REVUE FRANÇAISE D'INFORMATIQUE ET DE RECHERCHE OPÉRATIONNELLE. SÉRIE VERTE

OLIVIER SPIRE

Détermination de l'importance relative de différents paramètres descriptifs pour l'obtention d'un diagnostic

Revue française d'informatique et de recherche opérationnelle. Série verte, tome 4, n° V1 (1970), p. 85-99

<http://www.numdam.org/item?id=RO_1970__4_1_85_0>

R.I.R.O.
(4e année, V-1, 1970, p. 85-99)

DETERMINATION DE L'IMPORTANCE RELATIVE DE DIFFERENTS PARAMETRES DESCRIPTIFS POUR L'OBTENTION D'UN DIAGNOSTIC (1)

par Olivier SPIRE

La méthode que nous présentons a pour objet l'estimation de la probabilité de présence de certains diagnostics, compte tenu des valeurs prises par différents paramètres descriptifs.

Elle repose sur une analyse de l'importance respective de l'observation de chaque paramètre.

Ces paramètres doivent être qualitatifs ou quantitatifs à variations discrètes ; des paramètres continus ne peuvent être pris en compte que s'il est justifiable de discrétiser leurs variations par la création de classes appropriées.

Conçue dans le cadre d'une recherche concernant l'aide au diagnostic médical [A], effectuée en collaboration avec le Dr Naddor (2) pour l'hôpital Johns Hopkins de Baltimore, elle a donné des résultats très encourageants ; l'existence de nombreux autres domaines possibles d'application nous a incité à l'améliorer et à faire ici le point de notre travail.

INTERET DE LA METHODE

Avant d'en exposer les fondements théoriques, précisons par des exemples l'intérêt que présente la méthode dans différents domaines d'application.

Domaine médical

Lorsqu'un médecin procède à l'examen radiologique d'un patient, il observe les valeurs prises par des paramètres permettant de décrire les régions atteintes.

(1) Cet article a fait l'objet d'une conférence prononcée à Toulouse en Mars 1970 à l'occasion des Journées Internationales d'Informatique Médicale.
(2) Johns Hopkins University.

Supposons, pour être précis, que les patients examinés soient atteints d'une maladie de l'intestin grêle (maladie de Crohn, diverticulose, etc.) : voici quelques-uns des paramètres utilisés, et les valeurs qu'ils sont susceptibles de prendre :

PARAMETRES	VALEURS
Quantitatifs	
épaisseur moyenne des valvules iléo-coliques	1 : moins de 1.5 mm 2 : entre 1.5 et 2 mm 3 : plus de 2 mm
Qualitatifs	
existence d'irrégularités sur la paroi de l'iléon	0 : pas d'irrégularité 1 : quelques-unes 2 : nombreuses 3 : très nombreuses

Figure 1

L'expérience acquise par le médecin lui a permis d'établir des relations entre observations et maladies, qui lui servent à former son diagnostic lorsqu'il examine un nouveau patient.

Il éprouve toutefois des difficultés et ce pour deux raisons :

1° Il n'a pas pu appréhender avec suffisamment de précision des relations qui lient aux diagnostics les valeurs prises par certains paramètres ;

2° Les conclusions partielles qu'il peut tirer des examens successifs des valeurs prises par chaque paramètre sont souvent contradictoires.

C'est l'existence de ces difficultés qui nous a orientés vers la conception d'une méthode d'aide au diagnostic médical reposant :

1° sur l'analyse statistique d'une « banque de données » contenant les valeurs des paramètres et le diagnostic correspondant pour un ensemble de cas « confirmés » (par biopsie par exemple) ;

2° sur la détermination de l'importance respective de l'observation de chaque paramètre dans la formation du diagnostic.

Les résultats obtenus sont utilisés pour fournir au médecin une information chiffrée concernant l'éventualité de chaque diagnostic pour un nouveau patient.

Domaine industriel

Les problèmes auxquels la méthode est applicable sont, dans ce domaine, assez variés : contrôle de fabrication, maintenance, recherche de pannes ; citons, à titre d'exemple, le problème de la maintenance d'un réac-

teur d'avion : la complexité de son fonctionnement rend impossible la surveillance de tous les facteurs de panne : il serait donc intéressant, à partir des mesures que l'on peut régulièrement effectuer sur le réacteur, d'estimer la probabilité que le réacteur tombe en panne : ce diagnostic aiderait à déterminer quand il faut procéder à une révision.

