UNE APPROCHE NUMERIQUE / SYMBOLIQUE POUR L'EXTRACTION ET LA FORMALISATION DE CONNAISSANCES : APPLICATION A LA DESCRIPTION DE TACTIQUES DE PECHE ARTISANALE AU SENEGAL

Huyen Tong, Emmanuel Périnel [a]

I- INTRODUCTION

Ce travail s'inscrit dans le cadre d'un projet développé par l'ORSTOM intitulé "Interaction entre Analyse de Données, Intelligence Artificielle et Modélisation Mathématique pour la simulation de la pêche artisanale au Sénégal". Afin de mieux cerner la dynamique de ce type de pêcherie, plusieurs travaux ont été développés, en particulier à travers l'étude du comportement du pêcheur interagissant avec son environnement [Ferraris, Le Fur, 93]. Dans ce contexte, la notion de *tactique de pêche* [Ferraris, Samba 91] a été définie comme la combinaison de différents choix effectués par le pêcheur durant son activité, tels que : celui des espèces ciblées, du lieu de pêche, de la taille de l'équipage ou encore du type d'engin utilisé.

Dans ce travail, nous présentons une méthodologie de nature numérique / symbolique permettant de représenter et de formaliser la notion de tactique de pêche (figure n°1). A partir d'un traitement initial identique (constitution de groupes homogènes de comportements de pêche par une technique de classification automatique), une double approche est proposée afin de fournir une description explicite de chacun des groupes :

- la première est basée sur un *algorithme d'apprentissage,* dit supervisé, issu du domaine de l'Intelligence Artificielle (I.A.);
- la seconde approche consiste à utiliser une *méthode de segmentation* construisant un arbre de décision binaire.

(a) Laboratoire LISE CEREMADE, Université Paris-IX Dauphine

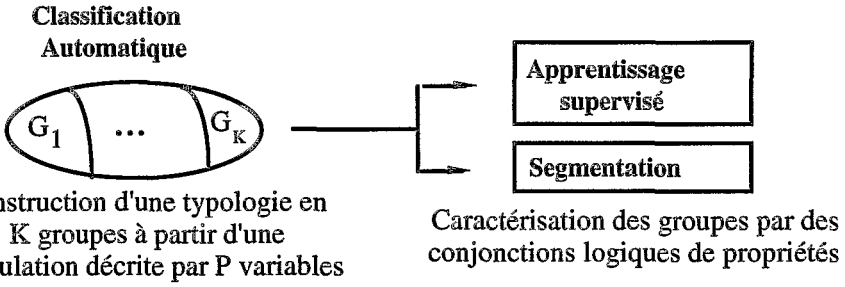

Figure n°1. Une double approche "numérique / symbolique".

Dans le cadre du projet MOPA [Le Fur 90], différents résultats ont déjà été obtenus par la première approche [Tong 94], [Périnel 91].

II- POSITION DU PROBLEME

Les méthodes de classification automatique possèdent une caractéristique essentielle : les groupes sont obtenus en agrégeant des individus présentant de fortes ressemblances, celles-ci étant évaluées *sur l'ensemble des attributs*. On qualifie parfois les groupes obtenus par ces techniques de *polythétiques* pour traduire le fait qu'il n'est pas possible de les caractériser par un ensemble de propriétés vérifiées par la totalité des individus du groupe (caractérisation *monothétique*).

L'interprétation fournie par les logiciels standards d'analyse de données qui en résulte est parfois délicate dans la mesure où il est difficile de retrouver comment sont combinées entre elles les différentes modalités caractéristiques de la classe. En d'autre termes, le problème peut être formulé comme suit : comment mettre en évidence au sein d'une classe de nature polythétique des liaisons ou dépendances logiques entre modalités - définissant des groupes de type monothétique - afin de fournir une description plus explicite de l'ensemble ?

III- APPROCHE PAR APPRENTISAGE SUPERVISE

La méthode présentée est une adaptation de l'algorithme *d'apprentissage supervisé* CABRO [Ho *et al.* 91], [Gettler *et al.* 93], [Morineau *et al.* 95]. Cette méthode, issue de l'I.A., vise à trouver des règles d'appartenance à une classe à partir d'un ensemble d'observations.

On dispose initialement d'une population partitionnée en deux groupes respectivement appelés *exemples* (*E*) et *contre-exemples* (*CE*). L'objectif est de fournir une description du goupe *E*. Le principe consiste à rechercher de manière itérative des sous-populations de type monothétique ; ceci jusqu'à avoir recouvert l'ensemble des éléments de *E*. La recherche de la première sous-population est effectuée comme suit :

1. De manière naturelle, on essaie de construire en priorité une description satisfaite par le plus grand nombre d'éléments de *E*. Pour cela, on sélectionne l'individu w* "le plus caractéristique" de *E*, i.e. celui dont les modalités composant sa description possèdent globalement un meilleur score au sens de la *valeur-test* [Morineau *et al.* 95] (indicateur statistique de significativité fournie par le logiciel SPAD.N au niveau de l'interprétation des classes).

