Classification et identification d'espèces par discrimination à partir de connaissances structurées

Noël Conruyt*— David Grosser* — Jacques Le Renard**

** IREMIA, Université de La Réunion*
97715 St Denis Messag. Cedex 9, LA REUNION

*** MNHN, Lab. de Bio. des Invertébrés Marins et Malacologie, CNRS URA 699*
55, rue de Buffon, 75005 Paris, FRANCE

RESUME. *Dans les Sciences de la vie, les connaissances que l'on veut représenter et traiter sont complexes. En systématique, discipline scientifique dont l'un des buts est de décrire et d'étudier la diversité des êtres vivants, les descriptions d'espèces et de spécimens sont le plus souvent structurées, variables, erronées, etc.. Du point de vue formel, les objets symboliques (objets de synthèse) permettent de représenter ces connaissances sur l'observable (modèle descriptif) et l'observé (instances du modèle). Néanmoins, l'analyse de ces objets symboliques doit être suffisamment « intelligente » pour tenir compte de la structure de ces descriptions. Dans cet article, nous présentons un algorithme de classification et d'identification d'espèces s'appuyant sur la connaissance du modèle descriptif. Une application sur les coraux du genre Pocillopora nous sert d'illustration.*

MOTS-CLES : *systématique, discrimination, connaissances structurées.*

1. Introduction

Dans la plupart des applications en sciences de la vie, les données à traiter sont plus complexes que celles que l'on trouve dans les domaines industriels. Par exemple, les variables ou attributs utilisés sont plus nombreux et de types plus variés (taxonomies, intervalles numériques, nominaux multi-valués exprimant la variation ou l'imprécision, etc.). La prise en compte de la structuration des connaissances en biologie [DAL 80], [ALL 84] est un progrès qui permet de prendre en charge une quantité de connaissances de fond (dites "de bon sens") utiles pour saisir, gérer et traiter les données complexes de manière plus cohérente et efficace.

Nous présentons un algorithme de classification et d'identification de données structurées pré-classifiées fonctionnant sur le principe de la discrimination (arbres de décision). L'algorithme est dirigé par les connaissances de base : il permet à la fois de tenir compte des dépendances entre les caractères du domaine (attributs inapplicables) et donc de gérer la cohérence entre les questions posées à l'utilisateur lors de la consultation, mais aussi d'opérer à une réduction du nombre de caractères à tester pour la mesure du gain d'information en fonction du contexte d'observation (composant observé présent).

2. Les connaissances de base

Afin de tenir compte de cette structuration des connaissances, nous avons introduit la notion de modèle descriptif [LE R 96] avec la mise en place de différentes logiques descriptives : (dé)composition, point de vue, spécialisation, itération (multi-instanciation), conditions contextuelles. Le modèle descriptif constitue le niveau sémantique de la représentation des connaissances (Fig. 1). Il définit tout ce qui est *observable* pour le domaine d'étude. Il se présente sous forme d'un schéma

structuré de tous les objets, attributs et valeurs possibles du domaine (*Pocillopora)*, ce dernier constituant la racine d'un *arbre de description*.