Domaine bancaire

Notre méthode pourrait par exemple être appliquée au problème des prêts personnels : l'analyse de la banque de données constituée par les dossiers des anciens clients permettrait de déterminer les critères sur lesquels il faut se fonder en priorité pour estimer les probabilités qu'un client soit bon ou mauvais payeur ; l'estimation, transmise par l'intermédiaire d'un système de télétraitement aussitôt après l'enregistrement des coordonnées du nouveau client, permettrait d'orienter la décision d'accorder ou non le prêt.

DEFINITIONS ET NOTATIONS

Soit $\mathcal{E}$ un ensemble d'éléments γ et $I = \{ 1, ..., i, ..., p \}$ un ensemble fini de paramètres supposés indépendants *(hypothèse 1)* utilisés pour décrire ces éléments.

$\forall\, i \in I$, $v_i(\gamma)$ désigne la valeur prise pour γ par le paramètre i et $\mathcal{V}_i$ l'ensemble des valeurs pouvant être prises par ce paramètre pour les éléments de $\mathcal{E}$:

$$\mathcal{V}_i = \{ v_i(\gamma) \}_{\gamma \in \mathcal{E}} ;$$

Soit $\mathcal{V} = \mathcal{V}_1 \times ... \times \mathcal{V}_i \times ... \times \mathcal{V}_p$.

Supposons *(hypothèse 2)* qu'il existe une application Φ faisant correspondre à tout élément γ de $\mathcal{E}$ un vecteur de $\mathcal{V}$:

$$\Phi(\gamma) = [v_1(\gamma), ..., v_i(\gamma), ..., v_p(\gamma)] ;$$

nous dirons que γ a été « observé » si $\Phi(\gamma)$ a été déterminé.

Soit $D = \{ 1, ..., d, ..., m \}$ un ensemble fini de diagnostics.

Supposons *(hypothèse 3)* qu'il existe une surjection Δ de $\mathcal{E}$ sur D ; le diagnostic $\Delta(\gamma)$, image dans D d'un élément γ de $\mathcal{E}$, est considéré comme « connu » lorsqu'il a été effectivement constaté, indépendamment de l'observation des valeurs $v_i(\gamma)$, $i \in I$.

Soit $\hat{D}$ l'ensemble des mesures de probabilité sur D et δ l'injection qui, à tout diagnostic $\Delta(\gamma)$ de D, fait correspondre le vecteur $\delta[\Delta(\gamma)]$ de $\hat{D}$ dont la $d^{\text{ième}}$ composante est :

$$\delta(d, \gamma) = \begin{cases} 1 & \text{si} \quad \Delta(\gamma) = d \\ 0 & \text{si} \quad \Delta(\gamma) \neq d, \forall\, d \in D, \forall\, \gamma \in \mathcal{E}. \end{cases} \qquad [1]$$

Appelons enfin P une application qui à tout vecteur $\Phi(\gamma)$ de $\mathfrak{V}$ fait correspondre un vecteur $P[\Phi(\gamma)]$ de $\hat{D}$:

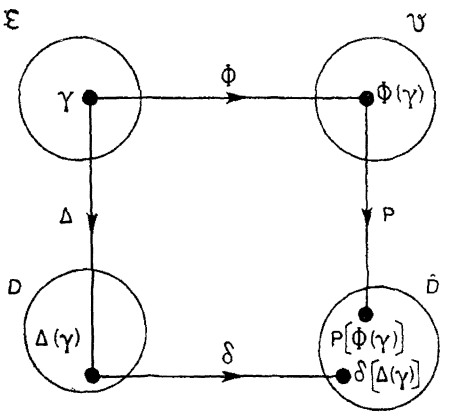

Figure 2

Il résulte de l'hypothèse 3 qu'il existe une partition de $\mathcal{E}$ en sous-ensembles $\mathcal{E}_d$, $d \in D$, telle que :

$$\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2 \ldots \cup \mathcal{E}_d \ldots \cup \mathcal{E}_m,$$

$\mathcal{E}_d$ étant composé exclusivement d'éléments auxquels correspond le diagnostic d ; mais nous ne connaissons pas cette partition, car le diagnostic n'est effectivement connu que pour quelques-uns des éléments de $\mathcal{E}$.

