X Symposio Internacional Métodos Matemáticos Aplicados a las Ciencias
Universidad de Costa Rica

Information & pronostic value of some variables on breast cancer

Jérome MANUCEAU, Marylène TROUPÉ, Jean VAILLANT

Université des Antilles-Guyane
UFR Sciences exactes et naturelles-Campus de Fouillole
Département de Mathématiques-Informatique
97159 Pointe-à-Pitre cedex

January 11, 1997

Abstract: In the context of survival analysis, mutual information theory between σ-algebras can be used for selecting qualitative or quantitative covariables with a high predictive value. The information rate carried out by covariables can be tested by means of a decomposition similar to the analysis of variance. Statistical units with some missing data are not excluded from the analysis so that a loss of information is avoided. These results are applied to survival data from 1304 patients with breast cancer followed over a period of ten years.

Mots-clés : Entropy, σ-algebra, Information, Survival analysis

1 Introduction

Let $(\Omega, \mathcal{F}, P)$ be a probability space. The entropy of the probability law P is an interesting tool for characterising uncertainty associated with P. When Ω is a countable set, the entropy of P is defined by

$$H(P) = -\sum_{\omega \in \Omega'} P(\omega) \ln(P(\omega)) \tag{1}$$

where $\Omega' = \{\omega \in \Omega \mid P(\omega) > 0\}$.

For a random variable X on a countable set Ω, the entropy of X denoted by $H(X)$ is the entropy of the image law P_X. If we now consider two random variables X and Y on Ω, conditionally to the event $Y = y$, the entropy of X denoted by $H(X \mid Y = y)$ is the entropy of P_X conditional on $Y = y$. The entropy of X conditional on Y is defined as follows :

$$H(X \mid Y) = \sum_{y \in Y(\Omega)} P(Y = y) H(X \mid Y = y) \tag{2}$$

This is the expectation of the random variable which maps $H(X \mid Y = y)$ to each y.

The notion of entropy seen above cannot be directly extended to non countable sets (see [8] for the formal link between *Shannon* entropy [9] and *Kullback* entropy [4] and their respective properties). Let μ be a measure on Ω, the generalized entropy with respect to this reference measure μ ([2],[8]) is defined as :

$$H(P; \mu) = -\int_{\Omega} f(\omega) Ln(f(\omega)) \mu(d\omega)$$

where f is the Radon-Nikodym derivative of P with respect to μ. If P is singular with respect to μ, we set $H(P; \mu) = \infty$. It is worth noticing that if we consider the generalized entropy of the approximating discrete distribution P_n with respect to the corresponding discrete approximation μ_n to the reference measure μ, then $H(P; \mu)$ is the limit of the sequence of discrete approximations $H(P_n; \mu_n)$.

In what follows, we refer to the same measured probability space $(\Omega, \mathcal{F}, \mu, P)$. We consider sub-$\sigma$-algebras of $\mathcal{F}$ generated by random variables. Let $\mathcal{A}$ be such a σ-algebra and X the random variable generating $\mathcal{A}$, we define the entropy of $\mathcal{A}$ denoted by $H(\mathcal{A})$ as follows : $H(\mathcal{A}) = H(P_X; \mu_X)$.

2 Information between σ-algebras

Let $\mathcal{A}$ and $\mathcal{B}$ be two σ-algebras. We denote by $\mathcal{A} \vee \mathcal{B} = \sigma(\mathcal{A} \cup \mathcal{B})$ the σ-algebra generated by $\mathcal{A} \cup \mathcal{B}$.

Definition 1 *Consider two σ-algebras $\mathcal{A}$ and $\mathcal{B}$. the entropy of $\mathcal{A}$ conditionally to $\mathcal{B}$ is defined as follows :*

$$H(\mathcal{A} \mid \mathcal{B}) = H(\mathcal{A} \vee \mathcal{B}) - H(\mathcal{B}). \tag{3}$$

This entropy represents the uncertainty on $\mathcal{A}$ when $\mathcal{B}$ is known.
We have the following properties :

