p_j \leqq 1$ is defined by:

$$\mathrm{U}(p_j) = -\frac{\sum_{j=1}^{N} p_j \log p_j}{\sum_{j=1}^{N} p_j}$$

Formulas (1), (2), (3) constitute an operating method to calculate the inductive similarity measures from the deductive ones.

5. REPETITION OF THE SAME TEST

Various authors (see Giasu [2] and Watanabe [10]), use the repetition of the same test to study the behaviour of the Bayesian inductive automata: this behaviour is mathematically expressed by the average inverse H-theorem which says that on an average the discriminating power of the inductive similarity measures increases from some stage of the process (i.e. when the number of tests is increasing) and converges at the infinity.

The deductive similarity measures $f(\mathbf{Y}_i \mid A_j)$ are known by learning, taking into account the structural hypothesis on the set of concepts $\mathscr{L}$ (see Fu [1]).

The use of a family of tests does not change the nature of the problem. The average inverse H-theorem is still true.

We shall show that, except for the unique case when $\varepsilon = 0$, which generalizes the independence notion of the conditional probabilities ($\mu(\mathbf{X} \wedge \mathbf{Y} \mid A) = \mu(\mathbf{X} \mid A) \cdot \mu(\mathbf{Y} \mid A)$, to the deductive similarity measures, the average inverse H-theorem is no longer true, although there is always the convergence at the infinity.

This result seems important because it shows that generally we must supervise the construction of the inductive similarity measures (choice of ε) by a learning on a set of objects known to verify some given concept (training set).

First let us suppose that we dispose of only one test $\mathbb{Y}$ with issues $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$, that we can repeat. The elements of $\mathbb{X}_p = \mathbb{Y} \times \ldots \times \mathbb{Y}$ (p times) will be sequences of issues to the test $\mathbb{Y}$ repeated p-times.

We are interested in a set of n concepts $\mathscr{L} = \{A_1, \ldots, A_n\}$ that we try to explain with the help of the elements $\mathbb{X}_p$ (p belonging to $\mathbb{N}$): thus we are looking for inductive similarity measures $f(A_j \mid \mathbf{X}_p)$ with $A_j \in \mathscr{L}$, $\mathbf{X}_p \in \mathbb{X}_p$.

Let us suppose that $f(\mathbf{Y}_i \mid A_j)$ is known for every issue $\mathbf{Y}_i$ to the test $\mathbb{Y}$, and every concept A_j of $\mathscr{L}$. Moreover we suppose the a priori weights on the concepts A_j are known.

For a sequence $\mathbf{X}_p \in \mathbb{X}_p$ of p issues $\mathbf{Y}_{i_1}, \ldots, \mathbf{Y}_{ip}$ to the test $\mathbb{Y}$ p-times repeated, the formulas (1) and (2) give:

$$f(A_j \mid \mathbf{X}_p) = \frac{[f(\mathbf{Y}_{i1} \mid A_j) * \ldots * f(\mathbf{Y}_{ip} \mid A_j] \, \mu(A_j)}{\sum\limits_{A_k \in \mathscr{L}} [f(\mathbf{Y}_{i1} \mid A_k) * \ldots * f(\mathbf{Y}_{ip} \mid A_k)] \, \mu(A_k)}$$

Because the composition law $*$ is associative and commutative, we can put together the issues $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$. Let $\alpha_1(p), \ldots, \alpha_m(p)$ be the frequences of the issues $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ in the sequence $\mathbf{X}_p$.

$$f(A_j \mid \mathbf{X}_p) = \frac{[f^{p\alpha_1(p)}(\mathbf{Y}_1 \mid A_j) * \ldots * f^{p\alpha_m(p)}(\mathbf{Y}_m \mid A_j)] \, \mu(A_j)}{\sum\limits_{A_k \in \mathscr{L}} [f^{p\alpha_1(p)}(\mathbf{Y}_1 \mid A_k) * \ldots * f^{p\alpha_m(p)}(\mathbf{Y}_m \mid A_k)] \, \mu(A_k)}$$

where

$$f^{p\alpha_i(p)}(\mathbf{Y}_i \mid A_k) = \underbrace{f(\mathbf{Y}_i \mid A_k) * \ldots * f(\mathbf{Y}_i \mid A_k)}_{p\alpha_i(p)\text{-times}}$$