Désignons par E un échantillon de $\mathcal{E}$, supposé représentatif *(hypothèse 4)*, dont tous les éléments ont une image connue dans D ; afin de les particulariser, notons c les éléments de cet échantillon ; soit d'autre part $n = |E|$.

E est la réunion de sous-ensembles E_d, $d \in D$, ne contenant que des éléments pour lesquels d est le diagnostic connu :

$$E = E_1 \cup E_2 \ldots \cup E_d \ldots \cup E_m.$$

Pour tout élément $c \in E$, $\Phi(c)$ (c'est-à-dire $[\varphi_1(c), \ldots, \varphi_i(c), \ldots \varphi_p(c)]$) et $\Delta(c)$ sont connus :

Nous appelons l'ensemble

$$B = \{ c, \{ \varphi_i(c) \}_{i \in I}, \Delta(c) \}_{c \in E}$$

la « banque de données ».

$\forall\, i \in I$, notons, d'autre part, V_i l'ensemble des valeurs prises par le paramètre i dans E :

$$V_i = \{ \varphi_i(c) \}_{c \in E} ; \; V_i \subset \mathfrak{V}_i.$$

POSITION DU PROBLEME

Lorsque le diagnostic $\Delta(\gamma)$ portant sur un élément observé de $\mathcal{E} - E$ est inconnu, il est intéressant d'estimer $\delta[\Delta(\gamma)]$ en utilisant $\Phi(\gamma)$.

Nous nous proposons donc de définir un critère d'estimation C, puis de déterminer l'application P de telle sorte que $P[\Phi(\gamma)]$ soit une estimation de $\delta[\Delta(\gamma)]$ selon ce critère.

Nous limiterons notre recherche à une classe particulière d'applications linéaires.

La méthode que nous utiliserons repose sur une analyse statistique du contenu de la banque de données ; sa caractéristique principale réside dans la détermination d'un ensemble de coefficients $\{ x_i \}_{i \in I}$, x_i reflétant l'importance relative de l'observation du paramètre i pour estimer $\delta[\Delta(\gamma)]$ selon C.

La $d^{\text{ième}}$ composante de $P[\Phi(\gamma)]$, $P(d, \gamma)$ pourra être interprétée comme une estimation de la probabilité que $\Delta(\gamma) \in \mathcal{E}_d$.

CHOIX DU CRITERE

Pour tout élément $c \in E$, définissons un écart $e(c)$ entre l'estimation $P[\Phi(c)]$ et le vecteur connu $\delta[\Delta(c)]$:

$$e(c) = \sum_{d \in D} [P(d, c) - \delta(d, c)]^2. \qquad [2]$$

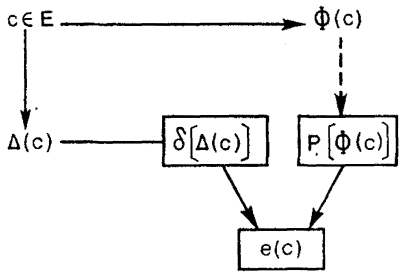

Figure 3

Nous proposons de déterminer P de manière à rendre minimale la somme y des écarts $e(c)$ pour tous les éléments de E :

$$C : \text{Min } y = \sum_{c \in E} e(c) = \sum_{c \in E} \sum_{d \in D} [P(d, c) - \delta(d, c)]^2. \qquad [3]$$

On reconnaît dans l'expression à minimiser l'espérance mathématique d'une fonction de coût quadratique, ce qui est la caractéristique de l'estimation bayésienne.

Pour illustrer la méthode, nous envisagerons un cas simple où

$$p = 3\,;\ m = 2\,;\ n = 5\,;\ V_1 = \{0, 1\}\,;\ \mathfrak{V}_1 = \{0, 1, 2\}\,;$$
$$V_2 = \mathfrak{V}_2 = V_3 = \mathfrak{V}_3 = \{0, 1, 2\}.$$

Nous raisonnerons sur la banque de données B1 de la figure 4 et calculerons $\{ P(d, \gamma) \}_{d \in \{1,2\}}$ pour les éléments de $\mathcal{E} - E$ décrits dans le tableau de la figure 5.

