INTERACTIVE COMPUTER PROGRAM FOR SELECTING VARIABLES IN MEDICAL DIAGNOSIS, BASED ON INFORMATION THEORY.

Martine PUZIN — The Hospital for Sick Children - Toronto
Jan KRYSPIN — University of Toronto - Inst. of Bio-Medical Electronics.

ABSTRACT:

In evaluating the data obtained by physiological experiments (multivariable systems) and clinical observations in intensive care units, criteria have to be established based on both quantitative (instrumental) and qualitative (clinical observations) data to determine which are the most relevant variables among the bulk of results we get. We demonstrate an interactive computer program using an IBM 1800, in an experimental monitoring and a clinical diagnostic system. With the help of information theory we are able to compute the different degrees of relationships between arbitrary combinations of observables which lead to the elimination of redundant variables with the estimate of percentage of information lost by such a reduction. A relevant subset of observables is derived from this method, for each collection of experiments.

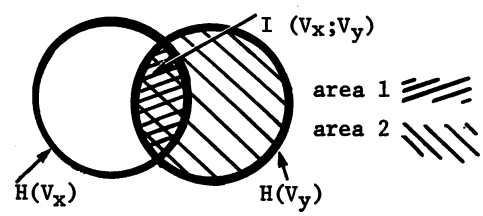
INTRODUCTION:

The applications of computers to help in medical diagnosis problems have been limited in the past either to 1) determine the probability of a diagnosis given specific symptoms and signs; or 2) classify a disease on the basis of similarity or dissimilarity of clusters of signs and data. Both these approaches follow more or less the Platonic approach to diagnosis, namely, that the universal entity-disease - is absolutely recognizable in every patient on the basis of his symptoms, signs and laboratory data. The last developments have too much stressed the importance of "objective" laboratory data in the final decision about the diagnosis. At the same time, the individual aspects of diseases, so important in the Hippocratic tradition of medicine have been relegated mainly to the field of the so-called medical "art" as though they were of less importance in the scientific approach to diagnosis and treatment. Only recently (e.g. FEINSTEIN 1969) the significance of the individual aspects of diseases was rediscovered and the need for new approaches in the taxonomy of illnesses (as opposed to diseases) was stressed. The illness of a particular patient is understood to be an intersection of the abstract entity called disease and of the physical entity of a host. The host is usually classified by its demographic data, personal attributes and environmental characteristics. The taxonomy of illnesses in computer compatible forms does not yet exist although it is very familiar to all experienced physicians. Its categories are e.g. 1) mode of detection, 2) cluster of symptoms and signs, 3) sequence of clinical manifestations, 4) timing of clinical manifestations, 5) comorbidity, 6) physiological characteristics, 7) typological characteristics or 8) higher nervous activity type of individual patients.

This approach to medical diagnosis calls for methods different from the usually applied Bayesian probability theorem or numerical taxonomy methods. Here we are more concerned with specific characteristics of clinical observation and experiment and with individual patients. The theoretical outlines of this approach have been studied by KRYSPIN (1968). The present paper deals with the possibilities and computer aspects of the theory of information in the study of individual illnesses. The model is based on data retrieved either in an experimental monitoring system or in a clinical ensemble of neurosurgical patients with brain tumors. The theory of information is suitable for clinical application because it allows the evaluation of relationships between symptoms, signs, laboratory data and disease.

INFORMATION THEORY AS THE BACKGROUND OF THE METHOD:

We give some definitions and summarize certain elements of Information Theory which form the background of the programs. The paper by MCGILL (1954) can be considered as the basis of the present approach. Let us consider a biological system defined by a set of N variables (V_i), i=1,...N. The values (measured or observed) taken by these variables at a certain time determine the "state" of the system. We can always choose two disjoint subsets of V such as VO and VI, which we call output subset and input subset. Transmitted or "shared" information measures the amount of association between the input and output subsets of the system. To illustrate the concept of shared information between two variables $I(V_x;V_y)$ let us use Venn diagrams. Let $H(V_x)$ marginal entropy of V_x and $H(V_y)$ marginal entropy of V_y be represented as the respective areas of two circles and $H(V_x,V_y)$ joint marginal entropy of (V_x,V_y) be represented by the total area enclosed by the two circles. See figure 1.



- figure 1.

