Classification par voisinages successifs sur des descriptions morphologiques complexes

David Grosser*, Noël Conruyt*
Henri Ralambondrainy*

*LIM-IREMIA, Université de la Réunion
Parc Technologique Universitaire, Bâtiment 2
2, rue Joseph Wetzell - 97490 Sainte-Clotilde
grosser, conruyt, ralambon@univ-reunion.fr

Résumé. Afin de classifier des descriptions morphologiques issues de bases de connaissances en biologie, nous proposons une méthode de fouille de données incrémentale, interactive et semi-dirigée. Cette méthode est fondée sur la construction itérative du voisinage de la description partielle de l'objet à classer. Nous proposons différents indices de similarité adaptés à la nature complexe des données considérées (multi-valuées, incomplètes et structurées), pour sélectionner les descriptions les plus proches. Les connaissances du domaine sont utilisées aux différentes étapes du processus de classification, notamment pour le choix de variables discriminantes. A partir de la base de connaissances sur les coraux des Mascareignes, une application montre l'intérêt de cette approche.

1 Introduction

Outre l'étude des liens de parenté entre espèces (phylogénie), la description de la morphologie de *spécimens* biologiques à des fins d'identification et de classification est une part essentielle du travail des systématiciens (phénétique). L'automatisation de ce processus par des outils informatiques dans le but de construire des systèmes classificatoires pose d'intéressants problèmes de représentation et de traitement de connaissances. Ceci est particulièrement vrai en biologie marine et pour certains taxons comme les coraux, où la nature polymorphe des individus (les colonies coralliennes) rend difficile leur description par des représentations classiques de type attribut-valeur. Les relations de dépendance entre caractères, induites par une forte variabilité morphologique engendrent notamment des difficultés d'observation et de description.

Pour pallier à ces contraintes et permettre la construction de bases de connaissances sur la biodiversité des milieux marins, une méthodologie d'acquisition des connaissances descriptives adaptée aux besoins particuliers des biologistes, s'appuyant sur *les logiques descriptives en sciences de la vie* a été proposée par Conruyt (1994) et Renard et al. (1996). Cette méthode permet de définir une connaissance d'ordre ontologique sur la nature des taxons et de décrire de manière structurée les spécimens utilisés pour la conception de la Taxonomie des milieux considérés. Elle offre un cadre signifiant pour représenter des objets biologiques complexes,

permettant notamment d'exprimer des relations de dépendance entre composants d'une description. Ces objets complexes sont caractérisés par des attributs de type nominal, ordinal ou continu, ensemble ou intervalle, dont les valeurs peuvent être manquantes, inconnues, variées ou imprécises selon les choix de définition des experts et les difficultés d'observation rencontrées (voir partie 2).

Les *logiques descriptives en sciences de la vie* sont à l'origine du modèle de représentation des connaissances par objets **Codesc** proposé par Grosser et al. (2003) et implanté au sein du système IKBS (*Iterative Knowledge Base System*) développé en Java par Grosser (2002). Ce système de gestion de bases de connaissances (SGBC) a notamment été utilisé pour la conception de la *base de connaissances sur les coraux des Mascareignes* grâce à des experts internationaux du domaine (Faure et al., 1999). Celle-ci rassemble plus de 150 taxons et 800 descriptions de spécimens, soit une représentation d'environ 80% des coraux de cet archipel du Sud-Ouest de l'Océan Indien. L'objectif premier de cette base de connaissances est de capitaliser une partie du savoir des rares experts traitant de la Taxonomie des coraux dans des monographies mais aussi du "savoir-faire" des systématiciens lié à l'observation et à la description de spécimens en collection et sur le terrain. L'objectif second est d'apporter une aide à la classification et à l'identification de nouveaux spécimens. IKBS dispose en effet de plusieurs algorithmes spécialisés adaptés à ce type de représentation, tels que la génération d'arbres d'identification, le calcul de similarités entre individus ou la recherche de voisinages d'une description partielle. Cette base de connaissances est aussi utile pour répondre à des problématiques de suivi de la biodiversité, d'identification d'espèces menacées ou encore de révision taxinomique des groupes zoologiques.

Pour identifier un spécimen et lui associer un nom, les biologistes utilisent traditionnellement des clefs d'identification (ou de détermination [1]). Cette méthode historique offre l'avantage d'être véhiculée sur support écrit et d'être très explicite, la diagnose pouvant être associée à l'identification. Avec l'informatisation des données systématiques et biologiques et l'essor de la taxonomie numérique initiée par Sokal et Sneath (1963), de nombreuses méthodes d'identification et de classification ont été proposées sur des données tabulaires. Parallèlement, en analyse discriminante et en apprentissage, d'autres méthodes de classification par arbre ont également été développées, telles que les arbres de classification (Breiman et al., 1984) ou de décision (Quinlan, 1986). Ces différentes méthodes peuvent être utilisées dans le cadre de la classification de données biologiques complexes, mais ne sont pas toujours satisfaisantes car elles ne prennent en compte ni les relations entre attributs, ni les données manquantes ou imprécises. Plus important encore, ces méthodes n'intègrent pas de connaissances *a priori* sur les domaines considérés, pré-requis indispensable pour construire des systèmes classificatoires, fiables et robustes (Conruyt, 1994). Or, les experts émettent des hypothèses sur l'objet étudié et procèdent généralement en deux phases pour déterminer la classe d'un spécimen. D'abord une phase *synthétique*, par observation globale des caractères les plus visibles permet de réduire le champ d'investigation en sélectionnant des descriptions présentant des similitudes. Puis une phase *analytique*, par l'observation fine de caractères discriminants permet d'affiner la recherche jusqu'à obtention du résultat.

La méthode de classification par voisinages successifs (CVS) développée dans cet article propose une méthode de fouille fondée sur ce type de raisonnement. Elle s'appuie sur la re-

1. succession d'alternatives dichotomiques portant sur les caractères d'un spécimen qui permet d'en déterminer le rang taxinomique.

cherche du voisinage d'une description partiellement renseignée, possédant éventuellement des erreurs. Cette méthode est rendue possible grâce aux différentes mesures de similarité qui tiennent compte de la structure et du contenu des descriptions (voir partie 3). Elle utilise une phase de sélection de variables discriminantes afin de compléter l'information et d'affiner progressivement le processus d'identification. La méthode est à la fois interactive et itérative. La partie 4 propose une implémentation de cette méthode au sein du système IKBS ainsi qu'une expérimentation sur différents groupes taxinomiques issus de la base de connaissances sur les coraux (partie 5).

2 Représentation des données morphologiques

Le modèle de représentation des données morphologiques Codesc propose deux entités de premier ordre pour la conception d'une base de connaissances : le **modèle descriptif** et les **descriptions**.

2.1 Le modèle descriptif

Un modèle descriptif permet de structurer les caractères observables d'un groupe taxinomique donné. Il est composé d'objets (ou composants) ainsi que d'attributs munis d'un domaine de valeurs et de relations. Celles-ci sont par défaut des relations de composition (*part-of*) entre composants ou entre un composant et des propriétés (caractères). Les attributs sont ainsi attachés aux composants qu'ils caractérisent. Par exemple sur la figure 1, le modèle descriptif de la famille des *Fungiidae* comporte cinq attributs (*label*, *descripteur*, *taxon*, etc.). Ces attributs sont attachés au composant *identification*. Dans cette représentation, les composants *identification* et *contexte* définissent les méta-données associées aux descriptions. Ils sont situés dans le même plan que ceux relatifs aux caractères morphologiques (branche *description*) afin de disposer d'une vision globale des données et méta-données des descriptions.

Formellement, la structure d'un modèle descriptif est une arborescence $\mathcal{M} = (\mathcal{A}, \mathcal{U})$, où l'ensemble des sommets $\mathcal{A}$ est un ensemble des objets et attributs. Les noeuds non-terminaux sont des objets, appelés également attributs structurés et sont notés $A_j :< A_1, \ldots, A_p >$ où A_j est la racine du sous-arbre dont les fils sont les $A_1, \ldots, A_p$,. Une arête $(A_j, A_l) \in \mathcal{U}$ exprime que A_l est un composant de A_j.