Simplifying the writing, we get:

$$(4) \qquad f(A_j \mid \mathbf{X}_p) = \frac{[\mathop{\times}\limits_{i=1}^{m} f^{p\alpha_i(p)}(\mathbf{Y}_i \mid A_j)] \, \mu(A_j)}{\sum\limits_{A_k \in \mathscr{L}} [\mathop{\times}\limits_{i=1}^{m} f^{p\alpha_i(p)}(\mathbf{Y}_i \mid A_k)] \, \mu(A_k)}$$

This last expression is nothing but the operating method of the automaton in the case of the repetition of one test only: the way we compute inductive similarities.

This process is not far from the weighing processes of [2].

Remark. 1 is a neutral element for $*$: $\forall x \in [0, 1]$ $1 * x = x * 1 = x$. Therefore we can suppose that if $\alpha_i(p) = 0$ we have $f^{\alpha_i(p)p}(\mathbf{Y}_i \mid A_j) = 1$ for all $A_j \in \mathcal{S}$. This justifies the simplified writting (4).

The use of laws $*$ allows us to compute the expression (4) completely.

5.1. Computation of the expression $f^{p\alpha_i(p)}(\mathbf{Y}_i \mid A_j)$

Different cases can occur:

1st *case*: $\alpha_i(p) \neq 0$ and $f(\mathbf{Y}_i \mid A_j) \in]\varepsilon, 1 - \varepsilon[$
then

$$f^{p\alpha_i(p)}(\mathbf{Y}_i \mid A_j) = \frac{(f(\mathbf{Y}_i \mid A_j) - \varepsilon)^{p\alpha_i(p)}}{(1 - 2\varepsilon)^{p\alpha_i(p) - 1}} + \varepsilon$$

2nd *case*: $\alpha_i(p) \neq 0$ and $f(\mathbf{Y}_i \mid A_j) \notin]\varepsilon, 1 - \varepsilon[$
then

$$f^{p\alpha_i(p)}(\mathbf{Y}_i \mid A_j) = f(\mathbf{Y}_i \mid A_j)\,.$$

5.2. Computation of the numerator of (4)

The concept A_j being given we can distinguish several kinds of sequences of issues to the test $\mathbb{Y}$ repeated p-times.
1th *case*: There exists in the sequence an issue $\mathbf{Y}_i$ of relative frequency $\alpha_i(p) \neq 0$ such that $f(\mathbf{Y}_i \mid A_j) \leqq \varepsilon$.

Then:

$$\mathop{\ast}_{i=1}^{m} f^{p\alpha_i(p)}(\mathbf{Y}_i \mid A_j) = \inf_{\mathbf{Y}_i \in U_j} (f(\mathbf{Y}_i \mid A_j)) \tag{5}$$

where U_j is the set of issues $\mathbf{Y}_i$ of frequency $\alpha_i(p) = 0$ verifying $f(\mathbf{Y}_i \mid A_j) \leqq \varepsilon$

2nd *case*: All the issues $\mathbf{Y}_i$ of frequency $\alpha_i(p) \neq 0$ are such that $f(\mathbf{Y}_i \mid A_j) \geqq 1 - \varepsilon$.
Then

$$\mathop{\ast}_{i=1}^{m} f^{p\alpha_i(p)}(\mathbf{Y}_i \mid A_j) = \inf_{Y_i \in V_j} (f(\mathbf{Y}_i \mid A_j)) \tag{6}$$

where V_j is the set of issues $\mathbf{Y}_i$ of frequency $\alpha_i(p) \neq 0$ verifying $f(\mathbf{Y}_i \mid A_j) \geqq 1 - \varepsilon$.