Property 1

1. $\mathcal{A} \subset \mathcal{B} \Rightarrow H(\mathcal{A}) \leq H(\mathcal{B})$.
2. $H(\mathcal{A} \vee \mathcal{B}) \leq H(\mathcal{A}) + H(\mathcal{B})$ *and* $H(\mathcal{A} \mid \mathcal{B}) \leq H(\mathcal{A})$.
3. *If $\mathcal{A}$ and $\mathcal{B}$ have finite entropies, then :*
   $H(\mathcal{A} \vee \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B})$ *and* $H(\mathcal{A} \mid \mathcal{B}) = H(\mathcal{A})$ *if and only if $\mathcal{A}$ and $\mathcal{B}$ are independant. We then write $\mathcal{A} \perp \mathcal{B}$.*

For a correct quantification $I(\mathcal{A},\mathcal{B})$ of the information between σ-algebras $\mathcal{A}$ and $\mathcal{B}$, some postulates are required concerning necessary properties verified by $I(\mathcal{A},\mathcal{B})$:

1. $\mathcal{A}\perp\mathcal{B} \quad \Leftrightarrow I(\mathcal{A},\mathcal{B}) = 0,$
2. $I(\mathcal{A},\mathcal{B}) = I(\mathcal{B},\mathcal{A}),$
3. $\mathcal{A} \subset \mathcal{C} \Rightarrow I(\mathcal{A},\mathcal{B}) \leq I(\mathcal{C},\mathcal{B}), \quad \mathcal{C}$ being a σ-algebra.

Choosing $I(\mathcal{A},\mathcal{B})$ as follows, all these postulates are verified :

Definition 2

$$I(\mathcal{A},\mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B}) - H(\mathcal{A}\vee\mathcal{B}), \tag{4}$$

Moreover,

$$I(\mathcal{A},\mathcal{A}) = H(\mathcal{A}). \tag{5}$$

Following the definition of mutual information between two σ-algebras $\mathcal{A}$ and $\mathcal{B}$ given by (4), we define the mutual information between $\mathcal{A}$ and $\mathcal{B}$ conditional on σ-algebra $\mathcal{C}$ by :

Definition 3

$$I(\mathcal{A},\mathcal{B}\mid\mathcal{C}) = H(\mathcal{A}\mid\mathcal{C}) + H(\mathcal{B}\mid\mathcal{C}) - H(\mathcal{A}\vee\mathcal{B}\mid\mathcal{C}). \tag{6}$$

We will now show how the entropy of a σ-algebra $\mathcal{A}$ can be decomposed by means of the mutual information between $\mathcal{A}$ and n other σ-algebras, say $\mathcal{B}_1$, ..., $\mathcal{B}_n$.

Theorem 1 *Let $\mathcal{A}$, $\mathcal{B}_1$ and $\mathcal{B}_2$ be 3 σ-algebras, then*

$$I(\mathcal{A},\mathcal{B}_1\vee\mathcal{B}_2) = I(\mathcal{A},\mathcal{B}_1) + I(\mathcal{A},\mathcal{B}_2) + I(\mathcal{B}_1,\mathcal{B}_2\mid\mathcal{A}) - I(\mathcal{B}_1,\mathcal{B}_2) \tag{7}$$

Proof

Let us calculate $Z = I(\mathcal{A},\mathcal{B}_1\vee\mathcal{B}_2) - \Big(I(\mathcal{A},\mathcal{B}_1) + I(\mathcal{A},\mathcal{B}_2) + I(\mathcal{B}_1,\mathcal{B}_2\mid\mathcal{A}) - I(\mathcal{B}_1,\mathcal{B}_2)\Big)$.

By using expressions (3), (4) and (6), it is straightforward that $Z = 0$, which leads us to the final result. •

We see that the mutual information between $\mathcal{A}$ and $\mathcal{B}_1\vee\mathcal{B}_2$ is the sum of the mutual informations $I(\mathcal{A},\mathcal{B}_1$ and $I(\mathcal{A},\mathcal{B}_2$ if and only if the mutual information between $\mathcal{B}_1$ and $\mathcal{B}_2$ does not depend on $\mathcal{A}$.

