Statistique et analyse des données

F. Murtagh

Une question de classifiabilité en classification automatique dans le cas particulier des rassemblements documentaires

Statistique et analyse des données, tome 5, n° 1 (1980), p. 77-89

<http://www.numdam.org/item?id=SAD_1980__5_1_77_0>

Statistique et Analyse des Données
1 - 1980 pp. 77-89

UNE QUESTION DE CLASSIFIABILITE EN CLASSIFICATION AUTOMATIQUE DANS LE CAS PARTICULIER DES RASSEMBLEMENTS DOCUMENTAIRES

F. MURTAGH

F. Murtagh, Laboratoire de Statistiques Mathématiques, Université de Paris VI.

INTRODUCTION

S'inspirant d'un article de Jones (1973), nous tentons de répondre au problème le plus ancien dans le domaine de la classification automatique : celui de la classifiabilité des données. Nous esquissons d'abord le problème que Jones a traité ; puis nous décrirons brièvement les données utilisées ; et ensuite nous exposerons la démarche à suivre.

Les travaux que Jones a menés appartiennent au domaine de la classification automatique, appliquée à l'interrogation des rassemblements documentaires. D'emblée, donc, nous avons affaire à des données d'une forme familière en ce qui concerne l'analyse des données : on aborde une analyse avec un tableau d'incidence binaire d'individus et de caractères, ou encore de documents et de leurs mots-clés. Quand on cherche des documents (des articles, des résumés, des autres spécifications) à partir de leurs contenus, on se trouve en présence d'un problème majeur : comment retirer du rassemblement documentaire les documents les plus pertinents. Afin de mettre en marche un système d'interrogation efficace, on affecte les documents à des classes de documents semblables. Cette démarche permet la réduction de la base des données à ces classes. Ensuite on suppose que l'on désire la sortie de toute la classe à la fois ; ou, s'il s'agit d'une hiérarchie de classes, que l'on désire la sortie d'une classe entière à un niveau déterminé de la hiérarchie. Afin d'édifier une classification, on peut se servir de toutes les diverses méthodes de la classification automatique : pour une revue de ce domaine, on se réfère à Salton et Wong (1978).

Une seule caractéristique distingue ce domaine des autres pour lesquels on utilise la classification automatique : cette différence est celle de la taille des données. A titre indicatif, l'échantillon que nous avons utilisé comprenait quelque 6000 documents et 1500 mots-clés, en

total. Cet échantillon appartenait au rassemblement documentaire de l'International food information service. Chaque mois cet organisme produit une bande magnétique qui comporte des renseignements documentaires, y compris des mots-clés (voir The IFIS tape manual, 1972 ; ou Murtagh, 1979).

Revenons maintenant au problème de Jones (1973) qui a comparé plusieurs rassemblements documentaires différents. Au fond, la question posée est la suivante : est-ce qu'un rassemblement est classifiable ou non ? Plus exactement, puisque l'on peut toujours faire passer un programme de classification automatique, cette question s'exprime comme suit : est-ce que l'on peut arriver à un indice de classifiabilité d'un rassemblement documentaire ?

Jones a discuté des propriétés des rassemblements comme des nombres rares, fréquents ou moyens de mots-clés associés à chaque document (on se rappelle que cela revient à dire, le nombre des unités dans la ligne du tableau d'incidence qui appartient au document ; ou encore, en langage ensembliste, l'effectif du document) ; des nombres de documents associés à chaque mot-clé ; etc. Nous allons adopter une telle approche pour caractériser le rassemblement.

Signalons, au préalable, que le rassemblement documentaire n'est autre, d'une façon générale, que l'ensemble fini des individus que l'on cherche à classer. La première étape est alors de trouver un résumé fondamental d'une (ou de quelques) propriété(s) du rassemblement documentaire ou de cet ensemble d'individus. Nous avons choisi comme une telle propriété le nombre de mots-clés qui appartiennent aux documents ; ou encore les effectifs des documents. Nous avons été amenés à cet aspect du rassemblement en vertu des faits suivants :

