A CLASSIFICATION PROCEDURE IN MEDICAL DIAGNOSIS

By

K. Ádám and M. Siminszky*

Department of Mathematics of the Faculty of Electrical Engineering,
Technical University, Budapest

Received June 20, 1978

Presented by Prof. Dr. I. Fenyő

The objective was set to develop a procedure for determining carcinoma exposure among leukoplakia patients.

Our data were obtained on a homogeneous sample of 500 leukoplakia patients treated during the past twenty years at the Dental Surgery Clinic of SOTE. Each patient had a syndrom-symptom vector of about 60 dimensions consisting of binary coded informations about the patient's condition and the results of laboratory examinations.

The patients were classified, on the basis of their syndroms, into categories of patients with more or less identical degree of carcinoma exposure. Naturally, we presume that the symptom vectors of similarly exposed patients are nearly "similar"; a concept that may be made precise by introducing a metric into the space of the possible symptom vectors.

Thereby the stage of exposure of a newly diagnosed patient will be stated on the basis of his category. Some possible procedures of classification will be described.

The procedure to be described seems to be the best possible theoretical answer to the question. We intend to treat the problem in a more general set-up, the above classification procedure of leukoplakia will be obtained as a particular case.

Let us denote a symptom vector by $\zeta = (\zeta_1, \dots \zeta_n)$ and the range of possible values of ζ by S, $S \subset R^n$. (We do not presume that the symptoms are necessarily binary coded). Let W indicate the parameter space $W = \{w_1, \dots, w_k\}$ and assume that to every value $w_i \in W$ there corresponds a distribution P_i on S, which has a density function f_i. In our application $w_1, w_2, \dots, w_k$ stand for the k classes with different degree of exposure; thus P_1 might be the distribution of the symptom vectors of patients with cancer, P_2 is that of patients with a high degree of exposure, etc., P_k is that of patients with no danger of cancer.

* Head pulpician, Weil Emil Hospital, Budapest

Let ω be a random variable, that takes its values in W according to the probability distribution $P(\omega = w_i) = p_i$, $i = 1, 2, \ldots k$. The proportion of the different classes in among all patients with leukoplakia is known (or can be determined from known data), that is to say, we know the *a priori* distribution $P = \{p_1, \ldots, p_k\}$. Suppose that we also know the conditional density functions $f(x \mid \omega = w_i) = f_i(x)$, i.e. the distributions of the symptom vectors of patients in the different classes. The possible decisions concerning the patients are represented by $d_1, d_2, \ldots, d_m$. $D = \{d_1, \ldots, d_m\}$ is the space of decisions. Especially, if $m = k$ and if d_i means that the patient is classified into the group i, we get the answer to our original problem. But d_i may also signify other decisions, e.g. decisions related to therapy.

Let $L : WxD \rightarrow R$ be the loss function, i.e. $L(w_i, d_j) = l_{ij}$ means the loss if the parameter value is w_i and our decision is d_j. $\varrho(P, d_j) = \sum_{i=1}^{k} l_{ij} p_i$ is our expected loss in case we decide for d_j. $\varrho(P,.)$ is called the risk function.

(If $k = m$, d_j means classifying into the class j, and

$$L(w_i, d_j) = l_{ij} = 1 - \delta_{ij}, \quad \text{i.e.} \quad l_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases}$$

then $\varrho(P, d_j) = \sum_{i \neq j} p_i$

is the probability of an erroneous classification.)

The decision $d^* \in D$ for which $\varrho^*(P) = \min_{j=1,\ldots,m} \varrho(P, d_j) = \varrho(P, d^*)$ (this is where the minimum of the risk function is attained), is called the Bayes decision corresponding to the distribution P. Let $\delta : S \rightarrow D$ be any function (so-called decision function). $\delta(x) = d_i \in D$ means that we decided for d_i if the observed value of the random variable was x.