Theorem 2 *Let $\mathcal{A}$, $\mathcal{B}_1$, ..., $\mathcal{B}_n$ be σ-algebras, the entropy of $\mathcal{A}$ can be decomposed as follows :*

$$H(\mathcal{A}) = \sum_{i=1}^{n} I(\mathcal{A},\mathcal{B}_i) + \sum_{i=2}^{n}\Big(I(\bigvee_{j=1}^{i-1}\mathcal{B}_j,\mathcal{B}_i\mid\mathcal{A}) - I(\bigvee_{j=1}^{i-1}\mathcal{B}_j,\mathcal{B}_i)\Big) + H(\mathcal{A}\mid\bigvee_{i=1}^{n}\mathcal{B}_i). \tag{8}$$

proof
We can prove this proposition by recurrence on n. From expressions (3) and (4), we derive

$$H(\mathcal{A}) = I(\mathcal{A}, \mathcal{B}_1) + H(\mathcal{A} \mid \mathcal{B}_1) \tag{9}$$

so that equality (8) is true for $n = 1$.

If we denote $\bigvee_{i=1}^{n} \mathcal{B}_i$ by $\mathcal{F}_n$ for any non null integer n, then we have

$$H(\mathcal{A}) = I(\mathcal{A}, \mathcal{F}_{n+1}) + H(\mathcal{A} \mid \mathcal{F}_{n+1}) \tag{10}$$

From (7), we obtain

$$I(\mathcal{A}, \mathcal{F}_{n+1}) = I(\mathcal{A}, \mathcal{F}_n) + I(\mathcal{A}, \mathcal{B}_{n+1}) + I(\mathcal{F}_n, \mathcal{B}_{n+1} \mid \mathcal{A}) - I(\mathcal{F}_n, \mathcal{B}_{n+1}).$$

These two last equations give us the decomposition of $H(\mathcal{A})$ with respect to $\mathcal{B}_1, ..., \mathcal{B}_{n+1}$. •

When we have no reasonable parametric models with respect to the observed phenomena (variable of interest Y and k covariables $X_1, \ldots, X_k$), a non parametric approach is possible from the information provided by contingency tables associated with the observations in a similar way as the one used by [5] et [6]. A maximum information principle with respect to a reference σ-algebra associated with the variable of interest Y is applied here.

3 Method of covariable selection

We use a method based on the estimation of the conditional and non conditional entropies of Y and covariables $X_1, \ldots, X_k$. The estimates are calculated by means of the empirical frequencies provided by the multidimensional contingency tables. For covariables having a continuous distribution, we then assume that the outcomes have been grouped into mutually exclusive intervals whose union contains the support of the random variable.

3.1 Preparing of contingency tables

To diminish low frequency occurrences in the contingency table, some covariable modalities can be grouped by means of a maximum entropy principle described below.

On the maximum entropy principle Let Ω be a set of m elements. The probability distribution on Ω with the highest entropy is the uniform distribution whose entropy is equal to $\ln(m)$.

The maximum entropy principle can be applied to a quantitative or ordered qualitative variable Z. It consists in reducing the initial number of modalities to a value k by grouping k classes of consecutive modalities such that their frequencies are almost equal.

For example, a variable Z with 10 modalities whose frequencies are given in table1 can have its modalities reduced to 3 as follows using the maximum entropy principle :

Initial modality	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}
Frequency	3	16	10	7	19	8	11	5	1	14
New modality	Z'_1			Z'_2			Z'_3			
New frequency	29			34			31			

Table 1: *Example of modality grouping with the maximum entropy principle*

3.2 Information carried by a covariable

Let Y be the variable of interest, X, any covariable, (n_{ij}) the contingency table associated with Y and X, and $p = (p_{ij})$ the matrix of expected relative frequencies.

The first step of the method of selection consists of quantifying the relationship between X and Y. Following this aim, we use a correlation coefficient based on the *Shannon* entropy defined as follows:

Definition 4 *The expression of the relative information of X on Y is*

$$I_r(X,Y) = \frac{I(X,Y)}{H(Y)} \tag{11}$$

Using expressions (3) and (4), we get :

$$I_r(X,Y) = \frac{H(Y) - H(Y \mid X)}{H(Y)} \in [0,1]. \tag{12}$$

$I_r(X,Y)$ can be seen as the percentage of uncertainty on Y conditionally to X.