L'objectif principal de cette modélisation est de permettre aux experts de bâtir une représentation d'un groupe taxinomique donné, dans l'exemple, une famille corallienne. La structuration est une donnée essentielle du modèle car elle permet de regrouper les caractères qui définissent le même objet observable (*le squelette*, *les différentes sortes de columelles*, *la muraille*, etc.). Certains objets peuvent être absents des descriptions. Par exemple, les *columelles* des différents centres calicinaux (axiaux et latéraux) peuvent être absentes et certains individus du groupe *Fungiidae* [2] peuvent en être dépourvus, ce qui est visualisé par la présence d'un signe - à gauche de l'objet. Ces composants sont qualifiés "absents possibles" ou contingents.

Les caractères sont définis comme des attributs complexes décrits par plusieurs facettes telles que le type, le co-domaine, la question associée, la pondération, un commentaire explicatif, des liens hypertextuels, des illustrations. Ces dernières offrent une représentation photo-

2. La famille des Fungiidae se distingue par le fait que les colonies sont constituées d'un organisme unique (polype) de grande taille, à l'inverse des autres familles.

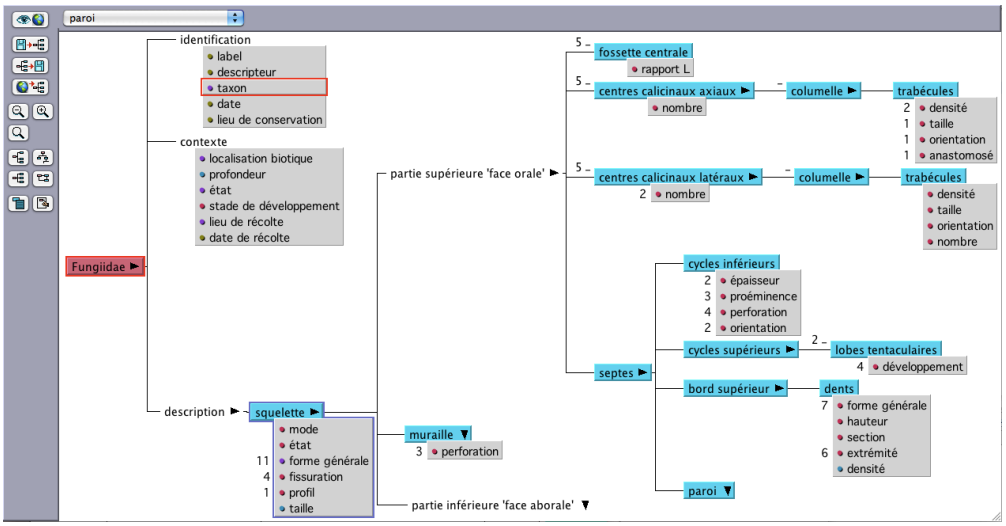

FIG. 1 – *Partie du modèle descriptif de la famille des Fungiidae, extrait d'IKBS.*

graphique des différents états de caractère (cf. figure 2). Dans cet exemple, le caractère *forme générale* de l'objet *squelette* est de type taxonomique (ou hiérarchisé), son co-domaine est organisé en une hiérarchie de spécialisations de valeurs.

Des règles de dépendance (ou d'implication) entre caractères peuvent être exprimées. Elles sont du type SI $c_1 = v_1$ ALORS $c_2 = v_2$ où les c_i sont des caractères et v_i des états. Ces règles permettent essentiellement de garantir la cohérence des descriptions.

2.2 Les descriptions

Une description est une sous-arborescence dérivée de celle de $\mathcal{A}$, obtenue par valuation de la présence/absence des objets et par valuation des caractères. Dans l'exemple de la figure 3, les caractères ont été valués par des valeurs complexes. Les valeurs peuvent être de type nominal, ordinal, taxonomique, numérique, discrète ou intervalle, conjonctive ou disjonctive, etc. Par exemple *densité des dents = [12 18]* indique une variation de densité par cm^2, *hauteur des dents = basse & moyenne* indique la présence simultanée des deux valeurs *basse* et *moyenne*, car c'est un ensemble de dents qui sont ici décrites et caractérisées. Les descriptions sont réunies au sein d'une base de cas.

La structure arborescente d'une description est appelée *squelette*. Un *squelette* décrit la structure morphologique d'une observation, il renseigne le statut de chaque composant, dont l'état peut être présent (+), absent (-) ou inconnu ($*$). Dans l'exemple, plusieurs composants sont absents : *centres calicinaux latéraux*, *lobes tentaculaires*, etc. Le symbole croix vient se superposer dans la figure. Certains composants comme la *columelle des centres calicinaux axiaux* sont inconnus car ils n'ont pu être observés. Le symbole inconnu permet d'exprimer le doute, l'imprécision.

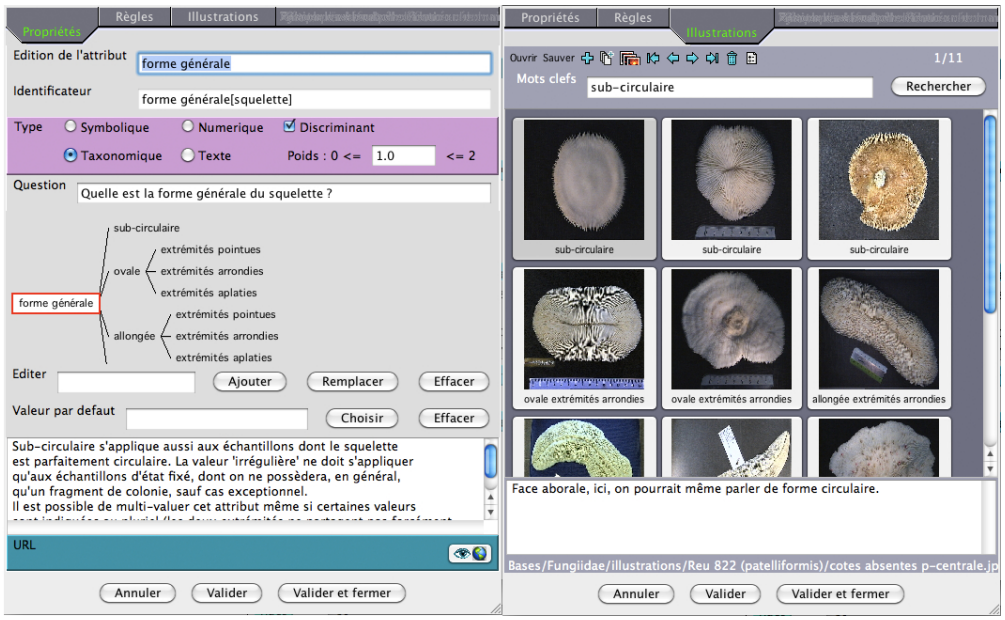

FIG. 2 – *Partie du modèle descriptif de la famille des Fungiidae, extrait d'IKBS.*

Notons S l'ensemble des symboles $\{+,-,*\}$. Une application $\sigma : \mathcal{A} \rightarrow S$ définit le squelette H_σ par l'arborescence annotée par S : $H_\sigma = (\mathcal{A}_\sigma, \mathcal{U})$ avec $\mathcal{A}_\sigma = \{(A_j, \sigma(A_j))_{j\in J}\}$, avec $A_j \in A$.

Sur l'ensemble des squelettes les règles de cohérence suivantes sont imposées. Pour tout attribut structuré $B :< B_l >_{l\in L}$, on a :

1. "Les fils d'un noeud absent sont absents" : $\sigma(B) = -$ alors $\sigma(B_l) = -$ pour $l \in L$,
2. "Les fils d'un noeud inconnu sont inconnus ou absents" : $\sigma(B) = *$ alors $\sigma(B_l) = *|-$ pour $l \in L$.
3. "Les fils d'un noeud présent peuvent être présents, absents ou inconnus" : $\sigma(B) = +$ alors $\sigma(B_l) = +|-|*$ pour $l \in L$.

