REVUE DE STATISTIQUE APPLIQUÉE

R. C. CROS

Construction automatique d'un questionnaire arborescent pour l'aide au diagnostic

Revue de statistique appliquée, tome 23, n° 3 (1975), p. 87-91

<http://www.numdam.org/item?id=RSA_1975__23_3_87_0>

CONSTRUCTION AUTOMATIQUE D'UN QUESTIONNAIRE ARBORESCENT POUR L'AIDE AU DIAGNOSTIC [1]

R. C. CROS
Laboratoire de Pathologie Digestive Inserm (2)

Pour l'élucidation des problèmes de pathologie et pour des raisons pratiques, il est intéressant d'essayer d'établir une hiérachie parmi des données cliniques et biologiques caractérisant une population comportant plusieurs groupes de malades différents (et éventuellement un groupe de sujets témoins). Pour simplifier, nous supposerons que les groupes sont disjoints, c'est-à-dire qu'aucun sujet de la population n'est atteint simultanément de deux ou plusieurs des affections à étudier. L'ensemble des groupes réunis définit la population totale qui permet, en retour, de définir un anti-groupe pour chaque groupe, à savoir l'ensemble complémentaire obtenu par soustraction d'un groupe à la population totale.

La distribution de chaque variable étant établie pour chaque groupe et chaque anti-groupe correspondant, on recherche tous les seuils à partir de la valeur la plus basse jusqu'à la plus élevée (ce qui se réduit dans le cas d'une donnée de type oui-non à l'étude de la différence entre oui et non) et l'on calcule toutes les valeurs de χ^2 définies par les quatre paramètres suivants :

A gvs : nombre de sujets du groupe g pour lesquels la variable v est inférieure ou égale au seuil s

B gvs : nombre de sujets du groupe g pour lesquels la variable v est supérieure au seuil s

C gvs : nombre de sujets de l'anti-groupe g pour lesquels la variable v est inférieure ou égale au seuil s

D gvs : nombre de sujets de l'anti-groupe g pour lesquels la variable v est supérieure au seuil s

Appelons K gvs la valeur de χ^2 ainsi obtenue,

S vs la moyenne arithmétique des K gvs de tous les groupes

P vs la moyenne géométrique des K gvs de tous les groupes

et R vs le produit S vs × P vs

En fonction d'une option préalable, nous choisissons de maximiser la fonction S vs, P vs ou R vs pour déterminer la "meilleure" question initiale séparant la population totale T en deux sous-populations U et V au moyen de la variable v et du seuil s retenu.

---

(1) Article remis en Mars 1974, révisé en juin 1974.

(2) 46 chemin de la Gaye, 13009 Marseille

Le principe itératif est immédiat : chacune des deux sous-populations prise isolément comporte un certain nombre de groupes, éventuellement moindre que la population T de départ, permettant de définir de nouveaux anti-groupes. On recherchera donc, selon le même processus, à séparer U_1 et U_2, V_1 et V_2, etc. Les calculs sont effectués avec le nombre nécessaire de dichotomies pour arriver à des sous-ensembles où tous les sujets appartiennent au même groupe, ou, à défaut, à un sous-ensemble hétérogène qui ne peut plus être subdivisé par aucun critère. Bien entendu, les valeurs S vs, P vs et R vs données à chaque subdivision permettent d'apprécier la valeur discriminative des critères sucessivement obrenus.

En général, la maximisation de S vs aboutit à l'établissement d'un arbre irrégulier ou les groupes de sujets très typiques (très différents de la moyenne) sont classés rapidement avec peu de critères, tandis que les branches correspondant aux groupes atypiques sont beaucoup plus ramifiées. Au contraire, la maximisation de P vs conduit à égaliser les niveaux auxquels les différents groupes sont décelés. La fonction R vs apporte un compromis entre ces deux tendances.

Lorsque deux ou plusieurs variables aboutissent au même maximum, l'algorithme retient celle qui a déjà été le plus souvent utilisée comme critère dichotomique. En cas d'égalité, on choisit la variable pour laquelle l'écart relatif est maximum (c'est-à-dire que la première sous-population ayant des valeurs comprises entre v_{min} et vs_1 et la seconde des valeurs comprises entre vs_2 et v_{max}, on recherche le rapport $\frac{vs_2 - vs_1}{v_{max} - v_{min}}$). Si les écarts relatifs sont à leur tour égaux, la première variable dans l'ordre de rangement de la matrice est arbitrairement sélectionnée.