The observation of X and Y through a contingency table leads to the following property which is a consequence of equalities (??) and (12):

Property 2 *An expression of the relative information of X on Y is:*

$$I_r(X,Y) = \frac{\sum_{i,j} p_{ij} \ln\left(\frac{p_{ij}}{p_{i.}p_{.j}}\right)}{\sum_{j} p_{.j} \ln(p_{.j})} \tag{13}$$

We then deduce the following property:
Property 3 *An estimator of the relative information of X on Y is:*

$$T = \widehat{I}_r(X,Y) = \frac{\sum_{i,j} \widehat{p_{ij}} \ln\left(\frac{\widehat{p_{ij}}}{\widehat{p_{i.}}\widehat{p_{.j}}}\right)}{\sum_{j} \widehat{p_{.j}} \ln(\widehat{p_{.j}})} \tag{14}$$

where $\widehat{p_{ij}}$, $\widehat{p_{i.}}$ and $\widehat{p_{.j}}$ are classical notations used for the empirical relative frequencies.
This estimator is convergent and follows asymptotically the gaussian distribution $\mathcal{N}\Big(I_r(X,Y), \mathrm{var}(T)\Big)$ where $\mathrm{var}(T)$ is the asymptotic variance of T. Its expression is

$$\mathrm{var}(T) = \frac{1}{n}\mathrm{var}\left(\sum_{i,j} \frac{\partial I_r(X,Y)}{\partial p_{ij}}(p)\mathbb{I}_{ij}\right) \tag{15}$$

where $n = \sum_{i,j} n_{ij}$ is the number of statistical units considered, and the $\mathbb{I}_{ij}$ are the indicator functions associated with the contingency table of (n_{ij}).
Using equalities (13) and (15), we get the following property:
Property 4 *An expression of the asymptotic variance of estimator T is:*

$$\mathrm{var}(T) = \frac{1}{n}\left[\frac{1}{H^2(Y)}\sum_{i,j} p_{ij}\left[\ln\left(\frac{p_{ij}}{p_{i.}\,p_{.j}}\right) + T\ln(p_{.j})\right]^2\right]. \tag{16}$$

$\mathrm{var}(T)$ can be estimated by $\widehat{var}(T)$ which is obtained by replacing p_{ij}, $p_{i.}$ and $p_{.j}$ by their empirical estimations $\widehat{p_{ij}}$, $\widehat{p_{i.}}$ and $\widehat{p_{.j}}$.

3.3 Information carried by several covariables

As in the previous paragraph, we can measure the combined relative information of several covariables $X_1, \ldots, X_k$ on Y. The valuation of the information variance permits to build confidence intervals and eliminates the less pertinent covariables.

Vocabulary: When Y is the survival variable, $I_r(X,Y)$ is called pronostic value of X and denoted by $I_r(X)$.

3.4 Practical method of selection

We select progressively the different covariables using an iterative process.
Let E be the set of covariables $X_1, \ldots, X_k$ considered in the study.

The first step consists of building the set E_1 defined by

$$E_1 = \left\{ X_l \in E \ / \ \widehat{I}_r(X_l, Y) + u_\alpha \sqrt{\widehat{\text{var}}\left(\widehat{I}_r(X_l, Y)\right)} \geq p_1 \right\}$$

and at a step h ($h \geq 2$), we get

$$E_h = \left\{ (X_l, x) \in E \times E_{h-1} \ / \ \widehat{I}_r\Big((X_l, x), Y\Big) + u_\alpha \sqrt{\widehat{\text{var}}\left(\widehat{I}_r\Big((X_l, x), Y\Big)\right)} \geq p_h \right\}$$

where p_1, p_h and α are thresholds fixed by the user, and u_α is such that $\Phi(u_\alpha) = 1 - \frac{\alpha}{2}$, Φ being the distribution function of the gaussian distribution $\mathcal{N}(0, 1)$.

Remarks

1. The thresholds series $(p_h)_{h \in I\!N^*}$ must be a strictly increasing sequence because the adding of a covariable necessarily increases the relative information.

2. The selection process described above is interesting since at any given step h, a covariable which hasn't been selected at a previous step can be included in the selection set E_h.

In practice, the stopping criteria are the following:

i) a maximum number of iterations is determined by the user,

ii) the user determined a maximum threshold beyond which he considers that the relative information is sufficient,

iii) any combination of *i)* and *ii)*,

iv) no covariables can be added.