If V_x and V_y are completely independent the two circles do not overlap and the common area or intersection of $H(V_x)$, $H(V_y)$ is zero. $I(V_x;V_y) = 0 = I(V_y;V_x)$ i.e., the amount of transmitted information between V_x and V_y is zero. On the other hand if one circle is completely contained by the other $I(V_x;V_y)$ is maximum and equal to the smallest circle. If the circles are equal, all the information contained by V_x is contained by V_y and we can discard one of the two variables as completely redundant, without any loss of information. By generalizing we define the N-dimensional transmitted information as being the joint entropy of the input subset, plus the joint entropy of the output subset, minus the joint entropy of the subset union of V_I and V_O. We need to use a multivariate model with multivariable input and multivariable output, because unlike communication theory where the transmission is restricted to a single source of information, the biological systems are of an extremely complex type of intricacies between all the sources or carriers of information. The Information Influence Coefficient Z gives the percentage of amount of information of the output subset which is due to the input subset, in other words, it shows the "influence" of the input onto the output, of course, the higher this coefficient the higher this influence, and the greater the dependency of the input subset and the output subset. In case of one variable as input and one as output this Z gives the degree of dependency of these two variables $Z(V_x;V_y) = I(V_x;V_y)/H(V_y)$. By looking again at the Venn diagram for this particular example (Figure 1) we see that $Z(V_x;V_y)$ is the ratio of area 1 to area 2, and obviously shows that Z is the amount of information of V_y due to the presence of V_x. It lies in the range $0 \leq Z \leq 1$, $Z=1$ when V_y is entirely dependent upon V_x. The information of V_y is completely redundant provided we know V_x. $Z=0$ when V_y and V_x are completely independent and $I(V_x;V_y) = 0$. Now that we have defined the three main Information functions I,H,Z, we shall describe the model(program) which makes use of them for solving our problem.

AN INTERACTIVE COMPUTER PROGRAM:

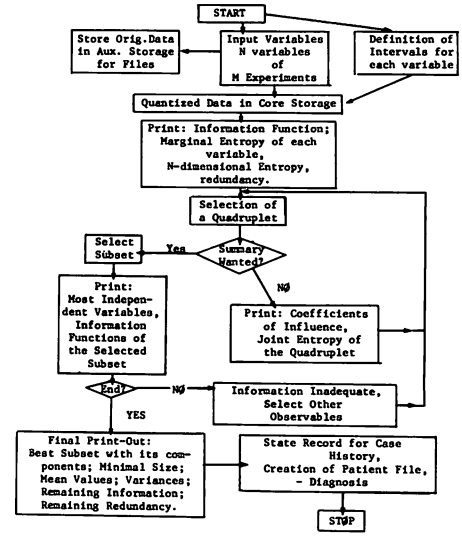
Let us consider a patient under observation as an information system composed of generators of information $V_1, V_2 \ldots V_n$. Each corresponds to one type of medical observation and their values specify the state of the system. They are time-dependent and a real-time interactive program is most suitable for this kind of system. The number n was chosen to be 40 as an acceptable compromise between sufficient number of variables and computer limitations. The 40 information sources are generally dependent on each other to a certain degree. We consider the union of the marginal entropies of the 40 variables as representing the total information generated by our system. We see that a simple study of the different marginal entropies is not satisfactory in itself because we would not know anything about the dependency of the variables. Calculation of Z functions is then necessary (including calculations of I) to evaluate the percentage of redundancy of the system.

Description of data:

Each experiment consists of the recording of up to 40 values, one for each variable. These 40 data are quantized according to the different intervals set by the user. What is available at the moment is data extracted from Hospital files recorded on DIFFERENT patients having a common disease. By monitoring data in real-time we shall have an "ideal data collection" technique.

Software - IMEDIS Package:

The programs have been written in FORTRAN IV for the first evaluation of the method. The tasks achieved by the different components of the IMEDIS package (Interactive Medical Diagnosis System) are presented here in the form of a flow-chart.



The final subset will contain the least redundant variables that are hopefully also the most relevant ones in the characterization of the patient's state. Further study will show to what extent this procedure may influence the decisions of a physician.

BIBLIOGRAPHY

-Feinstein, Alban R. "Taxonomy and Logic in Clinical Data". Ann. N.Y. Acad.Sci. vol.161, Art. 2, p. 450-459, 1969.

-Kryspin, J. "Dimensional Analysis of Electrophysiological Quantities and the Methods of Approach to the Biological Field Equations". U.of T.1968.

-McGill,W.J. "Multivariate Information Transmission". Transactions PGIT, 1954 Symposium on Information Theory, PGIT - 4, pp. 93 - 111.