BANQUE DE DONNEES B1

Elément c	$\Phi(c)$ Valeurs prises par les paramètres			Diagnostic $\Delta(c)$	Vecteur $\delta[\Delta(c)]$	
	$\nu_1(c)$	$\nu_2(c)$	$\nu_3(c)$		$\delta(1, c)$	$\delta(2, c)$
1	0	0	1	2	0	1
2	0	2	2	1	1	0
3	1	1	0	2	0	1
4	1	2	0	2	0	1
5	1	1	1	1	1	0

Figure 4

ELEMENT γ	$\Phi(\gamma)$ VALEURS PRISES PAR LES PARAMETRES			ESTIMATION $P[\Phi(\gamma)]$	
	$\nu_1(\gamma)$	$\nu_2(\gamma)$	$\nu_3(\gamma)$	$P(1, \gamma)$	$P(2, \gamma)$
6	1	1	2	?	?
7	0	0	0	?	?
8	2	1	1	?	?
9	1	2	2	?	?
10	1	2	2	?	?

Figure 5

PROBABILITES A POSTERIORI

Soit $h(d/\varphi_i)$ la probabilité *a posteriori* de présence du diagnostic d lorsque le paramètre i prend la valeur φ_i.

Supposons tout d'abord $\varphi_i \in V_i$: compte tenu des hypothèses faites, $h(d/\varphi_i)$ est la fréquence relative d'apparition dans B du diagnostic d, lorsque le paramètre i prend la valeur φ_i, et peut être calculée par la relation :

$$h(d/\varphi_i) = \frac{n(\varphi_i, d)}{n(\varphi_i)} \quad , \quad \forall\, \varphi_i \in V_i, \qquad [4]$$

dans laquelle :

— $n(\varphi_i, d)$ est le nombre d'apparitions dans B de la valeur φ_i quand le diagnostic est d :

$$n(\varphi_i, d) = |C_1|, \qquad [5]$$

avec :

$$C_1 = \{\, c : c \in E,\ \Delta(c) = d,\ \varphi_i(c) = \varphi_i \,\};$$

— $n(\varphi_i)$ est le nombre d'apparitions dans B de la valeur φ_i :

$$n(\varphi_i) = |C_2|, \qquad [6]$$

avec :

$$C_2 = \{\, c : c \in E,\ \varphi_i(c) = \varphi_i \,\};$$

puisque $\varphi_i \in V_i$, $n(\varphi_i) \neq 0$; d'autre part :

$$n(\varphi_i) = \sum_{d=1}^{m} n(\varphi_i, d). \qquad [7]$$

Pour toute valeur φ_i susceptible d'être prise par le paramètre i, mais qui n'apparaît pas dans B ($\varphi_i \in \mathfrak{V}_i - V_i$), il est impossible de calculer $h(d/\varphi_i)$ comme précédemment. Nous supposons cette probabilité égale à la fréquence d'apparition du diagnostic d dans B *(hypothèse 5)* :

$$h(d/\varphi_i) = \frac{n(d)}{n} \quad , \quad \forall\, \varphi_i \in \mathfrak{V}_i - V_i, \qquad [8]$$

où $n(d) = |E_d|$; ceci revient à supposer que les événements d et φ_i sont indépendants, ce qui semble logique, dans la pratique, puisque nous ne disposons d'aucune information mettant en évidence leur dépendance.

A titre d'exemple examinons, dans B1, les valeurs $\varphi_1(c)$, $c = 1,5$ prises par le premier paramètre, et les diagnostics correspondant $\Delta(c)$, $c = 1,5$:

$$\begin{cases} \varphi_1(1) = 0 & \Delta(1) = 2 \\ \varphi_1(2) = 0 & \Delta(2) = 1 \end{cases}$$

$$\begin{cases} \varphi_1(3) = 1 & \Delta(3) = 2 \\ \varphi_1(4) = 1 & \Delta(4) = 2 \\ \varphi_1(5) = 1 & \Delta(5) = 1 \end{cases}$$