4 Hierarchisation of a covariable modality

In the case where Y is the survival variable, the quotient of the number of patients alive at time t by the number of uncensored patients just before time t is called survival rate at time t. The survival curve is the curve of the survival rates as a function of t. The recovery rate of a sub-population is the survival rate at the end of the experimentation.

The different modalities of a covariable gives a partitionning of the patient population (*cf. examples presented in fig. 3 and 4*). These modalities are ordered in function of their recovery rates.
If the survival curves of two modalities are very close, they should be grouped in a single modality.

5 Application to the breast cancer

5.1 The data

The data come from an investigation realised in Marseille by Professor J.-M. SPITALIER and the Doctor D. HANS: 1304 patients were observed during 10 years. 45 quantitative or qualitative variables have been observed on each patient. A treatment was applied to the different patients according to the values of the variables.

15 out of 45 variables were retained after a preliminary data analysis: 1-age, 2-clinic class, 3-thermographic class, 4-senographic class, 5-echographic class, 6-clinic PEV, 7-clinic diameter, 8-clinic behaviour, 9-side, 10-histology, 11-N histology, 12-number of invading ganglions, 13-œustradiol receivers, 14-progesterone receivers, 15-UICC stage.

5.2 Selection of the covariables

Computations were made using programs written in APL or Turbo Pascal. The first step allowed to select 5 covariables which were: the thermographic class, the number of invading ganglions, the histology, the N histology and the UICC stage using the following thresholds: $p_1 = 5\%$ and $\alpha = 5\%$ (*cf. Fig. 1*).

The second step confirmed the choice with the following thresholds: $p_2 = 15\%$ and $\alpha = 5\%$ (*cf. Fig. 2*). Nevertheless, at the third step, the best combination of 3 covariables was the age, the histology and the number of invading ganglions: this combination gave about 40% of relative information relative to the death.

5.3 Hierarchisation of the modality of a covariable

If a covariable has a correct pronostic value, it is interesting to be able to assign a survival curve to the different possible values of this covariable (*cf. Fig. 3 and 4*). These curves represent the survival rate conditional to the value taken by this covariable on a patient. They are useful tools for medical pronostics.

The studies which were undertaken separately of the 4 covariables age, histology, number of invading ganglions and UICC stage indicates the modalities which have the best survival

expectation. So, for the covariable number of invading ganglions (*cf. Fig. 3*), 0, 1 or 2 invading ganglions give survival curves that are near and above the average survival whereas the modality "3 to 4 invading ganglions" is clearly less than average. The modality "more than 4 invading ganglions" is the worst.

If the covariable age is taken individually, it carries little information on the survival variable: this is shown by the entwining of the survival curves associated with its modalities (*cf. Fig. 4*).

6 Conclusions

We have presented an asymptotic method which can be applied when the number of statistical units is large. Using this method, some covariables carrying most of the information on the variable of interest can be selected. When applied to patients having breast cancer, four variables were selected. The plotted survival curves for each modality of a covariable gave more precise information on the pronostic values.

References

[1] ARCONTE,A., KHATTABI, MANUCEAU, J., MARTIAS, C., Mutual Information between σ-algebras, *Prépublication UAG*, **95/07**, 1995.

[2] DALEY & VERES-JONES, An introduction to the theory of point processes, *Springer-Verlag*, New-York, 1988.

[3] EL HASNAOUI,A., Le concept du gain d'information:une nouvelle approche en épidémiologie quantitative. *Thèse de doct. - univ. de Montpellier*, 1993.

[4] KULLBACK, S., Information theory and statistics. *Wiley*, 1959.

[5] RAO, C.R., Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhya Series A*, **44**,1-22, 1982.

[6] RAO, C.R., Generalization of ANOVA through entropy and cross entropy functions. In *Probability theory and mathematical statistics*, **2**,477-494, Sciences Press, Utrecht, 1986.

[7] RÉNYI, A., On measures of entropy and information. *Proc. of the fourth Berkeley Symposium on Mathematical Statistics & Probability* **1**,547-561,1961.

[8] ROBERT, C., An entropy concentration theorem: applications in artificial intelligence and descriptive statistics. *J. Appl. Prob.* **27**,303-313, 1990.

[9] SHANNON, CE., A mathematical theory of communication. *Bell System Technical Journal* **27**,379-423 & 623-656, 1948.