On note $\mathcal{H}$ l'ensemble des squelettes vérifiant les règles de cohérence précédentes.

3 Mesures de similarité

Comment mesurer la similarité entre des arbres étiquetés enracinés ? Nous allons en premier lieu examiner quelques mesures classiques proposées pour la comparaison d'arbres.

3.1 Mesures de similarité structurelle en Biologie

Soient les arbres $H_1, H_2 \in \mathcal{H}$ étiquetés par les fonctions λ_1 et λ_2. Notons :
– n le nombre total d'attributs,

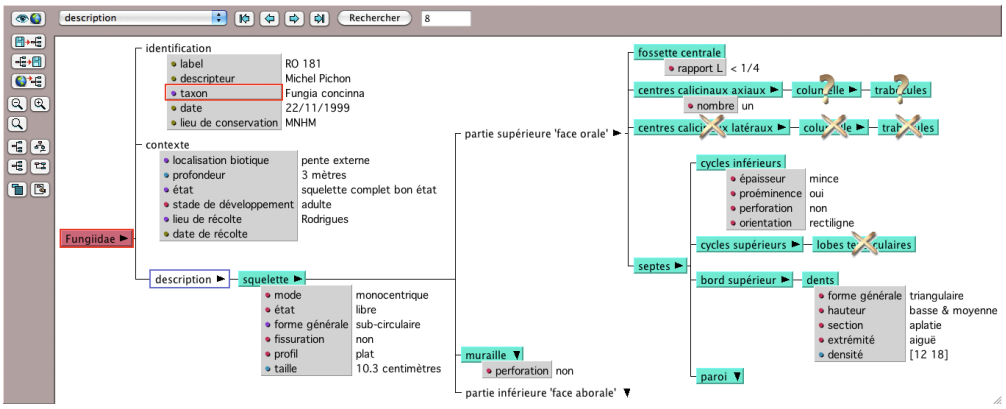

FIG. 3 – *Partie d'une description. Espèce Fungia concina, famille des Fungiidae.*

- $n_{=} = |\{\lambda_1(A_j) = \lambda_2(A_j) = +|-\}|$ le nombre de co-présence et de co-absence,
- $n_{**} = |\{\lambda_1(A_j) = \lambda_2(A_j) = *\}|$ le nombre de noeuds simultanément inconnus,
- $n_{*} = |\{\lambda_1(A_j)|\lambda_2(A_j) = *\}|$ le nombre de noeuds où un seul des deux noeuds est inconnu,
- $n_{\neq} = n - n_{=} - n_{**} - n_{*}$ le nombre de noeuds différents en ne tenant pas compte des valeurs inconnues.

En Biologie, Sokal et Sneath (1963) proposent l'indice de similarité suivant :

$$\zeta_{Sokal}(H_1, H_2) = \frac{n_{=}}{n}.$$

Le numérateur exprime que seules les co-présences et co-absences sont prises en compte, car on estime que la comparaison de deux noeuds inconnus n'est pas source de discordance (toutes les possibilités d'appariements +/- sont possibles). L'indice de dissimilarité associé à cet indice de similarité est :

$$d_{Sokal}(H_1, H_2) = \frac{n - n_{=}}{n} = \frac{n_{\neq} + n_{*} + n_{**}}{n}.$$

Estabrook et Rogers (1966) ont proposé une variante de l'indice de similarité de Sokal qui ne tient pas compte au dénominateur des valeurs inconnues :

$$\zeta_{Estabrook}(H_1, H_2) = \frac{n_{=}}{n - n_{**} - n_{*}}$$

3.2 Distance d'édition d'arbre

Les données se présentant comme des arbres étiquetés se rencontrent dans de nombreux domaines. Durant ces dernières années, elles ont conduit à différents travaux sur la distance entre arbres dite "distance d'édition" (Kaizhong et Shasha, 1989) (Bille, 2005), pour l'étude de la similarité des structures secondaires de l'ARN (Dulucq et Tichit, 2003), la classification de documents structurés XML (Dalamagas et al., 2006), par exemple.

La valeur de la distance d'édition entre deux arbres étiquetés et ordonnés est le coût minimum d'une séquence d'opérations qui transforme un arbre en un autre. Les opérations s_i sur

les noeuds sont : la suppression, l'insertion et le renommage. A chaque opération est associée une fonction de coût $\alpha(s_i)$ que l'on impose être une distance. Soit $S_{H_1H_2} = \{s, \ldots, s_n\}$ une séquence de transformation des arbres H_1, H_2, son coût est :

$$\alpha(S_{H_1H_2}) = \Sigma_i \alpha(s_i)$$

et la distance d'édition est alors

$$d_{edition}(H_1, H_2) = min_{S_{H_1H_2}} \alpha(S_{H_1H_2})$$

d_{ed} est une distance, car chaque fonction de coût est elle aussi une distance.

Dans notre cas, les arbres sont décrits par un modèle descriptif et le calcul de la "distance d'édition" s'en trouve simplifié car tous les arbres sont structurellement isomorphes. Afin de transformer de tels arbres, il n'y a pas lieu de supprimer ou d'insérer des noeuds. La seule opération reste le renommage. Notons $a \rightarrow b$ l'opération de renommer le noeud dont le label est a de l'arbre H_1 en b dans l'arbre H_2. Considérons le cas généralement adopté où le coût est constant : $\alpha(a \rightarrow b) = \frac{1}{n}$ pour tout noeud dont le label est $a \in H_1$ différent de $b \in H_2$. Notons $J_{\neq}$ l'ensemble des indices des noeuds dont les labels diffèrent dans les deux arbres : $J_{\neq} = \{j \in J | \lambda_1(A_j) \neq \lambda_2(A_j)\}$. La distance d'édition entre H_1 et H_2 s'écrit :

$$d_{edition}(H_1, H_2) = \Sigma_{j \in J_{\neq}} \alpha(\lambda_1(A_j) \rightarrow \lambda_2(A_j)) = \frac{n_{\neq} + n_*}{n}. \quad (1)$$

c'est-à-dire le pourcentage de labels qui diffèrent dans les deux arbres qu'ils soient inconnus ou non, d'où la relation suivante :

$$d_{edition} = d_{Sokal} + \frac{n_{**}}{n}$$

Compte tenu de la structure identique des arbres que nous traitons, l'utilisation des algorithmes complexes fondés sur la programmation dynamique (Kaizhong et Shasha (1989)), pour le calcul de cette distance d'édition ne se justifient pas. D'autre part, imposer que la fonction de coût de renommage des labels soit une distance ne permet pas de prendre en compte la structure hiérarchique des données dans le traitement des présences-absences de noeuds (voir paragraphe 3.3.2).

3.3 Mesure de similarité structurelle pondérée

Deux critiques essentielles peuvent être faites aux précédents indices. La première est qu'ils ne tiennent pas compte des liens de cohérence entre les noeuds (cf. paragraphe 2.2). La seconde est qu'un même poids est accordé aux différents noeuds, qu'ils soient en position élevée ou basse dans la hiérarchie du calcul de la similarité entre les arbres.