Les fréquences relatives d'apparition des diagnostics 1 et 2, lorsque $\varphi_1 = 0$, sont respectivement 1/2 et 1/2 ; quand $\varphi_1 = 1$, ces fréquences sont respectivement 1/3 et 2/3 :

$$h(1/\varphi_1 = 0) = 1/2 \quad \text{et} \quad h(2/\varphi_1 = 0) = 1/2 ;$$
$$h(1/\varphi_1 = 1) = 1/3 \quad \text{et} \quad h(2/\varphi_1 = 1) = 2/3.$$

Remarquons que la valeur $\varphi_1 = 2$ (qui apparaît pour l'élément 9 de la fig. 5) ne se trouve pas dans B1 ; les fréquences relatives d'apparition dans B1 des diagnostics 1 et 2 étant 2/5 et 3/5, nous supposons :

$$h(1/\varphi_1 = 2) = 2/5 \quad \text{et} \quad h(2/\varphi_1 = 2) = 3/5.$$

Appelons H_1 l'ensemble

$$\{ h(d/\varphi_i) \}_{i \in I,\ \varphi_i \in \mathfrak{V}_i,\ d \in D}.$$

Les éléments de H_1, calculés à partir de B1, sont donnés dans le tableau de la figure 6.

	$h(d/\varphi_i)$			
φ_i	$i = 1$	$i = 2$	$i = 3$	d
0	1/2 1/2	0 1	0 1	1 2
1	1/3 2/3	1/2 1/2	1/2 1/2	1 2
2	2/5 3/5	1/2 1/2	1 0	1 2

Figure 6

DETERMINATION DE L'APPLICATION *P*

Considérons un élément $\gamma \in \mathcal{E}$. A l'observation de chaque paramètre $i \in I$, correspond une valeur $\varphi_i(\gamma)$ et une probabilité $h[d/\varphi_i(\gamma)]$ que $\gamma \in \mathcal{E}_d$.

Envisageons l'hypothèse selon laquelle $P(d, \gamma)$ peut être obtenu par combinaison linéaire des probabilités $h[d/\varphi_i(\gamma)]$, $i \in I$ *(hypothèse 6)* :

$$P(d, \gamma) = \sum_{i \in I} x_i \cdot h[d/\varphi_i(\gamma)]. \qquad [9]$$

Pour que les relations

$$P(d, \gamma) \geqslant 0 \quad , \quad \forall\, d \in D \quad \text{et} \quad \sum_{d \in D} P(d, \gamma) = 1 \qquad [10]$$

soient vérifiées, les contraintes

$$x_i \geqslant 0 \quad , \quad \forall\, i \in I \quad \text{et} \quad \sum_{i \in I} x_i = 1 \qquad [11]$$

doivent être respectées.

Les coefficients x_i, $i \in I$, doivent d'autre part être tels que le critère C soit satisfait.

Nous devons donc, pour les obtenir, résoudre le problème suivant :

> Déterminer les coefficients x_i, $i \in I$, satisfaisant les contraintes :
>
> $$x_i \geqslant 0 \quad , \quad \forall\, i \in I$$
>
> et
>
> $$\sum_{i \in I} x_i = 1,$$
>
> qui rendent minimale la somme :
>
> $$y = \sum_{c \in E} \sum_{d \in D} [P(d, c) - \delta(d, c)]^2,$$
>
> dans laquelle, $\forall\, d \in D$ et $\forall\, c \in E$:
>
> $$P(d, c) = \sum_{i \in I} x_i \cdot h[d/\varphi_i(c)]$$
>
> et
>
> $$\delta(d, c) = \begin{cases} 1 & \text{si} \quad \Delta(c) = d \\ 0 & \text{si} \quad \Delta(c) \neq d \end{cases}$$

y peut s'écrire :

$$y = \underbrace{\sum_c \sum_d \left(\sum_i x_i \cdot h[d/\varphi_i(c)] \right)^2}_{Q} \underbrace{- 2 \sum_c \sum_d \left(\delta(d, c) \sum_i x_i \cdot h[d/\varphi_i(c)] \right)}_{L} + \underbrace{\sum_c \sum_d \delta^2(d, c)}_{K} \qquad [12]$$

— Q est une *forme quadratique* dont l'expression matricielle est $Q = X'RX$:

X est le vecteur colonne $\begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_p \end{bmatrix}$ et R la matrice symétrique d'ordre p associée à Q, dont les éléments r_{ij} sont donnés par :

$$r_{ij} = \sum_{c \in E} \sum_{d \in D} h[d/\varphi_i(c)] \cdot h[d/\varphi_j(c)] \quad , \quad \forall\, (i, j) \in I. \qquad [13]$$

(X' désigne la matrice transposée de X).