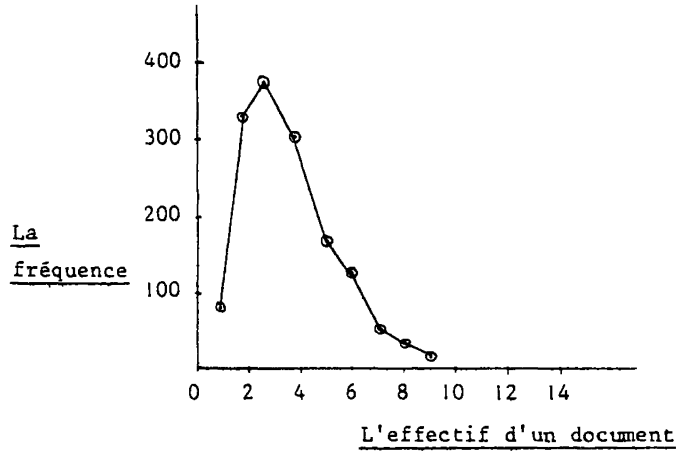
- d'abord, en ce qui concerne les données employées, on a constaté une courbe de fréquence des effectifs qui était presque la même pour tous les échantillons examinés ; on a appelé cette courbe de fréquence la "courbe caractéristique" des données (cf. fig. 1).

- Ensuite, quand les mots-clés étaient assignés aux documents, on a utilisé une liste de référence. Cette liste gouverne l'application des mots-clés aux documents et une liste différente impliquerait une autre matrice d'incidence. On conclut que la "courbe caractéristique" résulte de cette liste, appliquée au rassemblement documentaire particulier.

La fréquence des nombres de mots-clés qui s'associent à chaque document (pour un échantillon de 4 bandes) :

La bande	1	2	3	4	5	6	7	8	9	10	11	12	13	14	> 15
1	93	358	391	275	186	121	47	29	15	8	2	4	1	0	1
2	117	408	373	311	147	102	39	30	8	9	3	3	0	1	0
3	69	290	402	301	150	167	71	39	17	22	1	5	1	2	0
4	58	296	375	320	200	165	65	60	19	26	10	7	2	3	3
Moyenne	84	338	385	302	171	139	56	40	15	16	4	5	1	2	1

Exemple ; Sur la première bande, il arrive 93 fois qu'un seul mot-clé s'associe à un document ; c'est-à-dire qu'un document a un effectir d'un seul mot-clé.



La courbe de fréquence
(moyenne pour 4 bandes)
des effectifs des documents

FIGURE 1

LA "COURBE CARACTERISTIQUE" POUR 6228 DOCUMENTS DU RASSEMBLEMENT "IFIS".

D'une façon générale, donc, la liste de référence -quelque aveugle que ce soit- définit la correspondance entre les caractères (ou les mots-clés) et les individus (ou les documents) et on s'intéresse dans la courbe caractéristique dans la mesure où elle fournit une représentation de cette correspondance.

La deuxième étape de notre démarche consiste à définir une "classification aléatoire". On suppose comme modèle probabiliste que la matrice d'incidence est une matrice binaire aléatoire de loi uniforme parmi toutes les matrices qui sont compatibles avec la courbe caractéristique. Cette dernière n'est autre que les totaux des lignes (i.e. les individus ou les documents). Ce modèle revient à supposer qu'un document d'un effectif quelconque est une partie aléatoire de mots-clés, de loi uniforme parmi toutes les parties de son effectif.

Finalement, la troisième étape de la démarche est la comparaison entre la théorie et la pratique. Quant à celle-ci, on se situe dans le cadre assez général de la classification ascendante hiérarchique. On s'intéresse à un niveau quelconque de la hiérarchie et on le représente comme un graphe simple. La théorie probabiliste précédente fournit la probabilité d'une arête dans ce graphe. On n'aborde pas ici le problème de la variabilité aux divers niveaux de la hiérarchie : on considère seulement un niveau donné.

Le but de toute cette approche peut se résumer comme suit : fournir une mesure de classifiabilité des documents (individus) dans le but d'effectuer des comparaisons ultérieures entre des rassemblements documentaires différents. On met cette démarche en oeuvre par le moyen du même modèle posé pour les rassemblements ; du niveau déterminé de la classification hiérarchique ; et on examine les graphes simples associés.