Afin d'accorder une importance différente aux noeuds, selon leur degré de profondeur, nous définissons une fonction de poids m sur les noeuds. Le poids d'un noeud est le nombre de sommets du sous-arbre dont il est la racine. D'autres choix sont possibles comme le rapport entre le poids du père d'un noeud sur le nombre de fils. Puis, on se donne une mesure de similarité σ, sur les fonctions labels, à valeurs dans $[0, 1]$ c'est-à-dire :

1. $\sigma(\lambda_1(A_j), \lambda_2(A_j) = 1 \Longleftrightarrow \lambda_1(A_j) = \lambda_2(A_j)$

2. $\sigma(\lambda_1(A_j), \lambda_2(A_j)) = \sigma(\lambda_2(A_j), \lambda_1(A_j))$

La similitude structurelle de deux squelettes est alors évaluée comme la moyenne des valeurs de similitude des noeuds. La mesure de similarité structurelle pondérée, notée $\zeta_{SPondere}$ a pour expression :

$$\zeta_{SPondere}(H_1, H_2) = \frac{\Sigma_{j \in J} m(A_j)\sigma(\lambda_1(A_j), \lambda_2(A_j))}{\Sigma_{j \in J} m(A_j)}.$$

Le tableau 1 donne l'expression générale de la mesure de similarité σ pour un noeud donné. Cette mesure de similarité généralise celle de Sokal, et la distance d'édition d'arbre pour un choix particulier des poids des noeuds et des valeurs de σ.

3.3.1 Comparaison avec des indices classiques

Si on attribue à chaque noeud A_j le même poids $m(A_j) = 1$ et que la mesure de similarité choisie est σ_0 (tableau 2) alors la mesure de similarité structurelle pondérée est égale à l'indice de Sokal. En effet, elle s'écrit sous ces hypothèses :

$$\zeta_0(H_1, H_2) = \frac{\Sigma_{j \in J} \sigma_0(\lambda_1(A_j), \lambda_2(A_j))}{n} = \frac{n_=}{n} = \zeta_{Sokal}(H_1, H_2).$$

car seules les co-présence et les co-absence sont prises en compte par la mesure de similarité σ_0.

La distance d'édition du paragraphe 3.2 est un cas particulier de la distance associée à la mesure de similarité $\zeta_{SPondere}$. On considère que les noeuds ont toujours le même poids et la mesure de similarité choisie est $\sigma_{edition}$ (tableau 3). On a

$$\zeta_{edition}(H_1, H_2) = \frac{\Sigma_{j \in J} \sigma_1(\lambda_1(A_j), \lambda_2(A_j))}{n} = \frac{n_= + n_{**}}{n}.$$

car la similarité entre les termes non diagonaux est nulle. La valeur de la distance associée à la mesure de similarité $\zeta_{edition}$ est :

$$d(H_1, H_2) = 1 - \zeta_{edition}(H_1, H_2) = 1 - \frac{n_= + n_{**}}{n} = \frac{n - n_= - n_{**}}{n} = \frac{n_{\neq} + n_*}{n}$$

elle est bien égale à celle de la distance d'édition (1). En fait à $\zeta_{edition}$ est associée la distance d'édition dont la fonction de coût pour un noeud A_j est :

$$\alpha(\lambda_1(A_j) \rightarrow \lambda_2(A_j)) = \frac{1}{n}(1 - \sigma_{edition}(\lambda_1(A_j), \lambda_2(A_j))$$

Cette fonction de coût est bien une distance. Le calcul de la distance d'édition associée donne :

$$d(H_1, H_2) = \Sigma_{j \in J_{\neq}} \alpha(\lambda_1(A_j) \rightarrow \lambda_2(A_j)) = \Sigma_{j \in J_{\neq}} \frac{1}{n}(1 - \sigma_{edition}(\lambda_1(A_j), \lambda_2(A_j))$$

comme $\Sigma_{j \in J_{\neq}} \frac{1}{n} = \frac{n_{\neq} + n_*}{n}$, et que $\sigma_{edition}$ prend des valeurs nulles en dehors de la diagonale, c'est-à-dire pour les indices $j \in J_{\neq}$, on a :

$$d(H_1, H_2) = \frac{n_{\neq} + n_*}{n}$$

et on retrouve :

$$d_{edition}(H_1, H_2) = 1 - \zeta_{edition}(H_1, H_2)$$

TAB. 1 – *La mesure de similarité σ pour un noeud donné*

$\lambda_1 \diagdown \lambda_2$	$+$	$-$	$*$
$+$	1	α_j^1	β_j^1
$-$	α_j^2	1	γ_j
$*$	β_j^2	γ_j	1

TAB. 2 – *La mesure de similarité σ_0*

$\lambda_1 \diagdown \lambda_2$	$+$	$-$	$*$
$+$	1	0	0
$-$	0	1	0
$*$	0	0	0

3.3.2 Choix des valeurs de similarité

Nous donnons ci-dessous des indications pour le choix des valeurs de la mesure de similarité σ (tableau 1) entre les noeuds.

- $\sigma(\lambda_1(A_j) = +, \lambda_2(A_j) = -) = \alpha_j^1$. Si le noeud A_j est une feuille alors $\alpha_j^1 = 0$, sinon sa valeur est le rapport entre le nombre de fils absents du noeud A_j de H_1 sur le nombre de ses fils. α_j^1 mesure l'importance du sous-arbre absent du noeud présent dans H_1, plus ce noeud a des fils absents plus la similarité sera grande avec celui de H_2. Notons la dissymétrie suivante, il n'y a aucune raison pour que cette valeur α_j^1 soit égale à $\alpha_j^2 = \sigma(\lambda_1(A_j) = -, \lambda_2(A_j) = +)$. Cette valeur ne peut donc pas être un coût associé à une opération de renommage d'une distance d'édition pour laquelle on doit avoir : $\alpha(+ \rightarrow -) = \alpha(- \rightarrow +)$.
- $\sigma(\lambda_1(A_j) = +, \lambda_2(A_j) = *) = \beta_j^1$. La valeur inconnue $*$ correspond à une présence $(+)$ ou une absence $(-)$ du noeud donné. On calcule dans l'échantillon les probabilités de présence simultanée $Pr(\lambda_1(A_j) = +, \lambda_2(A_j) = +)$ et de présence-absence $Pr(\lambda_1(A_j) = +, \lambda_2(A_j) = -)$. Si la co-présence est la plus probable alors $\beta_j^1 = 1$ sinon $\beta_j^1 = \alpha_j^1$. La valeur par défaut de ce paramètre est 0.
- $\sigma(\lambda_1(A_j) = -, \lambda_2(A_j) = *) = \gamma_j$. Si la configuration $(\lambda_1(A_j) = -, \lambda_2(A_j) = -)$ est la plus probable dans l'échantillon alors $\gamma_j = 1$ sinon sa valeur est 0. Par défaut, sa valeur est 0.
- Les valeurs α_j^2, β_j^2 du tableau 1 sont calculées de manière symétrique.

Le tableau 7 donne des exemples de calcul de la mesure de similarité structurelle sur des observations.

3.4 Mesure de similarité structurelle récursive

Les mesures de similarité structurelle pondérée, en sommant sur tous les noeuds, introduisent une redondance pour ceux absents (resp. inconnus), car leurs descendants sont absents (resp. inconnus) selon les règles de cohérence présentées au paragraphe 2.2. L'idée est d'affiner cette valeur en considérant la similitude des descendants uniquement pour les noeuds

TAB. 3 – *La mesure de similarité* $\sigma_{edition}$

$\lambda_1 \diagdown \lambda_2$	$+$	$-$	$*$
$+$	1	0	0
$-$	0	1	0
$*$	0	0	1

TAB. 4 – *La mesure de similarité* σ_{SP}

$\lambda_1 \diagdown \lambda_2$	$+$	$-$	$*$
$+$	1	α_j^1	0
$-$	α_j^2	1	0
$*$	0	0	1

simultanément présents dans les deux squelettes. Soient les squelettes H_1 et H_2. Le degré de similitude est défini de manière récursive pour les noeuds simultanément présents, à l'aide d'une mesure de similitude σ_r définit comme suit :