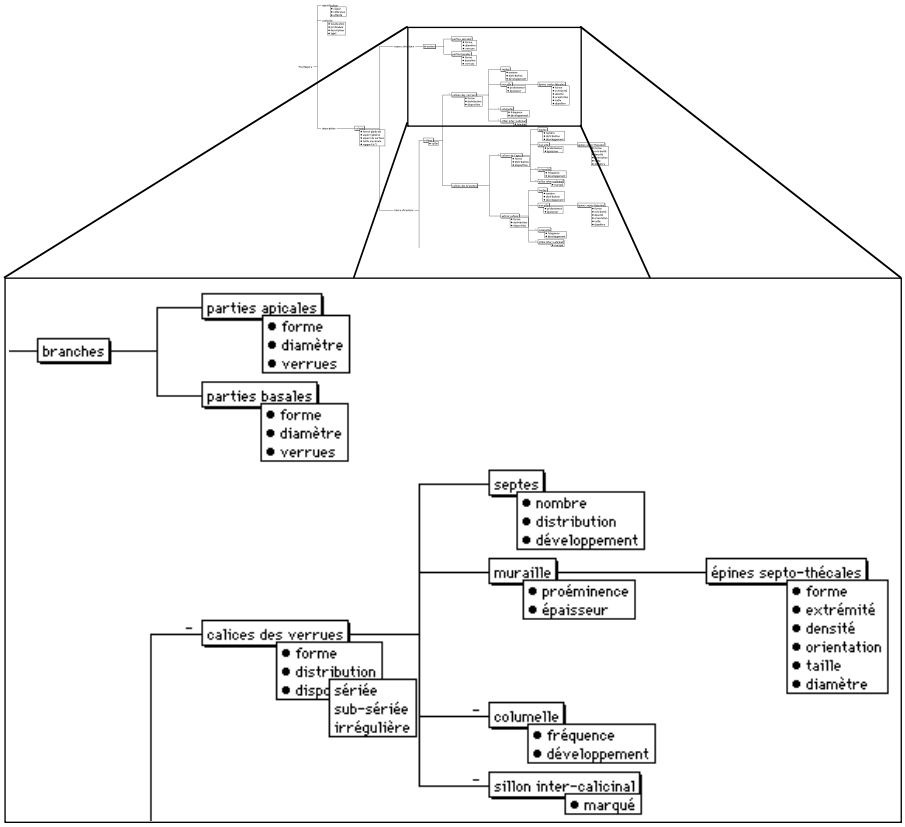

Figure 1. *Une partie du modèle descriptif des Pocillopora*

Les nœuds de l'arbre correspondent aux composants observables. Chacun d'eux est décrit à l'aide de caractères ou attributs typés, pouvant admettre des valeurs symboliques, numériques, uniques, multiples (variables), ordonnées, structurées ou imprécises. Sur la figure 1, on ne visualise que les valeurs nominales de l'attribut distribution des calices des verrues.

Certains composants comme "calices des verrues" peuvent être absents (signalé par un signe négatif). La présence d'autres objets comme les septes, la muraille dépend de la présence du premier : cette connaissance de fond (règle d'inapplicabilité) va être exploitée dans notre algorithme de traitement inductif des données structurées.

2. Les connaissances instanciées

Les descriptions observées sont des instances du modèle descriptif. Le niveau formel de cette représentation est celui des objets symboliques, plus précisément celui des objets de synthèse [DID 93].

3. Les connaissances dérivées

Pour la **classification**, un arbre de décision est construit. A partir des descriptions (représentation en extension) une méthode inductive fondée sur la mesure d'entropie et du gain d'information [SHA 49], [QUI 86] fabrique une caractérisation de ces classes par un ensemble de règles. Chaque chemin depuis la racine vers les feuilles de l'arbre de décision est une règle de classification (également appelée *diagnose*).

Pour **l'identification**, le seul chemin de l'arbre de décision correspondant au cas à identifier est construit. Etant donné un ensemble d'exemples (descriptions pré-classifiées), la méthode extrait dynami-

quement le critère le plus efficace (selon la mesure du gain d'information) à partir d'une liste ordonnée de tests, cela après chaque réponse de l'utilisateur. Les cas compatibles sont retenus en fonction de cette réponse. Si elle est inconnue, le second test le plus discriminant est proposé à l'utilisateur, et ainsi de suite [CON 94].

4. Algorithme de construction d'un arbre de décision

Soient $E = \Omega = \{\omega_1, ..., \omega_n\}$, ens. des exemples observés,
$M = \{N, Y\}$, ens. des composants et des attributs observables,
$N = \{N_1, ..., N_m\}$ ensemble des noms des composants structurés,
$Y = \{Y_1, ..., Y_p\}$ ensemble des attributs dépendants de N.

```
ArbreDécision (E, M)
    Y = SélectionnerAttributs (racine(M))
    ConstruireArbre(E, Y)
fin ArbreDécision

SélectionnerAttributs(r)
    Y' = ∅
    si (absencePossible(r) = "oui") Y' = Y' ∪ {exist(r)}
    sinon
          Y' = Y' ∪ Att(r) // attributs de r
          pour tout n_f ∈ fils(r)
                SélectionnerAttributs (n_f)
          fin pour
    fin si
    retourner Y'
fin SélectionnerAttributs

ConstruireArbre(E, Y)
    si critèreArrêt(E, Y) alors CréerFeuille(E)
    sinon
          A = meilleurTest(E, Y) // A = y(n)
          d_i = construireNoeud(A)
          Y = filtrerAttribut(A) // selon type de A
          partition(d_i) = R(E)
                // R(E) : ∀ ω ∈ E, A(ω) = v_i ⇔ ω ∈ E_i
          pour chaque E_i ∈ partition(d_i)
                créerBranche(v_i)
                si (A = exist(n)) ∧ (v_i = "oui")
                      Y = SélectionnerAttributs(n)
                fin si
                ConstruireArbre(E_i, Y)
          fin pour
    fin si
fin ConstruireArbre
```