— L est une *expression linéaire* dont la notation matricielle est $L = 2T'X$:

T est le vecteur colonne d'ordre p dont les éléments t_i sont donnés par :

$$t_i = \sum_c \sum_d \delta(d, c) \cdot h[d/\rho_i(c)] = \sum_c h[d(c)/\rho_i(c)] \quad , \quad \forall\, i \in I \qquad [14]$$

(il est facile de montrer que $t_i = r_{i,i}$).

— K est une constante valant n.

La fonction à minimiser est donc

$$y = X'RX - 2T'X + n, \qquad [15]$$

ou, ce qui revient au même :

$$y^* = X'RX - 2\,T'X. \qquad [16]$$

$\forall\, (i, j) \in I$, $r_{ij} \geqslant 0$; dans le domaine défini par $X \geqslant \tilde{0}$ et $\sum_{i \in I} x_i = 1$, R est donc semi définie positive et $X'RX$ est convexe ;

— $2T'X$, qui représente un hyperplan, est aussi convexe ; y^* est par conséquent *convexe.*

L'algorithme de programmation quadratique proposé par Dantzig [B], adapté au cas où la fonction à optimiser contient également des termes linéaires, peut être utilisé pour résoudre notre problème.

Chaque *coefficient* x_i *peut être interprété comme caractérisant l'importance* relative de l'information découlant de l'observation du paramètre i.

Dans notre exemple :

$$R = \begin{bmatrix} 8/3 & 5/2 & 17/6 \\ 5/2 & 3 & 5/2 \\ 17/6 & 5/2 & 4 \end{bmatrix} \quad , \quad T = \begin{bmatrix} 8/3 \\ 3 \\ 4 \end{bmatrix}$$

la résolution donne pour résultats :

$$x_1 = 0\,;\, x_2 = 1/4\,;\, x_3 = 3/4,$$

et nous obtenons les estimations suivantes pour les éléments de E :

Element c	Diagnostic connu $\Delta(c)$	Estimation $P[\Phi(c)]$	
		$P(1, c)$	$P(2, c)$
1	2	.375	.625
2	1	.875	.125
3	2	.125	.875
4	2	.125	.875
5	1	.500	.500

Figure 7

De manière à pouvoir comparer l'estimation obtenue avec d'autres estimations, il est intéressant de définir une « mesure d'efficacité » liée à la valeur de y : y variant entre 0 et $2n$, nous pouvons, par exemple, adopter :

$$E = 1 - y/2n. \qquad [17]$$

Quel que soit n, E varie dans l'intervalle [0,1].

Dans notre exemple, $y = 7/8$ et donc $E = 0.9125$; avec

$$x_1 = x_2 = x_3 \simeq .33,$$

nous aurions obtenu $y = 1.3608$, d'où $E = .8639$.

CALCUL DES ESTIMATIONS

Connaissant H_1 et $\{x_i\}_{i \in I}$, nous pouvons désormais déterminer $P[\Phi(\gamma)]$ pour tout élément observé $\gamma \in \mathcal{E} - E$.

Les résultats suivants sont obtenus pour les éléments de la figure 5 :

Element γ	Estimation $P[\Phi(\gamma)]$	
	$P(1, \gamma)$	$P(2, \gamma)$
6	.875	.125
7	0	1.000
8	.125	.875
9	.500	.500
10	.875	.125

Figure 8

Regroupons maintenant données et résultats sur une seule figure :