LES IDEES DE BASE

On considère un ensemble fini de documents (ou plus généralement des individus), I, Une hiérarchie, H, de parties (h, h', etc.) de I,

$$H \subset 2^I ,$$

est définie par les conditions suivantes :

(1) $\phi \notin H$. (H ne contient pas la partie vide).
(2) $\forall i \in I : \{i\} \notin H$. (Les parties formées d'un seul élément ne sont pas comprises et il en résulte que H n'est pas une hiérarchie totale).
(3) $I \in H$. (La partie de I constituée par I lui-même appartient à H).
(4) $\forall h, h' \in H$, on a $h \cap h' \in \{\phi, h, h'\}$. (Deux parties non comparables ne se coupent pas et il en résulte que l'on ne considère pas des classes empiétantes).

On définit un indice de partie comme une application de H dans R^+ (les réels non négatifs),

$$x : h \to R^+ \quad \text{ou} \quad x(h) \in R^+ ,$$

satisfaisant la condition suivante :

(5) $\forall\, h, h' \in H$: si $h' \subset h$ alors $x(h') > x(h)$. (L'indice est strictement décroissant).

Dans la pratique, on définit les indices de parties à partir des similarités, qui sont définies pour chaque paire d'individus. Si Card I = n, on vérifie bien que l'on a :

Card $S \leqslant n(n-1)/2$ où $S = \{s(i,i') \mid i \text{ et } i' \in I\}$, l'ensemble des similarités ; et

où Card dénote la cardinalité ;

Card $X \leqslant n - 1$ où $X = \{x(h) \mid h \in H\}$, l'ensemble des indices de partie.

Les diverses méthodes pour édifier une hiérarchie, H, se fondent sur des définitions différentes de X à partir de S. On peut définir une telle méthode ou stratégie de la classification par l'application croissante suivante :

$$C : S \to X .$$

On a donc :

$$C(s_1) = x_1 ,$$
$$C(s_2) = x_2 ,$$
et si $s_1 > s_2$,
alors $x_1 > x_2$.

On remarque que l'on se borne à considérer les hiérarchies qui sont édifiées en respectant seulement les propriétés ordinales des similarités.

Etant donné une similarité quelconque (ou de même un indice de partie quelconque) on se donne un niveau de la hiérarchie. Une représentation claire et nette, et commune à toutes les méthodes de la classification, est celle d'un graphe simple, $G_{s^*}(n,e)$, où :

s^* est la similarité (le nombre réel) qui détermine le niveau ;

n est le nombre de noeuds, Card I ;

e est l'ensemble des arêtes qui se définit comme suit : il existe une arête entre les noeuds i et i' (qui s'associent aux individus i et i') si $s(i,i') \geqslant s^*$.

Autrement dit, le graphe fournit une représentation d'une partie de l'ensemble ordonné des similarités ; cette partie-là qui se trouve supérieure ou égale au seuil, s^* .

Dans ce cadre on va poser le modèle suivant :

$$\text{Prob } ((i,i') \in e) = p = \text{constante},$$
$$\text{où } G_{s^*}(n,e) \text{ est le graphe du niveau déterminé par } s^* ,$$

c'est-à-dire la probabilité qu'une arête quelconque appartient à l'ensemble des arêtes e est une constante qu'il faut préciser.

Ensuite la question qui nous intéresse est de trouver la probabilité :

$$\text{Prob } (s_r \geqslant s^* \text{ et } s_{r+1} < s^*)$$

où s_r est la r-ième similarité dans l'ensemble ordonné des similarités. Cette probabilité s'écrit aussi comme suit :

$$\text{Prob (Card e = r)} ,$$

c'est-à-dire la probabilité que le nombre d'arêtes égale r .

Donc la prochaine étape est de préciser

$$\text{Prob } ((i,i') \in e) = p$$

et alors de montrer comment on trouve d'une façon générale et commode

$$\text{Prob (Card e = r)}$$
$$= B\left(\binom{n}{2}, p ; r\right) \text{ qui dénote la loi binomiale et qui est définie par :}$$

$$\binom{\binom{n}{2}}{r} p^r (1-p)^{\binom{n}{2}-r} .$$

Avant de traiter ce problème à fond, notons que pour interpréter les résultats des méthodes de classification hiérarchique on procède comme suit :

$$\text{muni de} \quad \text{Prob } ((i,i') \in G_{s^*}(n,e))$$
$$\text{ou} \quad \text{Prob } (s(i,i') \geqslant s^*) ,$$
$$\text{estimer} \quad \text{Prob } (C(s(i,i') \geqslant s^*)$$
$$\text{ou} \quad \text{Prob } (x(h) \geqslant s^*) .$$

Lorsque la relation entre S et X (c'est-à-dire la fonction C) est compliquée, on peut se poser des questions supplémentaires : des estimations de la grosseur de la partie, h ; de son taux d'accroissement pour les valeurs différentes de s^* ; etc. C'est dans le cadre de ces études que travaille Ling (1973, 1975) par exemple.