Si l'on désire construire un questionnaire permettant d'aboutir plus rapidement à la détermination d'un ou de certains groupes, on peut donner une pondération Qg pour chaque groupe et calculer les facteurs

$$K' \text{ gvs} = K \text{ gvs} \times Qg$$

qui serviront à déterminer les fonctions S, P et R.

Le programme Fortran que nous avons mis au point permet de construire un arbre comportant au maximum 100 nœuds à partir d'une matrice de 300 sujets caractérisés par 60 variables et appartenant à 10 groupes différents.

A partir de 6 groupes comportant respectivement 16, 16, 24, 47. 18 et 16 sujets (soit un total de 137) caractérisés par 24 variables numériques (soit environ 450 seuils), ce programme a construit les 3 arbres définis par les maximisations de S vs, P vs et R vs comportant respectivement 31, 28 et 33 nœuds en 38 secondes sur UNIVAC 1110. Notons cependant que la matrice des données comportait des trous et que le programme rejette automatiquement les sujets non classables (quand une variable retenue comme critère correspond à une donnée inconnue pour certains sujets de la sous-population étudiée). Il restait au niveau des branches terminales $10 + 13 + 21 + 42 + 17 + 5 = 108$ sujets. Si l'on refuse des S vs $\leqslant 3.00$, il y a $3 + 3 + 4 + 3 + 2 + 2 = 17$ sujets sur ces 108 qui restent inclassables.

On ne peut donc considérer cette méthode comme suffisante pour résoudre tous les problèmes de l'aide au diagnostic. Néanmoins, les arbres construits font clairement ressortir la subordination des différentes variables et séparant bien les sujets typiques d'un groupe de ceux qui posent des problèmes diagnostics difficiles.

On pourrait penser que la recherche systématique de dichotomies soit assez artificielle, car a priori il n'est pas impossible qu'une subdivision comportant plus de 2 branches soit utile à certaines discriminations. En fait, dans l'exemple cité plus haut, il y a deux exemples de divisions à 3 branches qui résultent de la sélection répétée de la même variable avec deux seuils différents à deux niveaux successifs de l'arbre. Bien que les résultats soient présentés comme une succession de divisions dichotomiques, ces nœuds multiples n'échappent guère à l'œil de celui qui examine les résultats.

La sortie des résultats sous forme tabulaire permet de tracer aisément à la main l'arbre obtenu. On peut envisager le recours à un algorithme de traçage pour automatiser cette phase terminale du processus.

L'exemple que nous soumettons appelle quelque commentaires :

1/ Il convient d'être très circonspect en ce qui concerne la validité des critères de dichotomie lorsque l'effectif des groupes résiduels tombe à 1 ou 2 sujets. Ces branches qui ne sont représentées qu'à titre illustratif et on peut considérer que la méthode a un pouvoir discriminant limité à un test situé en amont, conduisant à une population hétérogène qu'il convient éventuellement d'analyser à l'aide d'autres méthodes statistiques ou heuristiques.