La partie originale de l'algorithme est celle de sélection des attributs :

1) L'arbre de description est parcouru en profondeur depuis la racine, composant par composant. Si l'un d'eux *peut* être absent dans le modèle descriptif (cf. symboles négatifs sur la fig. 1), alors on va générer le test sur l'existence de ce composant (réponse oui ou non) et le placer dans la liste des candidats éligibles. Ensuite, on ne va pas parcourir le sous-arbre dépendant de la présence de ce composant de manière à ne disposer que des attributs pertinents pour l'identification. Par exemple, le test "existence (calices des verrues) = oui $\vee$ non" est généré et tous les attributs du sous-arbre ne sont pas sélectionnés dans la liste.

2) Par contre, si le composant est toujours présent (pas de symbole négatif), ses attributs sont rajoutés à la liste des tests candidats pour la

mesure de leur pouvoir de discrimination (les questions posées seront donc toujours cohérentes).

3) Lors de la procédure d'identification, si le test d'existence d'un composant est choisi comme « le meilleur », et que l'utilisateur répond qu'il est *réellement* présent, alors les attributs dépendants de la présence de cet objet sont recherchés dans l'arbre de description selon la procédure n°1 avec comme racine le composant observé par l'utilisateur.

5. Conclusion

L'algorithme proposé illustre la technique utilisée pour exploiter la connaissance du modèle descriptif, en tenant compte des dépendances entre composants. D'autres aspects de la connaissance du domaine, tel que le coût d'observation ou la pertinence des caractères (utilisés par exemple dans MAKEY [VIG 91]) peuvent être utilisés, de même que les taxonomies de valeurs explicitées dans le modèle descriptif. L'arbre de décision ainsi construit est utilisé pour classifier les espèces, ou bien est généré dynamiquement en cours de consultation pour l'identification d'un nouveau cas. Cette procédure de discrimination est originale par rapport à d'autres algorithmes développés comme par exemple dans KATE [MAN 93] : ce dernier travaille à partir d'un tableau de données non structurées et génère dynamiquement les tests d'existence des composants à partir de l'apparition d'une valeur inapplicable pour un attribut de ce composant dans la colonne du tableau. Notre approche est plus intéressante d'un point de vue sémantique, car elle assure que l'inapplicabilité d'une valeur d'un attribut dépend de l'absence possible d'un composant et non l'inverse.

6. Bibliographie

[ALL 84] ALLKIN R., "Handling taxonomic descriptions by computer", In; Allkin R and Bisby FA (eds.), Databases in systematics. Systematics Association London, Academic Press, (26) pp 263-278, 1984.

[CON 94] CONRUYT N., *Amélioration de la robustesse des systèmes d'aide à la description, à la classification et à la détermination des objets biologiques*. Thèse de doctorat en informatique, Univ. Paris-IX Dauphine, pp 1-281, 1994.

[DAL 93] DALLWITZ M.J., PAINE T.A., ZURCHER E.J., "User's guide to the DELTA System. A general system for processing taxonomic descriptions" Canberra: CSIRO, Div. Entomol., 4th ed., 1993.

[DID 93] DIDAY E., "An introduction to Symbolic Data Analysis", Rapport de recherche INRIA, 1993.

[LE R 96] LE RENARD J., LEVI C., CONRUYT N., MANAGO M., "Sur la représentation et le traitement des connaissances descriptives : une application au domaine des éponges du genre *Hyalonema*", vol. 66 suppl., Biologie, Recent advances in sponge biodiversity and documentation, P. Willenz (Ed), Bulletin de l'Institut Royal des Sciences Naturelles de Belgique, pp. 37-48, 1996.

[MAN 93] MANAGO M., ALTHOFF K.D., AURIOL E., TRAPHÖNER R., WESS S., CONRUYT N., MAURER F., "Induction and reasoning from cases", First European workshop on case-based reasoning (EWCBR-93), MM Richter, S Wess, KD Althoff and F Maurer (Eds.), Springer Verlag, (2), 1993.

[QUI 86] QUINLAN J.R., *Induction of decision trees*, Machine Learning 1 : 81-106, 1986.

[SHA 49] SHANNON C.E., *The mathematical theory of communication*, University of Illinois Press, Urbana, 1949.

[VIG 91] VIGNES R., *Caractérisation automatique de groupes biologiques*. Thèse de doctorat, Univ. Paris-VI, pp 1-260, 1991.