L'ENUMERATION

Dans le but de faire une comparaison entre les résultats d'une classification automatique obtenue dans la pratique, et les résultats aléatoires rapportés à la courbe caractéristique des données, il faut indiquer comment on trouve ces derniers. C'est-à-dire trouver

$$\text{Prob}\ (s_{ii'} \geqslant s^*) \ .$$

Nous procédons comme suit :

- on suppose que les effectifs des individus sont donnés ; c'est-à-dire on suppose que ce que l'on appelle la courbe caractéristique est donnée ;

- à partir de cela, on suppose que l'individu i est une partie aléatoire de loi uniforme parmi les $\binom{\text{Card C}}{\text{Card i}}$ parties d'effectif Card i (on a noté par C l'ensemble des caractéristiques) ;

- après avoir pris une classe générale de fonctions de similarité, on trouve que ces fonctions réduisent les combinaisons possibles pour arriver au but ; ce qui aide à une énumération de toutes les combinaisons.

Précisons ce dernier fait. Soit un indice de similarité défini par la condition suivante :

s est une fonction de Card i, Card i' et de a ,
où a est le nombre de mots-clés partagés par i et à i'.

Ensuite la condition suivante entraîne que $s_{ii'} \geqslant s^*$:

SIMO : $2a \leqslant \text{Card } i + \text{Card } i' \leqslant (1 + g(s^*))a$
où $g(s^*) \geqslant 1$ et a, Card i et Card i' sont des nombres entiers.
g est une fonction de s^* . Prenons par exemple l'indice de similarité de Jaccard.
On a la relation suivante :

$$1 \geqslant \frac{a}{\text{Card } i \ + \ \text{Card } i' \ - a} \geqslant s^*$$

d'où l'on trouve :

$$2a \leqslant \text{Card } i + \text{Card } i' \leqslant (1 + \frac{1}{s^*})\, a \ .$$

On met $\frac{1}{s^*} = g(s^*)$. On trouve de même que $g(s^*) = 2/s^* - 1$ nous donne l'indice de Dice-Sorensøn ; et que $g(s^*) = 2 + 1/s^*$ donne un indice de Sokal et Sneath.

A partir de la condition SIM0, on note les théorèmes suivants (dont les démonstrations sont sans détours ; elles se trouvent dans Murtagh, 1979).

Théorème 1 :

SIM0 $\Rightarrow$ SIM1 et SIM2 ,

où SIM1 est la condition suivante :

SIM1 : $a \leqslant \text{Card } i \leqslant \lfloor a\ g(s^*) \rfloor$

et SIM2 est comme suit :

SIM2 : $a \leqslant \text{Card } i' \leqslant \lfloor a\ g(s^*) - \text{Card } i + a \rfloor$

où $\lfloor\ .\ \rfloor$ signifie le nombre entier inférieur ou égal au nombre donné.

La preuve de ce théorème est facile quand on note que la valeur minima de Card i et de Card i' est donnée par a ; et que les valeurs maxima et minima de Card i sont les valeurs minima et maxima, respectivement, de Card i' .

Pour une énumération de toutes les combinaisons, les résultats suivants pourraient être utiles ; les preuves découlent d'un examen simple des combinaisons.

Théorème 2 :

Soient deux documents avec a mots-clés communs à tous les deux ; il existe

$$\sum_{r=1}^{b} r$$

combinaisons possibles des valeurs de Card i et Card i' pour satisfaire à SIM0 ;
b dénote $\lfloor a\ g(s^*) - a + 1 \rfloor$.

Théorème 3 :

Soient deux documents de cardinalité Card i et Card i' ; pour Card i $\leqslant$ Card i' , il existe

$$\left\lfloor \frac{(\text{Card } i)(g(s^*))}{1 + g(s^*)} \right\rfloor + 1$$

valeurs possibles de a pour satisfaire à SIM0 .