3rd *case*: There exists issues $\mathbf{Y}_i$ of frequency $\alpha_i(p) \neq 0$ such that $f(\mathbf{Y}_i \mid A_j) \in]\varepsilon, 1 - \varepsilon[$ and all the issues $\mathbf{Y}_i$ of frequency $\alpha_i(p) \neq 0$ are such that $f(\mathbf{Y}_i \mid A_j) > \varepsilon$.
Then

$$\mathop{\ast}_{i=1}^{m} f^{p\alpha_i(p)}(\mathbf{Y}_i \mid A_j) = \frac{\left[\prod_{\mathbf{Y}_i \in W_j} (f(\mathbf{Y}_i \mid A_j) - \varepsilon)^{\alpha_i(p)}\right]^p}{(1 - 2\varepsilon)^{p \sum_{\omega_j} \alpha_i(p) - 1}} + \varepsilon \tag{7}$$

where W_j is the set of issues $\mathbf{Y}_i$ of frequency different from 0 such that $f(\mathbf{Y}_i \mid A_j) \in$ $\in]\varepsilon, 1 - \varepsilon[$ when we are in the conditions of this 3rd case; ω_j is the set of indices i of $\mathbf{Y}_i \in W_j$.

5.3. Mean value of (4) and asymptotic value

To find the average behaviour of (4), when p is given, we must look at all the sequences of issues and define a "typical sequence" of length p.

For this purpose, we suppose that $\mathbb{Y}$ is a random variable which takes the values $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ with respective probabilities $\gamma_1, \ldots, \gamma_m$.

We look at the νth issues ($\nu \leqq p$) of all the sequences of p issues; we can define the "ensemble-average" frequency for the issue $\mathbf{Y}_i$: it is by definition the probability of $\mathbf{Y}_i$ i.e. γ_i, independent of ν.

Then the "time-ensemble average" frequency for $\mathbf{Y}_i$ that is the average of the ensemble average, when ν varies from 1 to p is also γ_i.

Then the typical sequence of length p will be characterized by γ_i as frequency of $\mathbf{Y}_i$ and then by inductive similarity measures denoted by $\bar{f}(A_j \mid \mathbf{X}_p)$ such that:

$$\bar{f}(A_j \mid \mathbf{X}_p) = \frac{\left[\bigtimes_{i=1}^{m} f^{p\gamma_i}(\mathbf{Y}_i \mid A_j)\right] \mu(A_j)}{\sum\limits_{A_k \in \mathscr{L}} \left[\bigtimes_{i=1}^{m} f^{p\gamma_i}(\mathbf{Y}_i \mid A_k)\right] \mu(A_k)}$$

Let us consider now $\lim\limits_{p \to \infty} \bar{f}(A_j \mid \mathbf{X}_p) = \bar{f}(A_j \mid \mathbf{X}_\infty)$ which represents the asymptotic mean-value of the inductive similarity measures.

We distinguish several kinds of concepts:

1st *case*: The subset U of $\mathscr{L}$ corresponding to the preceding first case.

If $A_j \in U$ then:

$$(5') \qquad \lim_{p \to \infty} \bigtimes_{i=1}^{m} f^{p\gamma_i}(\mathbf{Y}_i \mid A_j) = \inf_{\mathbf{Y}_i \in U_j} (f(\mathbf{Y}_i \mid A_j))$$

As a particular case in U we have the concepts A_j such that there exists an issue $\mathbf{Y}_i$ of probability $\gamma_i \neq 0$ verifying $f(\mathbf{Y}_i \mid A_j) = 0$. This subset of U will be denoted by U^*.

2nd *case*. The subset V of $\mathscr{L}$, corresponding to the preceding second case.