- Si A_j est une feuille alors $\sigma_r(\lambda_1(A_j) = +, \lambda_2(A_j) = +) = 1$,
- Sinon c'est un noeud non terminal $A_j :< A_k >_{k \in K}$ alors :

$$\sigma_r(\lambda_1(A_j) = +, \lambda_2(A_j) = +) = \frac{1 + \Sigma_k m(A_k) \sigma_r(\lambda_1(A_k), \lambda_2(A_k))}{m(A_j)} \tag{2}$$

qui vérifie bien les propriétés d'une mesure de similarité. Le tableau 5 donne l'expression générale de σ_r. La mesure de similarité entre deux squelettes est calculée par une "descente récursive" à partir de la racine A :

$$\zeta_{SRecursive}(H_1, H_2) = \sigma_r(\lambda_1(A), \lambda_2(A)) \tag{3}$$

Dans ce parcours, la détermination des valeurs de similarité entre les noeuds non simultanément présents n'est faite que pour les noeuds les plus élevés de l'arbre. Dans le tableau 5 sont données des valeurs de cette similarité entre des arbres pour la mesure de similarité σ_{r_e} (tableau 6),

TAB. 5 – *La mesure de similarité récursive* σ_r *pour un objet* $A_j =< A_k >_{k \in K}$

$\lambda_1 \diagdown \lambda_2$	$+$	$-$	$*$
$+$	$\frac{1+\Sigma_k m(A_k)\sigma_r(\lambda_1(A_k),\lambda_2(A_k))}{m(A_j)}$	α_j^1	β_j^1
$-$	α_j^2	1	γ_j
$*$	β_j^2	γ_j	1

TAB. 6 – *La mesure de similarité récursive* σ_{r_e}

$\lambda_1 \backslash \lambda_2$	$+$	$-$	$*$
$+$	$\frac{1+\Sigma_k m(A_k)\sigma_r(\lambda_1(A_k),\lambda_2(A_k))}{m(A_j)}$	α_j^1	0
$-$	α_j^2	1	0
$*$	0	0	1

3.5 Mesure de similarité sur les valeurs

Dans cette partie, nous nous intéressons aux similarités portant sur les valeurs relatives aux attributs. L'ensemble des attributs est noté : $\{(A_q, D_q)|q \in Q\}$, la valeur d'une observation pour un attribut A_q est dans son domaine D_q si et seulement si l'attribut est présent.

On suppose donnée une mesure de similarité s_q sur chaque domaine d'un attribut : $s_q : D_q \times D_q \rightarrow [0, 1]$. Dans le logiciel IKBS, diverses distances adaptées à différents types d'attribut sont disponibles : la distance euclidienne normalisée pour les attributs de type numérique, la distance du khi-deux pour des attributs qualitatifs, des distances adaptées aux variables intervalles (Grosser et al., 2000).

L'indice de "similarité valeur" entre deux observations portera sur les valeurs des attributs présents en commun. Soit une observation $o \in O$, dont la fonction label est λ_o, on note l'ensemble de ces indices des attributs $Q_+(o) = \{q \in Q|\lambda_o(A_q) = +\}$. L'ensemble des valeurs de o est :

$$v(o) = (v_q|v_q \in D_q, q \in Q_+(o)).$$

La mesure de similarité valeur $\zeta_V : O \times O \rightarrow [0, 1]$ est définie pour deux observations o et o' dont les valeurs sont $v(o) = (v_q), v(o') = (v'_q)$ comme :

$$\zeta_V(o, o') = \frac{\Sigma_{q \in Q_+(o) \cap Q_+(o')} s_q(v_q, v'_q)}{|Q_+(o) \cap Q_+(o')|}$$

pour $|Q_+(o) \cap Q_+(o')| \neq \emptyset$ et $\zeta_V(o, o') = 0$ sinon.

3.6 Mesures de similarité globale

Une observation $o \in O$ est décrite par son squelette H_o et ses valeurs $v(o) = (v_q)$. On suppose que l'on dispose d'une mesure de similarité structurelle $\zeta_S \in [0, 1]^{O \times O}$ (indice $\zeta_{SPondere}$ ou $\zeta_{SRecursive}$) et d'une mesure de similarité valeur $\zeta_V \in [0, 1]^{O \times O}$. Pour définir un indice de similarité unique qui tienne compte à la fois de la structure et du contenu, on peut considérer la moyenne pondérée des deux indices. Pour les observations $o = (H_o, v(o) = (v_q))$ et $o' = (H_{o'}, v(o') = (v'_q))$, la mesure de similarité globale moyenne s'écrit :

$$\zeta_G(o, o') = \alpha\zeta_S(H_o, H_{o'}) + (1 - \alpha)\zeta_V(o, o')$$

où $\alpha \in]0, 1[$ est un coefficient qui permet de pondérer l'importance accordée à la structure ou aux valeurs.

Cependant on remarque que les attributs présents dans deux squelettes, sont comptabilisés deux fois, une fois dans la mesure de similarité structurelle pondérée ou récursive, avec une

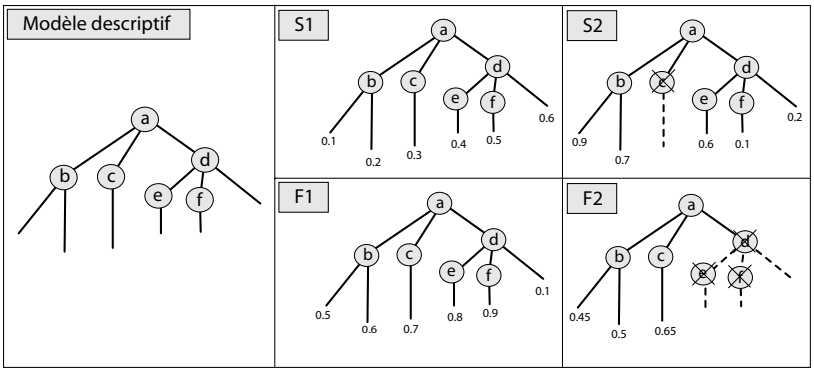

FIG. 4 – *Le modèle descriptif et les observations.*

TAB. 7 – *Valeurs des mesures de similarité.*

Similarité structurelle	$\zeta_{Edition}$	$\zeta_{SPondere}$	$\zeta_{SRecursive}$	ζ_V
Arbres $S1$ et $S2$	0.80	0.87	0.83	0.53
Arbres $F1$ et $F2$	0.40	0.33	0.50	0.93
Similarité globale	$\frac{(\zeta_{Edition}+\zeta_V)}{2}$	$\frac{(\zeta_{SPondere}+\zeta_V)}{2}$	$\frac{(\zeta_{SRecursive}+\zeta_V)}{2}$	ζ_{GR}
Arbres $S1$ et $S2$	0.66	0.70	0.69	0.64
Arbres $F1$ et $F2$	0.66	0.63	0.72	0.48

valeur égale à 1 et une autre fois avec la valeur de la mesure de similarité valeur. Pour éviter cette redondance, nous proposons la similarité globale ζ_{GR} en modifiant la mesure de similarité structurelle récursive proposée de la section 3.4 pour qu'elle prenne en compte la similitude au niveau des attributs présents en commun. Si A_j est une feuille alors on pose :

$$\sigma_r(\lambda_o(A_j) = +, \lambda_{o'}(A_j) = +) = s_j(v_j, v'_j),$$

sinon c'est un noeud non terminal $A_j :< A_k >_{k\in K}$, et l'expression (2) est appliquée. La mesure de similarité globale $\zeta_{GR}(o, o')$ est calculée récursivement à partir de la racine par la formule (3).

3.7 Exemples

Les observations $S1$ et $S2$ de la figure 4 ont des structures similaires mais des valeurs d'attributs différentes, c'est l'inverse pour $F1$ et $F2$. Le tableau 7 recensent pour ces arbres les valeurs des mesures de similarité structurelle : $\zeta_{Edition}, \zeta_{SPondere}, \zeta_{SRecursive}$, associées respectivement aux similarités $\sigma_{edition}$ (tableau 3), σ_{SP} (tableau 4), σ_{r_e} (tableau 6), et la mesure de similarité valeur ζ_V relative à la distance euclidienne de la valeur absolue. Les mesures globales moyennes sont calculées à partir de ces valeurs, de même que la mesure de similarité récursive ζ_{GR}. Les valeurs de la similarité entre $S1, S2$ (resp. $F1, F2$.) pour $\zeta_{SPondere}$ est supérieure (resp. inférieure) à celle de la mesure d'édition $\zeta_{Edition}$, et $\zeta_{Edition}(S1, S2) <$

$\zeta_{SRecursive}(S1, S2)$. Ces résultats illustrent la pertinence des mesures structurelles proposées qui font jouer aux noeuds élevés un rôle plus important dans la mesure de la ressemblance des structures. Si les moyennes sont des bons compromis de mesures globales pour prendre en compte la similarité de la structure et des valeurs, la mesure ζ_{GR} évite le choix délicat du coefficient de pondération α.