Théorème 4 :

Si SIM0 est satisfait par deux documents qui partagent a mots-clés, alors (a-1) de ces mots-clés permettent également la satisfaction de SIM0 si
$a \leqslant \text{Card } i \leqslant \lfloor (a-1)\ g(s^*) - 1 \rfloor$, et
$a \leqslant \text{Card } i' \leqslant \lfloor (a-1)\ g(s^*) - \text{Card } i + a - 1 \rfloor$.

Ces théorèmes ont pour but de rendre visible les combinaisons indépendantes, soit de Card i et Card i' à partir de a, soit de a à partir de Card i et Card i' ; pour les deux cas, les combinaisons que l'on désire doivent satisfaire à SIMO. A titre d'exemple, on montre dans la figure 2 les premiers couples de Card i et Card i' ; les quatre valeurs initiales de a ; tout pour une valeur de $g(s^*) = 2$.

Signalons maintenant le résultat suivant qui constitue le résultat principal de l'énumération.

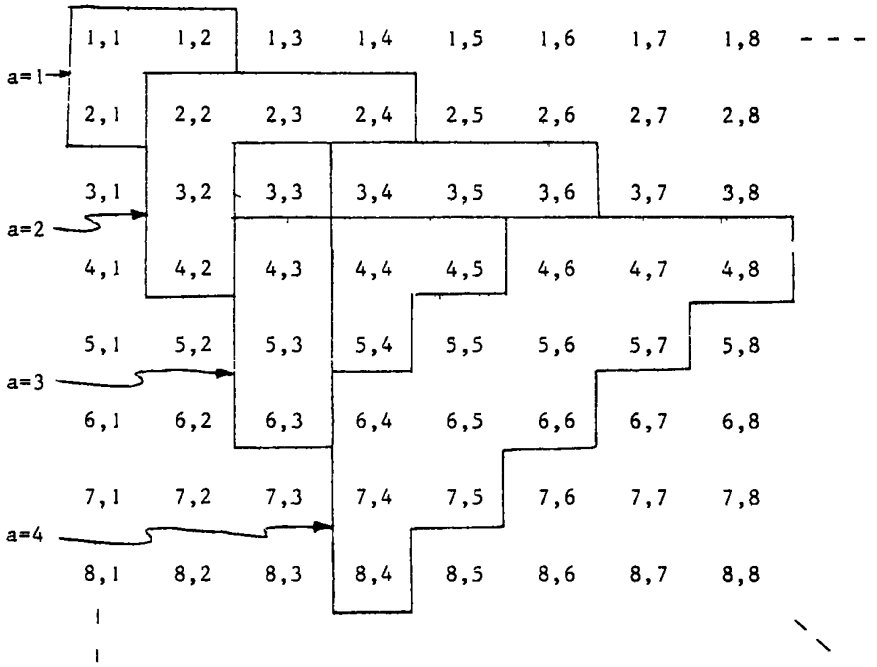
Théorème 5 :

La probabilité que la similarité de deux documents, i et i' , ($s_{ii'}$) partageant a mots-clés à la fois, soit supérieure ou égale au seuil (s^*) est donnée par :

PROI :
$$P(a) = \binom{M}{a} \sum_{\substack{\text{Card } i \\ = a}}^{v} \sum_{\substack{\text{Card } i' \\ = a}}^{w} B(\text{Card } i, 1/M ; a)\ B(\text{Card } i', 1/M ; a)$$
$$\phi(\text{Card } i)\ \phi(\text{Card } i')$$

où M est le nombre total de mots-clés

$\binom{M}{a}$ est le coefficient binomial usuel ;



Deux documents avec Card i, Card i' mots-clés ; pour qu'ils satisfassent à SIMO il faut qu'ils partagent a mots-clés : $g(s^*) = 2$.

FIGURE 2 : SOUS-ENSEMBLES DE L'ESPACE (CARD i, CARD i') POUR a = 1, 2, 3 & 4

v dénote $\lfloor a\ g(s^*) \rfloor$;
w dénote $\lfloor a\ g(s^*) - \text{Card } i + a \rfloor$;
B est la distribution binomiale : B(Card i, 1/M ; a) est la probabilité que a mots-clés particuliers se trouvent au sein de Card i mots-clés ; chacun avec une probabilité de 1/M ;

ϕ est la fonction de répartition des effectifs des documents et ϕ est fourni par la courbe caractéristique : $\phi(x) = \text{Prob (Card i} = x)$, ce que l'on a noté ci-dessus par ϕ (Card i) (par un léger abus de notation).