2. On retient les *K* meilleures modalités m_k décrivant w*. Chacune d'elle va constituer un point de départ possible pour former la première meilleure caractérisation de *E*.

3. On évalue tout d'abord la qualité (ou pouvoir discriminant) d'une modalité m_k de la façon suivante : chaque modalité m_k définit une propriété qui est satisfaite par un sous-ensemble d'individus de *E*, appelé extension de m_k dans *E* et noté $Ext_E(m_k)$, ainsi que par un sous-ensemble d'individus de *CE* noté $Ext_{CE}(m_k)$; on calcule alors le rapport suivant :

$$R = \frac{\text{card}\,(\text{Ext}_E(m_k))}{\text{card}\,(\text{Ext}_E(m_k)) + \text{card}\,(\text{Ext}_{CE}(m_k))}$$

ce rapport est d'autant plus élevé que la proportion d'individus de *E* possédant la propriété m_k dans la population totale est grande.

4. Si *R* ne dépasse pas un seuil donné *a* (choisi par l'utilisateur), on cherche alors une seconde modalité m_j parmi les *K-1* restantes de telle sorte que la conjonction $m_j \wedge m_j$ maximise le rapport *R*. On réitère ainsi la phase **4** tant que $R<a$.

5. Les étapes **3** et **4** sont effectuées pour chacune des *K* modalités m_k et on retient parmi les *K* conjonctions obtenues celle, notée a^* qui permet d'obtenir le plus grand recouvrement de *E* (i.e. celle dont l'extension sur *E* est la plus importante).

6. On retire de l'ensemble *E* les éléments vérifiant la conjonction a^* et on réitère les étapes **1** à **5** jusqu'à avoir entièrement recouvert l'ensemble des exemples.

Exemple : A partir de cinq groupes homogènes de comportements de pêche obtenus par classification automatique pour les données de Kayar, l'application de cet algorithme a permis d'extraire comme caractérisation du quatrième groupe les deux descriptions suivantes :

a_1 = [*cible 1* = thiof] Λ [*engin* = PML]

a_2 = [*cible 1* = thiof] Λ [*cible 2* = aucune] [*équipage* = 4]

La tactique de pêche associée au groupe 4 s'exprime ainsi à travers deux conjonctions logiques de propriétés ; elles sont formalisées ici dans le langage des *objets symboliques* de type booléens [Diday 92].

IV- APPROCHE PAR SEGMENTATION

Les méthodes de segmentation (par exemple [Gueguen, Nakache 1988]) constituent une alternative intéressante et une approche différente pour fournir une interprétation explicite des classes d'une partition. Elles se distinguent cependant des méthodes d'apprentissage telles que CABRO car elles permettent dans un même temps "d'organiser la connaissance" sous la forme d'un arbre binaire.

Le principe général d'une segmentation est le suivant : on procède à des dichotomies itératives de la population de manière à obtenir des sous-populations qui soient le plus homogènes possible vis-à-vis des groupes à caractériser. Les coupures ou dichotomies sont obtenues à partir de questions binaires portant sur les différents descripteurs ; par exemple, la première coupure de l'arbre de la figure n°2 affecte aux nœuds gauche et droit les individus ayant respectivement ciblé les espèces {48 ou ... ou 116} et {90 ou 125}. Par ailleurs, à chaque étape, la meilleure coupure est celle qui optimise un critère évaluant la discrimination des groupes au sein des deux nœuds créés (critère d'information, de Kolmogorov-Smirnov, ...).

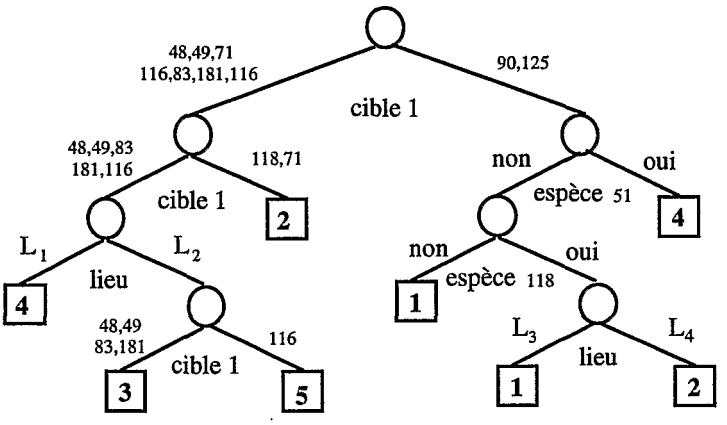

Figure n°2. Description des groupes par arbre de segmentation binaire

Chaque chemin de l'arbre (de la racine à une feuille) est une conjonction de propriétés décrivant une sous-population ; celle-ci est étiquetée par le groupe représenté majoritairement dans la feuille. On parle de nœud terminal "pur" lorsque la totalité des individus de la feuille sont issus du même groupe. Dans le cas général, il subsiste cependant un certain mélange des groupes G_k qui peut être représenté et formalisé au sein d'une assertion de type probabiliste [Diday 92]. On décrira par exemple la feuille à l'extrême gauche de l'arbre, étiquetée par 4, par