3.8 Conclusion sur les mesures de similarité

Après avoir montré les limites des indices classiques pour mesurer les similarités des données morphologiques complexes, nous avons proposé des mesures de similarité qui prennent en compte à la fois la structure hiérarchique des données, les contraintes de cohérence entre les objets, et les valeurs des attributs : numériques, qualitatives, multiples, manquantes ou inconnues. Ces différentes mesures sont mises en oeuvre dans la méthode de fouille que nous présentons maintenant.

4 Classification par voisinages successifs

Afin de déterminer l'appartenance à une classe d'une description partielle décrite par un utilisateur, nous proposons une méthode itérative et interactive de classification par voisinages successifs (CVS). La méthode est hybride, elle met en oeuvre conjointement une stratégie de type k-plus-proches voisins ainsi qu'une étape de sélection de variables discriminantes de type arbre de décision. Dans la suite, on note $(\mathcal{E}, d)$ l'espace métrique des descriptions où $\mathcal{E} = \mathcal{H} \times \mathcal{V}$ avec $\mathcal{H}$ l'ensemble des squelettes, $\mathcal{V}$ l'ensemble des valeurs des observations et $d = 1 - \zeta$ l'indice de dissimilarité dérivé d'un indice de similarité global ζ précédement défini.

4.1 Principe de la méthode

La méthode proposée consiste dans un premier temps à compléter la description du spécimen en appliquant des règles d'inférence liées au domaine (cf. ci-dessous). Puis le système sélectionne un ensemble de spécimens proches de la description en cours. Il donne différentes informations relatives aux classes d'appartenance de ces "voisins" pour aider à la prise de décision. Si l'utilisateur n'est pas satisfait des suggestions du système, ce dernier calcule un ensemble d'attributs discriminants parmi l'ensemble d'attributs non encore renseignés et les propose à la valuation. Puis, il recommence à sélectionner un ensemble de voisins pour permettre la prise de décision.

4.2 Cohérence des descriptions

On note par $e \in \mathcal{E}$ la description partielle d'un spécimen dont la classe est à déterminer. Les règles d'inférence suivantes sont d'abord appliquées :

1. Les descendants d'un noeud non renseigné sont valués à "$*$",
2. Les descendants d'un noeud absent sont valorisés à "$-$",
3. Les ancêtres d'un noeud présent ou d'une feuille valuée sont considérés comme "présent".

Si d'autres règles de cohérence liées aux connaissances sont disponibles, elles sont mises en oeuvre à ce stade. Après cela, nous obtenons une description $e \in \mathcal{E}$.

4.3 Principes de l'algorithme CVS

L'algorithme CVS permet de calculer la classe d'appartenance d'un spécimen donné e. Il est itératif et comporte les étapes suivantes :

1. Initialiser la valeur du rayon Δ à la distance maximum de e à l'ensemble des observations.
2. Déterminer l'ensemble des voisins comme l'ensemble des objets à l'intérieur de la sphère de rayon Δ et de centre e,
3. Calculer les scores de classification des classes *a priori*,
4. Calculer la nouvelle valeur du rayon Δ à partir de l'ensemble des voisins,
5. Répéter les étapes 2,3,4 jusqu'à ce que les critères d'arrêt soient satisfaits,
6. A partir des scores de classification, proposer une classe d'appartenance de la description à l'utilisateur, si satisfaction alors fin.
   - sinon à partir de l'ensemble des voisins, calculer le pouvoir discriminant des attributs non encore valués,
   - L'utilisateur est invité à compléter la description avec une ou plusieurs valeurs des attributs suggérés.
   - Réitérer l'algorithme de discrimination par voisinage successif avec cette description complétée.

4.4 Calcul du voisinage d'une description partielle

L'ensemble des voisins de e à l'itération m est défini comme l'ensemble des objets dans la sphère de rayon Δ_m et de centre e :

$$voisin_{(m)} = \{o \in O | d(e, o) < \Delta_m\}.$$

la valeur du rayon est déterminée comme suit : soit la mesure de dissimilarité maximum entre e et un ensemble A :

$$D_{max}(e, A) = max_{a \in A} d(e, a)$$

on choisit pour valeur de Δ_m la distance maximum :

$$\Delta_m = D_{max}(e, N_{(m-1)}). \tag{4}$$

La suite de nombre positifs Δ_m est décroissante, car par définition de l'ensemble des voisins

$$\Delta_{m+1} = D(e_m, voisin_{(m)}) = max_{o \in voisin_{(m)}} d(e_m, o) < \Delta_m$$

Si à chaque objet o_i est associé un poids normalisé p_i, alors la distance moyenne $D_{moy}(e, A) = \frac{1}{\Sigma_{o_i \in A} p_i} \Sigma_{o_i \in A} p_i d(e, o_i)$ ou celle de l'inertie $D_{iner}(e, A) = \frac{1}{\Sigma_{o_i \in A} p_i} \Sigma_{o_i \in A} p_i d^2(e, o_i)$ peuvent être aussi choisies, car on peut montrer que ces suites sont décroissantes.

4.5 Suggestion de classes d'appartenance à partir de différents scores

Soient $\{C_l\}_{l \in K}$ l'ensemble des classes a priori d'appartenance possibles.

On note $Pr(C_l|N_{(m)}) = \frac{|C_l \cap N_{(m)}|}{|N_{(m)}|}$ la probabilité de la classe C_l dans le contexte de voisins $N_{(m)}$ ou la fréquence relative de C_l dans l'ensemble de voisins $N_{(m)}$. La classe qui sera proposée comme celle d'appartenance du spécimen e sera choisie à partir des classes telles que les probabilités $Pr(C_l|N_m)$ seront significativement différentes des probabilités à priori des classes $Pr(C_l|O) = \frac{|C_l|}{|0|}$. Les tests statistiques usuels de calcul de fréquence peuvent être appliqués à ce stade. Le score de classification de la classe C_l à l'itération m est :

$$R_l = \frac{Pr(C_l|N_m)}{Pr(C_l|O)}. \tag{5}$$

4.6 Sélection d'attributs discriminants

La méthode de sélection est interactive, car la liste des variables discriminantes est calculée à chaque étape du processus d'identification, en fonction de plusieurs critères et d'une réponse demandée à l'utilisateur. Elle est semi-dirigée dans le sens où l'utilisateur peut choisir une autre variable parmi la liste proposée ou bien apporter une réponse inconnue. Dans ce cas, la variable en seconde position sera automatiquement choisie.

La méthode utilise une combinaison de plusieurs critères de choix, de façon à minimiser le nombre de questions et prendre également en compte certaines connaissances de fond du domaine relatives à la qualité des variables, exprimées dans le modèle descriptif :

1. Construction de la liste des attributs éligibles. Cette liste est contextuelle, constituée des attributs pouvant être sélectionnés. Elle dépend aussi de la structure du modèle et des attributs déjà renseignés. La présence des attributs éligibles doit notamment être vérifiée pour chaque composant contingent (cf. section 2.1). Une question booléenne portant sur sa présence/absence est ajoutée à la liste.
2. Choix de critère classique de calcul du gain d'information utilisé en apprentissage, tel que l'*entropie de Shannon* ou le *Gini Index*. Ce type de critère permet de minimiser la longueur de la diagnose en privilégiant les attributs les plus discriminants.
3. Pondération des attributs. Chaque attribut peut être pondéré par l'expert sur une échelle réelle de 0 à 1. Le poids 0 est affecté à un attribut non discriminant. Les pondérations sont très utiles pour introduire une connaissance d'ordre stratégique sur les caractères. Certains attributs sont en effet particulièrement difficile à observer (nécessite du matériel spécifique), à décrire ou sujet à interprétation. L'expert peut ainsi choisir de minimiser leur occurrence.