If $A_j \in V$ then:

$$(6') \qquad \lim_{p \to \infty} \bigtimes_{i=1}^{m} f^{p\gamma_i}(\mathbf{Y}_i \mid A_j) = \inf_{\mathbf{Y}_i \in V_j} (f(\mathbf{Y}_i \mid A_j))$$

3rd *case*: The subset W of $\mathscr{L}$ corresponding to the preceding third case:

If $A_j \in W$ then:

$$(7') \qquad \lim_{p \to \infty} \bigtimes_{i=1}^{m} f^{p\gamma_i}(\mathbf{Y}_i \mid A_j) = \varepsilon$$

Returning to $\bar{f}(A_j \mid \mathbf{X}_\infty)$ when A_j belongs respectively to U, V, W, we distinguish the situations when $\varepsilon \neq 0$ and $\varepsilon = 0$:

5.4. Computation of $\bar{f}(A_j \mid \mathbf{X}_\infty)$

5.4.1. Case $\varepsilon \neq 0$

(i) If $A_j \in U^*$ then for every $p\, \bar{f}(A_j \mid \mathbf{X}_p) = 0$ and then $\bar{f}(A_j \mid \mathbf{X}_\infty) = 0$. The concept A_j is said logically refutable.

(ii) If $A_j \in U - U^*$ (resp. V) then

$$\bar{f}(A_j \mid \mathbf{X}_\infty) = \frac{\inf\limits_{\mathbf{Y}_i \in U_j (\mathrm{resp.}\, V_j)} (f(\mathbf{Y}_i \mid A_j))\, \mu(A_j)}{D} \neq 0$$

where

$$D = \sum_{A_k \in U} \inf_{U_k} (f(\mathbf{Y}_i \mid A_k))\, \mu(A_k) + \sum_{A_k \in V} \inf_{V_k} (f(\mathbf{Y}_i \mid A_k))\, \mu(A_k) + \varepsilon \sum_{A_k \in W} \mu(A_k)\,;$$

D is different from 0 because ε is different from 0 except if $\mathscr{L} = U^*$

(iii) If $A_j \in W$ then

$$\bar{f}(A_j) \mid \mathbf{X}_\infty) = \frac{\varepsilon\, \mu(A_j)}{D}$$

5.4.2. Case $\varepsilon = 0$

When $\varepsilon = 0$, which corresponds to the case when $*$ is the multiplicative law, the behaviour of $\bar{f}(A_j \mid \mathbf{X}_\infty)$ is different because $U = U^*$, $V = \emptyset$ and $D = 0$.

(i) If $A_j \in U$, then $\bar{f}(A_j \mid \mathbf{X}_p) = \bar{f}(A_j \mid \mathbf{X}_\infty) = 0$ and A_j is logically refutable.

(ii) If A_j belongs to the subset W_1 of W such that:

$$W_1 = \{A_j \in W / \prod_{i=1}^{m} f^{\gamma_i}(\mathbf{Y}_i \mid A_j) = \max_{W} \prod_{i=1}^{m} f^{\gamma_i}(\mathbf{Y}_i \mid A_j))\}$$

then

$$\bar{f}(A_j) \mid \mathbf{X}_\infty) = \frac{\mu(A_j)}{\sum\limits_{A_k \in W_1} \mu(A_k)}$$

(iii) If $A_j \in W - W_1$ then $\bar{f}(A_j \mid \mathbf{X}_\infty) = 0$

Remark. Here we have only supposed that the test $\mathbb{Y}$ was a random variable, but generally the similarity measures are not conditional probabilities except for the case when $\varepsilon = 0$, where they can be so.

In this last case $\varepsilon = 0$ expresses the independence of any sequence of iterated tests $\mathbb{Y}$.

6. CONCLUSION

6.1. We see that the Bayesian model of induction when $\varepsilon = 0$ is the only one to take into account the information provided by the successive issues to the tests, when we decide after an infinite sequence of tests. This decision is taken when we choose among the concepts $\mathscr{L}$ one of the concepts A_j verifying:

$$\bar{f}(A_j \mid \mathbf{X}_\infty) = \max_{A_k \in \mathscr{L}} \bar{f}(A_k \mid \mathbf{X}_\infty)$$

In other cases where ε is different from zero, the experiment does not provide any information. This means that the infinite iteration of tests does not allow us to "learn" by elimination of cases. There are no more concepts refuted by the process.