$$[\ \textit{cible 1} \in \{48,49,83,181,116\}\ ] \wedge [\ \textit{lieu} \in L_1\ ]$$
$$\wedge\ [\ \textit{groupe} = \textit{0.97}\ (G_4),\ \textit{0.03}\ (G_5)\ ]$$

pour traduire le fait que 97% des individus de la feuille proviennent du groupe 4 mais que 3% sont des individus du groupe 5.

V- DISCUSSION

A partir d'un traitement statistique multidimensionnel (numérique), nous avons proposé deux approches, qualifiées de symboliques permettant de construire des descriptions explicites des classes d'une partition d'objets. En effet, dans la deuxième phase, à la différence des traitements "classiques" de l'analyse des données (classifications, analyses factorielles), les synthèses d'information sont obtenues sur la base d'une propriété couramment utilisée en

I.A., la notion *d'héritage* ; celle-ci consiste à caractériser des groupes en allant *du plus général au plus spécifique*.

La structure parfois complexe des informations extraites par ces techniques (prise en compte de valeurs multiples, fréquences pondérant les modalités) est clairement exprimée à travers le formalisme des objets symboliqes (booléens ou probabilistes). De manière plus générale, cette approche peut se plonger dans le cadre théorique très général de l'analyse des données symboliques dont l'objectif est d'étendre les traitement usuels de l'analyse des données à des objets plus complexes tels que les objets tactique de pêche construits dans le cadre de cette application.

BIBLIOGRAPHIE

DIDAY (E.) 1992. Probabilist, possibilist and belief objects for a knowledge analysis. Cahiers du ceremade, Université Paris ix-Dauphine.

FERRARIS (J.) 1994 Compte-rendu 90L0709 MRT - opération de recherche, février 94.

FERRARIS (J.), LE FUR (J.) 1993. Méthodes d'analyse et de représentation d'un système d'exploitation : synergies et redondances. *In "Les recherches françaises en évolution quantitative et modélisation des ressources et des systèmes halieutiques"* Gascuel (D), Durand (JL) et Fonteneau (A), Ed. Colloques et Séminaires ORSTOM. pp. 347-373.

FERRARIS (J.), SAMBA (A.) 1991. Variabilité de la pêche artisanale sénégalaise et statistique exploratoire. *In "SEMINFOR 5 statistique impliquée"* Laloë (F), Ed. Colloque et Séminaire ORSTOM. pp. 169-190.

GETTLER-SUMMA (M.), PERINEL (E.), FERRARIS (J.) 1994. Automatic aid to symbolic cluster interpretation. *In New approaches in classification and data analysis*. Springer Verlag, E. Diday et al. editors. pp. 405-413

GUEGUEN (A.), NAKACHE (J.P.). Méthode de discrimination basée sur la construction d'un arbre de décision binaire. *R.S.A.* vol. 36, 1, pp. 19-38.

HO TU BAO, DIDAY (E.), GETTLER-SUMMA (M.) 1988. Generating rules for expert system from observations. *Pattern Recognition Letters* 7, pp. 265-271.

LE FUR (J.) 1990. Projet MOPA : Modélisation de la pêche artisanale au Sénégal. Document Multig. Orstom, 27 p.

MORINEAU (A.), GETTLER-SUMMA (M.), TONG (H.) 1995. Marquage sémantique des axes et des classes. xxviièmes journées de l'asu. Jouy-en-Josas 15-19 mai 1995.

PERINEL (E.) 1991. Analyse numérique-symbolique des tactiques de pêche artisanale au Sénégal. Mémoire de DEA, Université Paris ix-Dauphine. 74 p.

TONG (H.) 1994. Interprétation symbolique pour l'analyse factorielle la classification et le graphe de matrice de transition d'états. Mémoire de DEA de l'Université Paris ix-Dauphine. 86 p.

ALEVIZOS (P.), Morineau (A.) 1992. Tests et Valeurs-Tests. Application à l'étude des mastics utilisés dans la fabrication des vitraux. *R.S.A.*, vol. 40, n° 4, pp. 27-43.