4.7 Critères d'arrêt

Plusieurs critères d'arrêt sont considérés :

– Le score de classification d'une classe C_l (cf. formule 5) supérieure à un seuil fixé par l'utilisateur.
– Le nombre minimal de voisins, car le rayon Δ_m est une suite décroissante (cf. formule 4).

– Le nombre maximal d'itérations.
– La description est complète ou bien il n'y a plus de variables discriminantes disponibles.
Notons que ces différents critères sont exclusifs et ne peuvent être vérifiés simultanément.

5 Applications

Dans cette partie, nous procédons à différents tests de validation en grandeur réelle sur des données structurées issues de la base de connaissances sur les coraux des Mascareignes. L'objectif poursuivi est double. D'une part (section 5.1), illustrer l'exécution de la méthode CVS dans l'environnement IKBS sur un exemple réel. D'autre part (section 5.2), comparer les performances relatives de la méthode CVS par rapport à un classement par arbre d'identification et un classement par la méthode des k-plus-proches voisins.

5.1 Classement d'une description

L'exemple choisi est l'illustration du processus de classement d'une description de référence e_r appartenant à l'espèce *Fungia concinna*, de la famille des *Fungiidae*, donné pour exemple dans la partie 2. La famille des *Fungiidae* est structurée en quatre genres (*Cycloseris*, *Fungia*, *Herpolita* et *Podabacia*) et onze espèces (cf. figure 5). La base *Fungiidae* compte 63 descriptions (cas) comportant 94 caractères (attributs) et 15 taxons (classes).

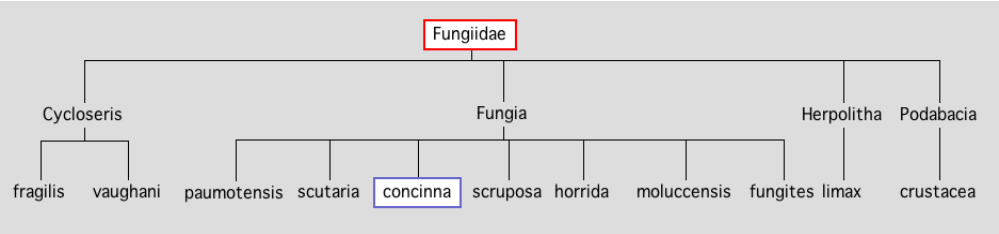

FIG. 5 – *Taxonomie de la famille des Fungiidae, comprenant quatre genres et onze espèces.*

L'algorithme cherche à associer à e_m (le cas test, initialement vide) un élément de la hiérarchie de taxons (fig. 5), en exploitant les informations du modèle descriptif et de la base de cas de référence. Afin de simuler l'interactivité avec l'utilisateur, la source de données de e_m est e_r. A chaque itération, e_m est alimenté par un attribut de e_r. L'étiquette de e_r est utilisée à la fin du processus pour vérifier que le classement est correct ou non. La mesure du gain d'information utilisée dans cette application est la mesure de l'entropie de Shannon (gain d'information). Dans l'exemple, l'algorithme CVS a conduit au bon classement de e_m en 21 itérations.

La table 8 illustre un sous-ensemble choisi d'itérations qui ont conduit au bon classement de e_t. Pour chaque itération, les informations suivantes sont renseignées :
– le numéro de l'itération,
– l'attribut sélectionné et le gain d'information associé,
– la valeur correspond à l'attribut pour le cas de référence affectée à e_m, ainsi que des informations relatives au voisinage de e_m.

TAB. 8 – *Illustration du classement d'une description structurée par la méthode CVS.*

Num iter	Attribut	Valeur	Voisinage			
			cas	classe	valeur	Dist
1	forme[squelette]	circulaire	62	fungites	circulaire	0,2216
	gain = 0,628		46	fungites	circulaire	0,2239
			51	scruposa	circulaire	0,2259
7	taille[squelette]	10.3 cm	62	fungites	9,5 cm	0,1905
	gain = 0,423		46	fungites	9 cm	0,1930
			60	fungites	15,5 cm	0,1974
8	densité[épines]	[10 16]	62	fungites	[6 10]	0,1912
	gain = 0,423		46	fungites	6	0,1946
			60	fungites	[6 8]	0,1978
9	profil[squelette]	plat	60	fungites	plat	0,1924
	gain = 0,394		62	fungites	convexe	0,1935
			46	fungites	plan convexe	0,1968
19	dév[côtes]	sub égales	46	fungites	sub égales	0,1901
	gain = 0,211		**39**	**concinna**	sub égales	0,1915
			60	fungites	inégales	0,1930
20	lobes-tentaculaires	absent	46	fungites	absent	0,1901
	gain = 0,173		**39**	**concinna**	absent	0,1915
			60	fungites	absent	0,1930
21	forme[épines]	cylindre	39	**concinna**	cylindre	0,1912
	gain = 0,153		46	fungites	conique	0,1935
			62	fungites	cylindre	0,1941

Pour chaque voisinage, sont donnés les numéros des cas de la base de référence les plus proches, le taxon (classe) associé, la valeur de l'attribut sélectionné, ainsi que la distance du cas e_m. Notons que pour des raisons pratiques, seuls les trois cas les plus proches du voisinage sont affichés.

Différents types d'attributs sont sélectionnés lors de cette application : type hiérarchique (forme[squelette], cf. fig. 2) nominal, continu, intervalle ou booléen portant sur la présence d'un composant (présence des lobes-tentaculaires). Pour les besoins de notre application, le critère d'arrêt qui s'applique ici est la correspondance entre la classe du cas de référence et la classe du premier voisin. Dans un processus réel de classement pour lequel l'étiquette n'est pas *a priori* connue, ce critère s'applique selon les modalités développées à la section 4.5.

L'information la plus intéressante à observer dans ce tableau est la liste des cas les plus proches. Les variations de position des voisins permet d'évaluer intuitivement dans quelle mesure l'adjonction d'une information supplémentaire modifie les distances des éléments du voisinage au cas e_m et par conséquent modifie l'ordre des voisins. Ainsi par exemple, lors du passage de l'itération 8 à 9, le cas 60 passe en première position du fait que la valeur du profil du squelette (valeur *plat*) correspond à la valeur de référence. A l'itération 19, apparait dans la liste des trois candidats le cas 39 en position 2, dont la valeur de classe correspond à la classe recherchée. Enfin, à l'itération 21, le cas 39 passe en première position devant le cas 46, du fait

d'une correspondance avec l'attribut forme[épines], ce qui aboutit à un bon classement.

5.2 Performances de la méthode CVS

Dans cette seconde expérience, nous souhaitons tester les performances relatives de l'algorithme CVS par rapport à deux méthodes de référence disponibles dans IKBS : les arbres d'identification et les k-plus-proches voisins. La méthode des arbres d'identification (Grosser (2002)) est une extension de la méthode de classification supervisée C4.5 (Quinlan (1993)) adaptée aux données complexes de notre modèle de représentation. La méthode est monothétique (les caractères sont proposés un à un) et interactive. Le modèle inductif utilisé pour le classement des attributs discriminants est identique à celui utilisé par la méthode CVS lors de la phase de sélection. Le second algorithme est de type k-plus-proches voisins. Dans l'expérience, le facteur k est simplement positionné à 1. La mesure de similarité utilisée est la mesure récursive ζ_{GR} exposée précédemment. La classe du cas le plus proche est affectée au cas test et la totalité de l'information du cas de référence est utilisée. La méthode est polythétique, un ensemble de caractères devant être renseignés *a priori*. La mesure de similarité utilisée est la même que celle de l'algorithme CVS. La méthode de validation utilisée, de type "Leave-on-out", consiste à classer chaque cas de la base en utilisant les autres cas comme base de référence. La méthode est appliquée pour les trois algorithmes dans des conditions identiques. Les paramètres de classement sont identiques à ceux utilisés pour la première expérience.