6.2. In the Bayesian inductive processes supposing the independence of the conditional probabilities used as similarity measures (i.e. $\varepsilon = 0$), it has been noticed that the process forgets at first and then learns to reach a total knowledge: it is the average inverse H-theorem (Watanabe [10] and Yousri [13]).

In the Bayesian inductive processes built on the probabilistic independence, we can say from the last result that any decision taken after the pth stage would be better (if p is big enough).

But in the case when $\varepsilon \neq 0$, and when we use any deductive similarity measures, the average H-inverse theorem is no longer true; nevertheless we still have the convergence of Shannon's entropy $\overline{\mathbf{U}}(p)$ when $p \to \infty$ defined as:

$$\overline{\mathbf{U}}(p) = - \frac{\sum_{A_j \in \mathscr{L}} \bar{f}(A_j \mid \mathbf{X}_{\mathrm{n}}) \log \bar{f}(A_j \mid \mathbf{X}_p)}{\sum_{A_j \in \mathscr{L}} \bar{f}(A_j \mid \mathbf{X}_p)}$$

but the decrease is no more assured (see Fig. 2).

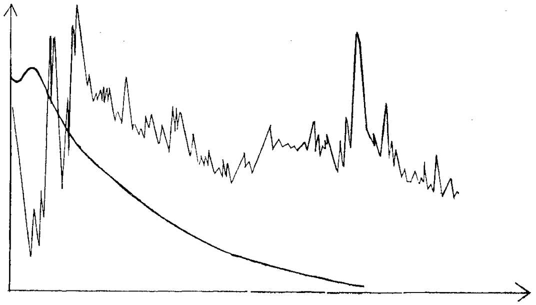

Fig. 2. ——— graph of $\overline{U}(p)$ for $\varepsilon = 0$
——— graph of $\overline{U}(p)$ for $\varepsilon = 0{\cdot}3$

6.3. From the points 1 and 2, we can say that the methods when $\varepsilon = 0$, are singular. Thus the hypothesis $\varepsilon = 0$ must be justified, for instance when there is the probabilistic independence.

Otherwise, the experiment justifies the operating method i.e. the choice of ε and the number of iterations: that is why the strategies are supervised by a learning that is to say by a set of objects known to verify a given concept (training set).

6.4. In this paper, we use the entropy in two different ways:

- entropy as an operating tool to construct an inductive process;
- entropy as a mathematical tool to evaluate a mathematical hypothesis, here the hypothesis of independence between tests.

7. APPLICATION

The inductive Bayesian processes can be used to extract the "best" sequences of tests to discriminate between concepts, when we have no a priori knowledge on the relationship between tests and when we have not enough objects and all the possible sequences of issues to more than one test.

We first give an algorithm using the principles developed in the preceding sections. This algorithm extracts some production rules which could be used to determine the relationship between a new object and a concept. Then we use it on an example.

7.1. Algorithm

Step 1: Compute on a training set the deductive similarities for any concept A_j and any issue $\mathbf{Y}_i$ to any test $\mathbb{Y}$.

Step 2: Give a value to the number N of tests.

Step 3: Choose a training set T for induction.

Step 4: Give a value to ε.

Step 5: For every element x of the training set do Steps 6 and 7.

Step 6: Select the N tests using the formulas (1), (2), (3) which express the inductive process. Compute the inductive similarities and their divergence.

Step 7: Affect x to one of the concepts, using the computed inductive similarities.

Step 8: Compute the divergence between the concepts determined in Step 7 and the a priori known concepts.

Step 9: Choose the ε's which give the lowest divergence.

Step 10: Determine for the above values of ε, what the issues of the used tests were, and what their links to the concepts were (extraction of production rules).

Example

We explain on an example how to extract short sequences of tests.