Let Δ be the space of all decision functions. For each $\delta \in \Delta$ the value of the risk (the expected loss) can be determined as $\varrho(P, \delta)$.

$$\varrho(P, \delta) = \sum_{j=1}^{m} \int_S L(w_j, \delta(x)) f_j(x) p_j \, dx \tag{1}$$

(assuming a given *a priori* distribution P).

Our aim is to minimize the loss $\varrho(P, \delta)$, that is to select the decision function $\delta^* \in \Delta$ for which $\varrho(P, \delta^*) \leqslant \varrho(P, \delta)$ for every $\delta \in \Delta$.

Let $\varrho^*(P) = \inf_{\delta \in \Delta} \varrho(P, \delta)$. $\varrho^*(P)$ is called the Bayes risk. If there exists a $\delta^* \in \Delta$, for which $\varrho(P, \delta^*) = \varrho^*(P)$ holds, it will be called the Bayes decision function.

Our aim is to construct a Bayes decision function δ^* in case of a given *a priori* distribution P.

Under our assumptions the order of the summation and the integration in (1) can be changed:

$$\varrho(P, \delta) = \int\limits_S \left(\sum_{j=1}^{m} L(w_j, \delta(x)) f_j(x) p_j \right) dx .$$

Now, let us minimize the above sum Σ for each given $x \in S$. Let $f(x) = \sum_{j=1}^{m} f_j(x) p_j$ be the marginal distribution of ζ. For those $x \in S$ where $f(x) = 0$ there is no need to define $\delta(x)$, that is it may be defined in an arbitrary fashion. Thus, instead of minimizing Σ, let us minimize the expression

$$\sum_{j=1}^{m} L(w_j, d) \frac{f_j(x) p_j}{f(x)} = \sum_{j=1}^{m} L(w_j, d) P(\omega = w_j \mid \zeta = x)$$

for each given x (for which it makes sense at all). This is nothing but the expected value of the loss function under the condition $\zeta = x$. The distribution of ω under the condition is called *a posteriori* distribution.

So $\delta^*(x)$ is the Bayes decision corresponding to the distribution $P_x = \{P(\omega = w_j \mid \zeta = x)\}$. The determination of $\delta^*(x)$ means in this case that we specify some sets E_i, $i = 1, 2, \ldots, m$ in such a manner that $\bigcup_{i=1}^{m} E_i = R^n$, $E_i \cap E_j = \emptyset$ and $\delta(x) = d_i$ holds for every $x \in E_i$.

Especially, if $m = k$ and the loss function $L(w, d)$ is chosen as $l_{ij} = 1 - \delta_{ji}$ then the risk corresponding to the Bayes function gives the probability of a false classification. Later, when we have a larger population sample and more experience, we probably shall be able to construct a better loss function than $l_{ij} = 1 - \delta_{ij}$ the one given above. The need is there, since we might make a larger error by misclassifying a patient of class 1 into class 3 than in the opposite case or by taking him into some other class.

In principle, this is a perfect answer to our problem. If we have to find the class into which the patient should be classified on the basis of the value of the symptom vector ζ, the possibility of a false decision has to be minimized by calculating the Bayes decision function δ^*. (This has to be calculated only once, later the optimal decision arises by substituting the symptom vector of the new patient.) Still the problem is that the *a priori* distribution and the *a posteriori* density function are unknown. These and the conditional density functions $f_i(x)$ should be estimated from the sample. The relevant methods developed up to now require, however, relatively large samples. The sides of the samples required for the estimation grows exponentially with the dimension — so it is difficult to estimate the original 60 dimensional distributions on the basis of the available samples. The considerable lessening of the dimension would facilitate the application of any method and lessen the computer running time. The possible loss of information would not necessarily yield a less reliable solution, for it is obvious that a clastering of the points of a set, such that it would reflect reality, is possible only if the elements are dense enough in space.