	élément γ	Diag-nostic $\Delta(\gamma)$	$-\Phi(\gamma)-$ valeurs prises par les paramètres $v_1(\gamma)$	$v_2(\gamma)$	$v_3(\gamma)$	Estimation $P[\Phi(\gamma)]$ $P(1,\gamma)$	$P(2,\gamma)$	
$\gamma=c\in E$	1	2	0	0	1	.375	.625	$y=7/8$
	2	1	0	2	2	.875	.125	
	3	2	1	1	0	.125	.875	$E=.9125$
	4	2	1	2	0	.125	.875	
	5	1	1	1	1	.500	.500	
	6	?	1	1	2	.875	.125	
	7	?	0	0	0	0	1 000	
	8	?	0	1	0	.125	.875	
	9	?	2	1	1	.500	.500	
	10	?	1	2	2	.875	.125	
			$x_1=0$	$x_2=1/4$	$x_3=3/4$			

Figure 9

(la partie hachurée représente B1).

REVISION DE L'HYPOTHESE 6

$\forall\ \alpha \leqslant i_m$, soit $\{ I_k^\alpha \}_{k=1, C_d^\alpha}$ l'ensemble des combinaisons des p paramètres de I pris α à α et $h[d/\{ v_i \}_{i\in I_k^\alpha}]$ la probabilité *a posteriori* que le diagnostic soit d, étant donné l'ensemble de valeurs $\{ v_i \}_{i\in I_k^\alpha}$:

Soit $$H_\alpha = \{ h[d/\{ v_i \}_{i\in I_k^\alpha}] \}_{k=1, C_p^\alpha\ ;\ v_i\in \mathcal{V}_i,\ \forall i\in I_k^\alpha\ ;\ d\in D'} \qquad [18]$$

et $\forall\ \gamma \in \mathcal{E}, \forall\ d \in D$: [19]

$$H_\alpha^{d,\gamma} = \{ h[d/\{ v_i(\gamma) \}_{i\in I_k^\alpha}] \}_{1, C_p^\alpha}.$$

D'après l'hypothèse 6, les seuls éléments pris en compte pour le calcul de $P(d, \gamma)$ sont ceux de $H_1^{d,\gamma}$; une meilleure estimation serait obtenue en combinant linéairement les éléments de tous les ensembles $H_\alpha^{d,\gamma}$, $\alpha = 1, i_m$: pratiquement l'importance des calculs nous limite à la prise en compte des éléments de $H_2^{d,\gamma}$.

L'expression de $P(d, \gamma)$ est alors :

$$P(d, \gamma) = \sum_{i=1}^{p} x_i \cdot h[d/\varphi_i(\gamma)] + \sum_{i=1}^{p} \sum_{j=i+1}^{p} x_{ij} \cdot h[d/\varphi_i(\gamma), \varphi_j(\gamma)]. \qquad [20]$$

Les éléments de H_2 sont calculés de la manière suivante : lorsque $\varphi_i \in V_i$ et $\varphi_j \in V_j$:

$$h(d/\varphi_i, \varphi_j) = \frac{n(\varphi_i, \varphi_j, d)}{n(\varphi_i, \varphi_j)}, \qquad [21]$$

relation dans laquelle n est définie comme précédemment ; si $\varphi_i \in V_i$ et $\varphi_j \notin V_j$: nous supposons (1) :

$$h(d/\varphi_i, \varphi_j) = h(d, \varphi_i) ; \qquad [22]$$

si $\varphi_i \notin V_i$ et $\varphi_j \notin V_j$: nous supposons (1)

$$h(d/\varphi_i, \varphi_j) = \frac{n(d)}{n}. \qquad [23]$$

Le temps de traitement sur ordinateur dépend du nombre des probabilités *a posteriori* qu'il faut effectivement calculer, mettre en mémoire, et rechercher chaque fois que besoin est : lorsque l'on calcule seulement les éléments de H_1, ce nombre est $n_{H_1} = m \left[1 + \sum_{i=1}^{p} n_i \right]$ (n_i étant le nombre d'éléments de V_i) ; [24]

Si l'on prend en compte les éléments de H_2, le nombre de calculs supplémentaires est :

$$n_{H_2} = \frac{1}{2} m \sum_{i=1}^{p} \left(n_i \sum_{\substack{j=1 \\ j \neq i}}^{p} n_j \right) = m \sum_{i=1}^{p} \left(n_i \sum_{j=i+1}^{p} n_j \right). \qquad [25]$$

en supposant $n_i = s$, $\forall\, i \in I$, nous aurions :

$$n_{H_1} = m(1 + p \cdot s) \quad \text{et} \quad n_{H_2} = \frac{1}{2} m(p - 1)p \cdot s^2 ;$$

avec 5 diagnostics, 10 paramètres et 20 valeurs possibles pour chaque paramètre, n_{H_1} est de l'ordre de 1 000 tandis que n_{H_2} vaut déjà 90 000.