Preuve :

Pour qu'un document possède a mots-clés particuliers, on trouve une probabilité de B(Card i, 1/M ; a). Idem pour le second avec une probabilité indépendante. On désire les mêmes mots-clés -ce qui veut dire qu'il se trouve $\binom{M}{a}$ possibilités pour ces a mots-clés. Reste les combinaisons valides pour Card i et Card i' -la courbe caractéristique (ϕ) fournit les probabilités ; et SIM1 et SIM2 (théorème 1) fournissent les limites des sommes.

Ce théorème alors donne la probabilité d'une arête dans le graphe aléatoire $G_{s^*}(n,e)$ -c'est-à-dire la probabilité que $s_{ii'} \geqslant s^*$ pour i,i' quelconques- à partir d'un procédé commode pour traiter l'indice de similarité : on réduit ce procédé à un calcul pour les mots-clés communs à deux documents. Donc, pour deux documents ou noeuds du graphe, la probabilité qu'il existe une arête est donnée par la série :

PRO2 : $P_{total} = P(a=1) + P(a=2, a \not< 2) + P(a=3, a \not< 3) + \ldots$

(On remarque que c'est pour trouver tous les évènements indépendants -par exemple, a = 2 mais a $\not<$ 2 à la fois- que l'on a formulé théorème 4).

Il reste un résultat secondaire mais tout de même important : pour démontrer que la série ci-dessus (PRO2) converge, souvent très vite. On vérifie d'abord les lemmes suivants.

Lemme 1 : $\binom{b}{a} \leqslant b^a/a!$

La démonstration provient de la définition de $\binom{b}{a} = \frac{b!}{a!(b-a)!}$ et la division de b! par (b-a)! .

Lemme 2 : $a! > a^a\ e^{-a}\ \sqrt{2\pi a}$

Il s'agit ici de l'approximation bien connue de Stirling.

Lemme 3 :

La borne maxima de $\binom{\text{Card } i}{a}\binom{\text{Card } i'}{a}$ se donne quand

$$\text{Card } i = \text{Card } i' = \frac{a\, g(s^*) + a}{2}$$

En vertu de la symétrie des valeurs de i et de i' et en utilisant le développement du terme initial , on trouve le terme général comme

$$\frac{1}{a!}\ \frac{1}{a!}\ (\text{Card } i - m)\ (\text{Card } i' - m)$$

où $0 \leqslant m \leqslant a-1$.

En employant SIM2 et en posant la dérivé par rapport à Card i (disons) égale à zéro, on trouve la valeur maxima comme ci-dessus.

Abordons maintenant le théorème.

Théorème 6 :

P(a) approche zéro pour a suffisamment grand ; autrement dit, la série définie par PRO2 décroît de façon monotone.

Preuve :

Pour une valeur de a quelconque, et pour satisfaire à SIM0, il existe
$(a\, g(s^*) - a + 1)(a\, g(s^*) - a + 2)/2$
combinaisons, au plus, de Card i et Card i' en vertu du théorème 2.
Pour aucune combinaison entre elles, on a l'expression suivante :

$$\binom{M}{a}\ B(\text{Card } i, 1/M\ ;\ a)\ \ B(\text{Card } i', 1/M\ ;\ a)\ \ \phi(\text{Card } i)\ \ \phi(\text{Card } i')$$

en vertu du théorème 5.
Les deux derniers termes sont toujours inférieurs à 1 .

En utilisant le développement de la binomiale, c'est-à-dire

$$B(\text{Card } i, 1/M\ ;\ a) = \binom{\text{Card } i}{a}\ \left(\frac{1}{M}\right)^a\ \left(1 - \frac{1}{M}\right)^{\text{Card } i - a}$$

et en se servant des lemmes, on trouve comme limite supérieure d'une combinaison quelconque :

$$\left(\frac{.25\ (g(s^*) + 1)^2\ e^3\ M}{(M-1)^2\ a}\right)^a \qquad a^{-3/2} \qquad (2\pi)^{-3/2} \quad .$$

On voit alors, en multipliant ce terme par le nombre maximum des combinaisons, que P(a) décroît d'une façon monotone et que $P(a) \to 0$ quand $a \to \infty$.
(D'autant plus vite que M est grand).