Let us take a formarly diagnosed population of leukoplakia patients and divide them into two groups: one with carcinoma, the other group of non-carcinoma patients (such data are at our disposal and we want to see how well the symptom vectors characterize the disease, that is: how significantly the symptom vectors of the patients with or without carcinoma differ). This can be ascertained by a test. Our zero hypothesis is that the distribution of the symptom vectors in the two groups are identical, i.e. the entire sample can be regarded homogeneous. Our counter-hypothesis is the opposite. In case the test shows (on a given significance level) that we have to accept the hypothesis H_0, the entire statistical analysis, at least on the basis of the involved symptom vector, is of no use.

If the hypothesis H_0 is rejected, our next task is to decrease the dimension of the symptom vectors, to omit the unessential components. Therefore we have to replace the symptom vectors $\zeta \in R^n$ with new symptom vectors $\xi \in R^r (r < n)$. This procedure is called factor analysis.We have to apply a simple transformation $T : R^n \rightarrow R^r$ such that the procedure based on the observation of the new symptom vector $\xi = T(\zeta)$ contains a minimum of errors.The theoretically well developed methods of extracting the essential data cannot be used here, because of the character of the sample at our disposal. We can use, however, some more or less heuristical procedures:

1. Let us examine now the symptom vectors of the carcinoma patients alone, and find the distribution of the comparing these values to the distribution on the entire sample we can select those components, which are considerably less in the carcinoma sample, than in the full sample. Evidently these must be the most important factors. But the question is when to regard the variance of a component "small enough" and when nct, can of course be decided on the basis of experience and a series of tests.

2. After omitting some coordinates from the symptoms vectors divided into two groups we apply χ^2 tests at each step and see to what extent the distributions of the reduced vectors differ. This method permits to single out the number $r < n$ of those components, which cause the greatest differences, that is, which are the most characteristic of the carcinoma state of the patient. In this case we must also rely on our experience, because on the basis of the χ^2 test it cannot be judged exactly which one of the two patterns of different dimensional vectors is less homogeneous.

3. Let us examine every coordinate of the symptom vector one by one, from the point of view, wether there is a different diffusion characteristic for

carcinoma and non-carcinoma patients. In the case where two coordinates are found to have nearly identical distributions they can be omitted as they do not help the classification.

Regarding the χ^2 test mentioned in paragraphs 2 and 3, we add that means identical distribution in both cases, and means different distributions.

In 1, (or in 2, 3) two types of errors may occur: a) we reject H_0, though it is valid (error of type I), or we accept H_0, though it is not valid (error of type II).

The probability of the second type of error is most important to us, for it means the likelihood that a component of great importance has been left out from the symptom vector. Naturally also the probability of the first type of error has to be reduced, because it means dealing with superfluous components. The tables used for statistical analysis contain generally the connection between the error of the first type $p = P(\chi_k \mid H_0)$ and the critical zone. χ_k — the critical zone — is the set of those statistical values where H_0 is rejected. On the contrary the second type of error generally exhibits a complementary occurrence, the force function being $E = P(\chi_k \mid H_1)$. The average size of the force function depends on the given sample size, the error of the first type and also on the difference of the distributions. The average value of the test can only be estimated if the *a priori* distribution is known. However, we know, that the greater, the first type error, the closer the force function is to 1. The χ^2 test is known to be consistent. That is, for any first type of error $\varepsilon > o$ despite the counter-hypothesis, the value of the force function will approximate 1, when the elements of the sample are increasing. The two facts guarantee that in case of a bigger sample the second type of error can be reduced by employing the right first type of error. These three methods help to select the appropriate number $r < n$ of components, so that they contain all essential information.

The sample thus prepared may analysed by the described method of decision theory, but also cluster analysis will produce reliable results.

Summary

A method is described for the determination of carcinoma exposure among leukoplakia patients using the theory of Bayes decisions.

On the basis of the observed symptom vectors the patients are classified into groups characterizing their exposures so that the expected error — weighted by its occasional consequences should be minimal. For this purpose the optimal Bayes decision function has to be determined.

Besides, some procedures are contained to reduce the dimension of symptom vectors.

Katalin ÁDÁM
Mario SIMINSZKY } H-1521 Budapest