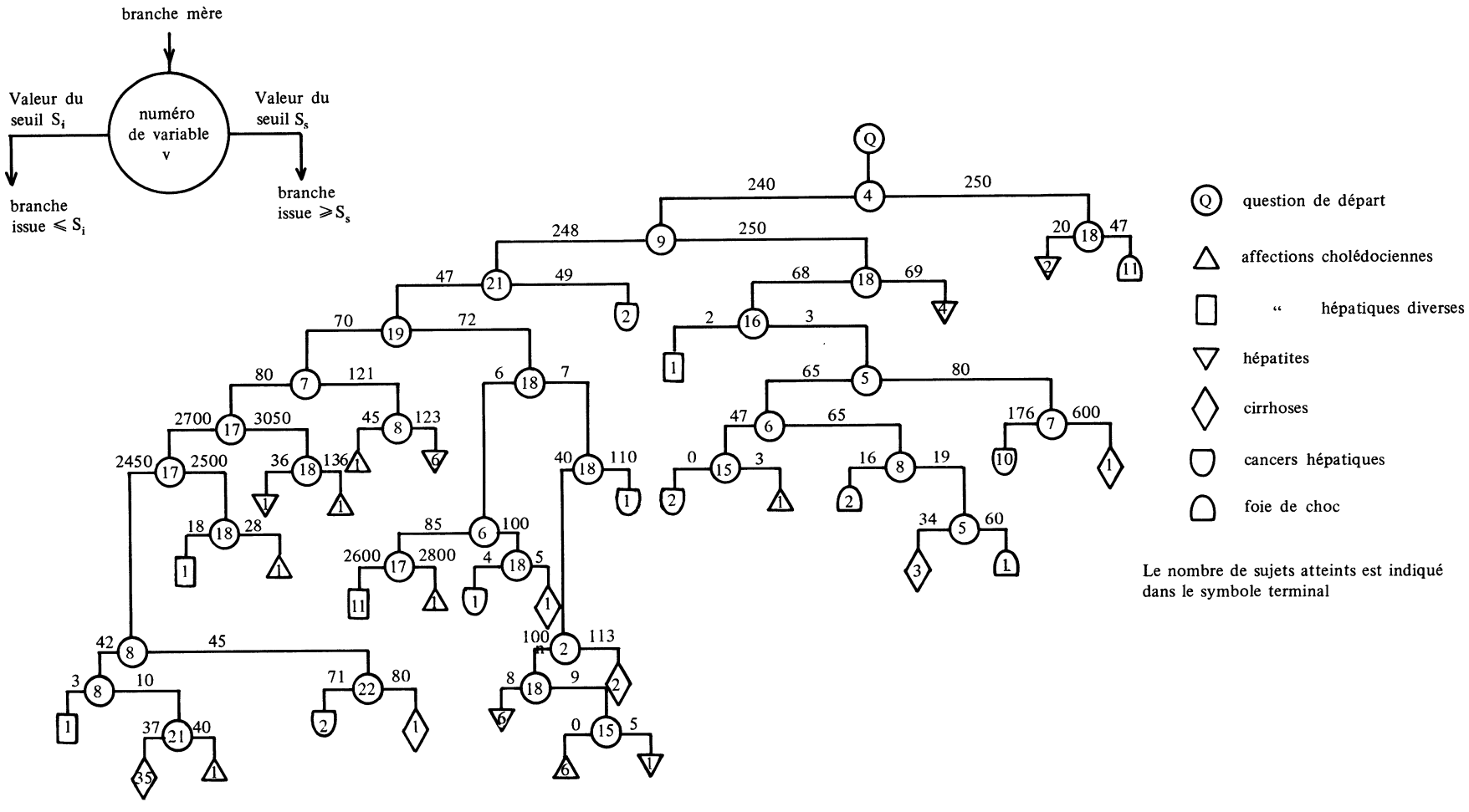
2/ Comme il est malheureusement fréquent que les données soient partiellement manquantes, il y a un risque que l'algorithme ne sélectionne pas les bonnes variables si la distribution des données manquantes à l'intérieur de chacun des groupes diffère appréciablement de la distribution des données connues. C'est pourquoi on ne saurait trop se méfier des résultats obtenus à partir de matrices quelque peu lacunaires.

3/ En dépit de cette limitation, nous avons appliqué la méthode à un ensemble de données comportant à peu près 40 % d'inconnues ! . Les calibres de 20 artères hépatiques et pancréatiques de 82 sujets normaux + 69 pancréatites chroniques + 33 cancers du pancréas. Bien que le programme ne permette pas de classer plus de 24 + 13 + 7 de ces sujets, les critères majeurs de distinction entre les groupes (en particulier l'hypervascularisation des branches hépatiques fines) correspondent bien à l'opinion des experts de ce domaine.

4/ La méthode que nous proposons peut dans une certaine mesure être rapprochée d'un algorithme testé aux Etats-Unis il y a quelques années et dont le principe est le suivant : partant du même ensemble de variables et de sujets, on recherche la variable possédant une valeur seuil en deça ou au delà de laquelle tous les sujets font partie du même groupe. Ayant éliminé par un seul critère le plus grand nombre de sujets possible, on procède à nouveau à la même analyse sur la population réduite etc. Cette méthode très simple d'érosion progressive du problème posé peut, et c'est un avantage très considérable, être appliquée sans ordinateur à des matrices qui ne sont pas trop volumineuses. Mais on peut penser qu'elle est très sensible aux données manquantes car à chaque étape d'élimination la connaissance d'une seule donnée manquante peut modifier le seuil ou la variable retenue.

Dans un cas comme dans l'autre, écrémage des distributions par les extrémités ou coupure au voisinage du centre, on aboutit à des résultats très satisfaisants pour les sujets les plus typiques de chaque groupe et pour les groupes les plus caractéristiques.

A nos yeux, cependant, l'intérêt principal de ces méthodes est qu'elles proposent une hiérarchie de variables discriminantes, génératrices d'hypothèses nouvelles pour le médecin et le biologiste.


branche mère
Valeur du seuil S_i
numéro de variable v
Valeur du seuil S_s
branche issue $\leqslant S_i$
branche issue $\geqslant S_s$
Q
question de départ
affections cholédociennes
" hépatiques diverses
hépatites
cirrhoses
cancers hépatiques
foie de choc
Le nombre de sujets atteints est indiqué dans le symbole terminal