En ce qui concerne les données que l'on a utilisées, on a trouvé des probabilités négligeables quand $a > 2$. En utilisant l'équation PRO2, donc, on a trouvé la probabilité d'une arête quelconque dans le graphe qui nous intéresse. On a utilisé l'indice de similarité de Jaccard et un niveau de la hiérarchie du saut minimum qui, tous les deux, ont indiqué que $g(s^*) = 2$. S'il s'agit de N documents on aura affaire à $\binom{N}{2}$ arêtes possibles dans le graphe, chacune munie de la probabilité que fournit l'équation PRO2. (En fait, on s'est servi des sous-ensembles de 75 documents). On utilise alors la distribution binomiale pour trouver la moyenne ou d'autres statistiques afin de comparer aux résultats trouvés dans la pratique. Enfin on a trouvé que les résultats pratiques se sont écartés de plus de 20 écarts-types de la moyenne des arêtes, cette dernière étant ce que l'on attend après avoir supposé le modèle probabiliste. Comme nous l'avons dit dans l'introduction, ce n'est pas le problème de la classification. C'est pour cette raison que nous donnons le résultat comme ci-dessus. Il s'agit d'une mesure de la classification quand on se donne un niveau de la hiérarchie.

CONCLUSION

A partir de ce que l'on appelle la courbe caractéristique des documents, on a montré comment les résultats d'une classification automatique peuvent se rapporter à une "classification aléatoire". Il s'agit d'abord de trouver empiriquement la courbe caractéristique. A partir de là, on pose le modèle probabiliste (qui propose qu'un document comporte le même nombre de mots-clés que précédemment mais que ces mots-clés soient tirés au hasard dans la population des mots-clés). Enfin on compare les résultats obtenus dans la pratique à ceux indiqués par cette démarche theorique. Un encadrement commode pour traiter cette dernière étape est de se situer dans le cadre des graphes simples. Quant aux résultats obtenus dans la pratique, on considère un niveau quelconque de la hiérarchie ; et la démarche probabiliste que l'on a esquissée fournit la distribution des nombres d'arêtes dans le graphe aléatoire associé à ce niveau.

Comme but ultérieur de cette démarche, on envisage en premier lieu des comparaisons entre divers exemples de mots-clés (caractères), tout en se servant du même ensemble de documents (individus). On peut chercher de cette façon l'ensemble de mots-clés dont résulte la meilleure classification. Autrement dit, on peut comparer des diverses listes de référence. Dans la pratique de systèmes documentaires, la liste de référence doit être régulièrement mise à jour et nous espérons que notre démarche soit utile à cette fin.

Dans cet article on a cherché à mettre en oeuvre cette démarche. Outre son extension aux autres indices de similarité (on en a considéré une classe) et aux données non binaires (qualitatives) on peut se poser les questions suivantes :

- en ce qui concerne les classifications hiérarchiques, est-ce qu'un ou quelques seuils (niveaux) suffisent pour faire l'analyse probabiliste ? Ou faut-il plutôt considérer le taux de l'agrégation (c'est-à-dire les taux d'accroissement des parties) ?

- la question la plus importante est d'évaluer l'utilité de cette approche dans le domaine de l'interrogation des systèmes documentaires et, plus généralement, dans tous les domaines auxquels s'applique la classification automatique.

BIBLIOGRAPHIE

The IFIS Magnetic Tape Manual, Version 2. Zentralstelle für Maschinelle Dokumentation : Frankfurt-am-Main, 1972.

Jones K.S. Collection properties influencing automatic term classification performance. Information Storage and Retrieval, Vol. 9, 1973, 499-513.

Ling R.F. A probability theory of cluster analysis. Journal of the American Statistical Association, Vol. 68, 1973, 159-164.

Ling R.F. An exact probability distribution on the connectivity of random graphs. Journal of Matnematical Psychology, Vol. 12, 1975, 90-98.

Murtagh F. Cluster analysis and bibliographie information retrieval. Thèse de MSc, Department of Computer Science, Trinity College Dublin, 1979.

Salton G. et Wong A. Generation and search of clustered files. ACM Transactions on Database Systems, Vol. 3, 1978, 321-346.