TAB. 9 – *Tests de validation "Leave-on-out" de différentes bases*

Bases	nb	nb	nb	Arbre ident		K-voisins		CVS	
	class	cas	attr	score	taux	score	taux	score	taux
Faviinae	36	92	146	65	70,65%	85	92,39%	84	91,30%
Montastreinae	15	24	118	17	70,83%	22	91,66%	19	79,16%
Fungiidae	15	63	94	47	74,60%	58	92,06%	55	87,30%
Mussidae	15	56	28	49	87,50%	54	96,42%	51	91,07%
Poritidae	28	28	87	22	78,57%	24	85,71%	19	67,85%
Siderastreidae	14	60	99	49	81,67%	56	93,33%	57	95,00%

La table 9 illustre les résultats des trois méthodes de classification sur 6 bases extraites de la base coraux. Pour chaque base, les informations suivantes sont affichées : nombre de taxons pour le modèle (nb class), nombre de cas (nb cas), nombre d'attributs (nb attr). Ensuite pour chaque méthode est donné le nombre (score) et le taux de cas bien classés.

Globalement, la méthode la moins robuste est la méthode par arbre d'identification. Les erreurs de classement sont fréquentes, de 12.5% sur la base *Mussidae* à $29,35\%$ pour la base *Faviinae*. La méthode la plus robuste est la méthode de type k-voisins, qui exploite la totalité de l'information disponible lors du calcul des similarités 2 à 2. Les taux d'échec pour celle-ci ce situent entre 14.29% (base *Poritidae*) et 3.58% (base *Mussidae*). La méthode CVS offre un score intermédiaire relativement proche de la méthode K-voisins et parfois très supérieur à la méthode par arbre. Observer par exemple les résultats de la base *Faviinae* qui montrent un écart de plus de 20% de bonnes identifications avec la méthode CVS.

La méthode K-voisins donne de bons résultats, mais est difficilement exploitable par des utilisateurs non-experts dans des cas réels d'utilisation. Il est en effet très difficile de choisir *a priori* un sous-ensemble de caractères discriminants, sans une connaissance approfondie du domaine. Il est donc très utile de disposer d'un processus interactif qui guide l'observation et suggère les caractères à décrire. L'approche monothétique (un caractère à la fois) proposée par la méthode des arbres d'identification offre cette facilité. C'est la raison pour laquelle cette méthode est très utilisée, par les utilisateurs d'IKBS, experts et biologistes et par toutes les personnes qui ont besoin d'identifier des organismes, en particulier pour l'évaluation et la conservation de la biodiversité. Nous pensons que la méthode CVS offre une solution alternative intéressante, monothétique et interactive, plus performante que les méthodes par arbres.

6 Conclusion

Dans le cadre de nos travaux sur la représentation et le traitement de connaissances en Systématique, plusieurs méthodes d'identification et de classification de données complexes issues de bases de connaissances ont été développées. Ces méthodes ont été implantées et évaluées à l'aide de l'outil IKBS. Ce logiciel a été utilisé par un groupe d'experts pour la conception d'une base de connaissances sur les coraux des Mascareignes. Les données que nous considérons dans cet article sont des descriptions morphologiques d'objets biologiques. Elles sont structurées en arborescences selon un modèle descriptif construit par les experts et peuvent être multi-valuées et incomplètes. Le modèle descriptif présente un plan d'organisation des variables descriptives (les caractères) utilisées pour décrire les objets. Cette modélisation offre l'avantage de représenter les relations de dépendances entre caractères, ainsi que de permettre l'expression de connaissances de fond relative au domaine (poids des caractères, illustrations, valeurs par défaut, commentaires, etc.).

Afin de comparer ce type de données structurées, un ensemble de mesures de similarités prenant en compte à la fois la structure et le contenu des descriptions a été développé et évalué. Une méthode itérative originale de classification par construction de voisinages successifs a également été présentée. C'est une méthode hybride qui utilise conjointement ce type d'indice pour la construction de l'espace des voisins (stratégie de type k-plus-proches voisins) ainsi qu'une méthode de sélection de variables discriminantes (stratégie de type arbre de décision) pour le choix des variables à chaque itération. La structure du modèle descriptif est utilisée à chaque étape du processus pour guider le choix des variables, garantir la cohérence des descriptions partielles et calculer les similarités. Une expérimentation sur différentes bases montre que la méthode présente une assez bonne résistance aux bruits comparativement aux méthodes de classement par arbres d'identification, tout en offrant des caractéristiques intéressantes d'interactivité et d'explication.

Références

Bille, P. (2005). A survey on tree edit distance and related problems. *Theoret. Comput. Sci. 337*(1-3), 217–239.

Breiman, L., J. H. Freidman, R. A. Olshen, et C. J. Stone (1984). *Classification and regression trees*. Wadsworth, Belmont.

Conruyt, N. (1994). *Amélioration de la robustesse des systèmes d'aide à la description, à la classification et à la détermination des objets biologiques*. Thèse de doctorat, Université Paris IX-Dauphine.

Dalamagas, T., T. Cheng, K. Winkel, et T. K. Sellis (2006). A methodology for clustering xml documents by structure. *Information System 31*(3), 187–228.

Dulucq, S. et L. Tichit (2003). RNA secondary structure comparison: exact analysis of the zhang-shasha tree edit algorithm. *Theor. Comput. Sci. 306*(1-3), 471–484.

Estabrook, G. F. et D. Rogers (1966). A general method of taxonomic description for a computed similarity measure. *Bioscience 16*, 789–793.

Faure, G., N. Conruyt, M. Pichon, M. Guillaume, D. Grosser, et Y. Geynet (1999). Development of a knowledge base for the corals of the mascarene archipelago. *Int. Conf. on Scientific Aspects of Coral Reef Assessment, Monitoring and Restoration*.

Grosser, D. (2002). *Construction de bases de connaissances descriptives et classificatoires avec la plate-forme à objets IKBS. Application à la systématique des coraux des Mascareignes*. Ph. D. thesis, Université de la Réunion.

Grosser, D., N. Conruyt, et Y. Geynet (2003). Représentation de connaissances descriptives et classificatoires: le modèle codesc. In *Actes des 9èmes journées francophones "Langages et modèles à objets", Revue Sciences et Technologies de l'information (RSTI), série l'Objet, Hermès (Ed.)*.

Grosser, D., J. Diatta, et N. Conruyt (2000). Improving dissimilarity functions with domain knowledge, applications with ikbs system. In *PKDD*, pp. 409–415.

Kaizhong, Z. et D. Shasha (1989). Simple fast algorithms for the editing distance between trees and related problems. *SIAM J. Comput. 18*(6).

Quinlan, J. (1986). Induction of decision trees. *Machine Learning 1*, 81–106.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Series in Machine Learning.

Renard, J. L., C. Lévi, N. Conruyt, et M. Manago (1996). Sur la représentation et le traitement des connaissances descriptives : une application au domaine des éponges du genre hyalonema. *Bulletin de l'Institut Royal des Sciences 66*, 37–48.

Sokal, R. et P. Sneath (1963). *Principles of numerical taxonomy*. San Francisco et Londres: W.H. Freeman et Cie.

Summary

The formalization of scientific knowledge in life sciences by experts in Biology or Systematics produces arborescent representations whose values could be present, absent or unknown. To improve the robustness of the classification process of those complex objects, often partially described, we propose a new classification method which is iterative, interactive and semi-directed. It combines inductive techniques for the choice of discriminating variables and search for nearest neighbors based on various similarity measures which take into account structures and values of the objects for the neighborhood computation.