In a set of hydrocarbure molecules built with benzenic cycle we want to study how the geometrical properties of their frame support the discrimination between the cancerogenous ones and the other ones.

More specifically, we want to express some production rules to discriminate the two classes, dealing with only 3 tests ($N = 3$).

7.2.1. Deductive similarities (Step 1)

Table 1 (resp. Table 2) gives the deductive similarities between the issue 0 or 1 of each of the nine tests $T_1, \ldots, T_9$, and the concept "to be cancerogeneous" denoted by A (resp. "to be non cancerogeneous" denoted by $\bar{A}$).

Table 1. Deductive similarities for A.

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
Issue 0	0·83	0·33	0·17	0·58	0·67	0·58	0·67	0·33	0·58
Issue 1	0·17	0·57	0·83	0·42	0·33	0·42	0·33	0·67	0·42

Table 2. Deductive similarities for $\bar{A}$.

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
Issue 0	0·35	0·75	0·40	0·85	0·35	0·95	0·55	0·85	0·85
Issue 1	0·65	0·25	0·60	0·15	0·65	0·05	0·45	0·15	0·15

7.2.2. Inductive similarities (Steps 5, 6, 7)

For every ε (reported on the first line of Table 3) we have found the best sequence of three tests for any object of the training set: this sequence is characterized by its discriminating power between A a $\bar{A}$, computed from the inductive similarities and valued from -5 (in this case the object is more likely to belong to $\bar{A}$) to 5 (in this case the object is more likely to belong to A): this is reported in Table 3 for each of the 32 objects of the training set.

On the training set, we select the issues which are used for the selected test.

The expression of the inductive similarity for these specific answers correspond to the so called production rules: *evaluation on inductive inference.*

7.2.3. Productions rules (Step 11)

The production rules are the following:

For $\varepsilon = 0{\cdot}1$:

$$
\begin{aligned}
f(A|(T_8 = 0) \wedge (T_6 = 0) \wedge (T_2 = 0)) &= 0{\cdot}8\\
f(A|(T_8 = 1) \wedge (T_6 = 1) \wedge (T_3 = 1)) &= 0{\cdot}2\\
f(A|(T_8 = 1) \wedge (T_6 = 0) \wedge (T_1 = 1)) &= 0{\cdot}5\\
f(A|(T_8 = 1) \wedge (T_6 = 0) \wedge (T_1 = 0)) &= 0{\cdot}2
\end{aligned}
$$

Table 3. Discriminating power for every object when ε varies.

0	0·001	0·01	0·02	0·05	0·1	0·2	0·3	0·4	0·5
−4	−4	−3	1	1	3	1	0	0	0
−4	−3	−3	−4	−4	−3	−4	−4	−4	−4
4	4	4	4	4	3	3	3	3	3
1	1	1	1	1	3	1	0	0	0
1	1	−2	−2	−1	0	−2	−3	−3	−3
−4	−3	−3	−4	−4	−3	−4	−4	−4	−4
−5	−3	−3	−4	−4	−3	−4	−4	−4	−4
−4	−3	−3	−4	−4	−3	−4	−4	−4	−4
−4	−3	−3	−4	−4	−3	−4	−4	−4	−4
−5	−3	−3	−4	−4	−3	−2	−3	−3	−3
1	1	1	1	1	3	1	0	0	0
−5	−3	−3	−4	−4	−3	−2	−3	0	0
0	0	0	4	4	3	3	3	3	3
−4	−4	−3	1	1	0	1	0	0	0
−4	−4	−3	1	1	3	1	0	0	0
0	0	0	4	4	3	3	3	3	3
4	4	4	4	4	3	3	3	3	3
4	4	4	4	4	3	3	3	3	3
0	0	0	4	4	3	3	3	3	3
4	4	4	4	4	3	3	3	3	3
−5	−3	−3	−4	−4	−3	−2	−3	−3	−3
−4	−4	−3	1	1	3	1	0	0	0
−4	−4	−3	1	1	3	1	0	0	0
0	0	0	4	4	3	3	3	3	3
4	4	4	4	4	3	3	3	3	3
0	0	0	4	4	3	3	3	3	3
4	4	4	4	4	3	3	3	3	3
−4	−4	−2	−2	−1	0	−2	−3	−3	−3
−5	−3	−3	−4	−4	−3	−4	−4	−4	−4
0	0	0	4	4	3	3	3	3	3
1	1	1	1	1	0	1	0	0	0
4	4	4	4	4	3	3	3	3	3