Le nombre θ d'éléments du tableau à traiter par pivotage au cours de l'optimisation influe d'autre part de façon cruciale sur le temps de traitement : s'il faut déterminer seulement les x_i, $i \in I$:

$$\theta = p(2p + 1) ; \qquad [26]$$

(1) cf. plus haut : justification de l'hypothèse 5.

mais s'il faut aussi calculer les x_{ij}, $i \in I$, $j = i + 1, p$:

$$\theta = \frac{p(2p+1)(2p^2+p+1)}{2}. \qquad [27]$$

CONCLUSION

Le schéma de la figure 9 met en évidence les liens entre les principales étapes de la méthode.

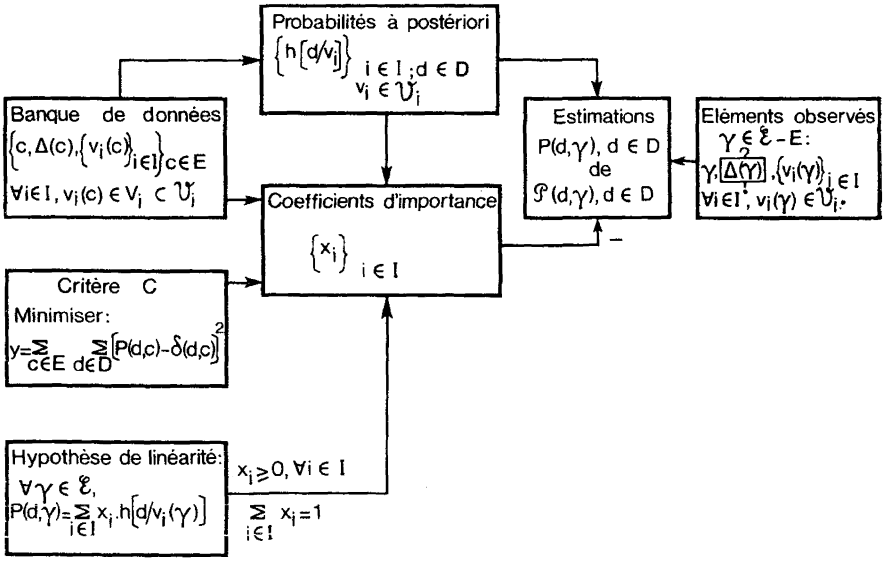

Figure 10

La validité de la méthode repose de façon cruciale sur le respect de l'hypothèse 4 (représentativité de l'échantillon E) : un soin extrême doit donc être apporté à la constitution de la banque de données.

Par ailleurs, une analyse préliminaire est indispensable pour sélectionner les paramètres qui peuvent être considérés comme indépendants (hypothèse 1).

Malgré ces précautions, les hypothèses 1 et 4 ne sont jamais vérifiées *stricto sensu* dans la pratique : le biais qui en résulte se compose aux biais découlant

— de la forme particulière choisie pour exprimer $P(d, \gamma)$ (hypothèse 6),

— des suppositions faites à propos de certains éléments de H_1 et H_2 (hypothèse 5).

Pour appréhender l'importance du biais résultant, il faut tester les estimations en les calculant pour des éléments observés de $\mathcal{E}$-E auxquels correspond un diagnostic connu.

REFERENCES

[A] : Communication au Congrès ORSA-ORSIS, Tel-Aviv, juillet 1969.

[B] : Dantzig G. B., Linear Programming and Extensions, Princeton University Press, Princeton N.J., 1963.

[C] : Ferguson T. S., Mathematical Statistics (A decision theoratic approach), Academic Press, N.Y. and London, 1967.

Le Directeur de la Publication : Georges Dunod. — *Imprimé en France.*
Dépôt légal : 2e trimestre 1970. N° 6479.
Imprimerie Nouvelle, Orléans. — N° 6155.