For $\varepsilon = 0{\cdot}02$:

$$\begin{aligned}
f(A|(T_8 = 0) \wedge (T_6 = 0) \wedge (T_1 = 1)) &= 0{\cdot}9\\
f(A|(T_8 = 0) \wedge (T_6 = 0) \wedge (T_1 = 0)) &= 0{\cdot}6\\
f(A|(T_8 = 1) \wedge (T_1 = 0) \wedge (T_3 = 1)) &= 0{\cdot}1\\
f(A|(T_8 = 1) \wedge (T_1 = 1) \wedge (T_2 = 1)) &= 0{\cdot}3\\
f(A|(T_8 = 1) \wedge (T_1 = 0) \wedge (T_3 = 0)) &= 0{\cdot}1
\end{aligned}$$

These production rules could give a new description of the objects of the training set, which would take into account the relationship between objects and concepts.

In the same way, we know that, in statistical pattern recognition, the a posteriori conditional probabilities give an optimal description in the sense of the Bayes error.

(Received April 29, 1982.)

REFERENCES

[1] K. S. Fu: Sequential Methods in Pattern Recognition and Machine Learning. Academic Press, New York 1968.

[2] S. Guiasu: Information Theory with Applications. Mc Graw Hill, New York 1977.

[3] J. Kampe de Feriet, B. Forte and P. Benvenuti: Forme générale de l'opération de composition continue d'une information. C. R. Acad. Sci. Paris, Sér. A—B, *269* (1969), 529—534.

[4] D. Kayser: Vers une modélisation du raisonnement approximatif. Congrés de Reconnaissance des Formes, Saint-Maximin 1979.

[5] C. H. Ling: Public. Mathematicae *12* (1965), 189.

[6] J. C. Simon, E. Backer and J. Sallantin: A structural approach of pattern recognition. Signal Processing *2* (1980).

[7] J. Sallantin: Représentation d'observations dans le contexte de la Théorie de l'Information. Thèse d'Etat. Publication Structures de l'Information, CNRS n° 8, Paris 1979.

[8] J. Sallantin and Th. Van der Pyl: Analyse des processus inductifs bayésiens. C. R. Acad. Sci. Paris Sér. A *290* (1979), 389—392: see also: 5th IJCPR, Miami Beach 1980.

[9] M. Terrenoire, D. Tounissoux and B. A. Coche: Méthodes séquentielles pour l'aide au diagnostic. Journées Internationales "Récents dévelopements en reconnaissance des formes", Lyon, Mai 1979.

[10] S. Watanabe: Knowing and Guessing. John Wiley, New York 1969.

[11] S. Watanabe: Le principe de la moindre entropie. Séminaire IRIA classif. aut. et percept. par ordin.

[12] S. Watanabe, F. Takanori and Kamakura: A model of accelerated inductive learning. Informational Conference on Cybernetics and Society, Tokyo 1979, p. 1492.

[13] M. Yousri El-Fattah and C. Foulard: Learning Systems: Decision, Simulation and Control. (Lectures Notes in Control and Information Sciences 9.) Springer-Verlag, Berlin—Heidelberg—New York 1978.

Dr. Jean Sallantin, Dr. Thierry Van der Pyl, Groupe de Recherche C. F. Picard du C.N.R.S.-Structures de l'Information, Tour 45 — Université Paris VI, 4, place Jussieu, 75230 Paris Cédex 05. France.