Rapport INRIA 1994 — Programme 5

Classification automatique et reconnaissance des formes

PROJET CLOREC

3 mai 1995

PROJET CLOREC

Classification automatique et reconnaissance des formes

Localisation : *Rocquencourt*

1 Composition de l'équipe

Responsable Scientifique

Edwin Diday, professeur, université Paris 9 Dauphine

Responsable Permanent

Yves Lechevallier, directeur de recherche, INRIA

Secrétariat

Martine Cornélis

Personnel INRIA

Sami Bochi, ingénieur de recherche
Christian Désarménien, ingénieur de recherche

Chercheurs extérieurs

Patrice Bertrand, université Paris 9 Dauphine
Marc Csernel, université Paris 9 Dauphine
Gérard Govaert, UTC, Compiègne
Jacques Lebbe, université Paris 6
Yvette Ok, université Paris 9 Dauphine
Régine Vignes, université Paris 6

Chercheurs invités

Francisco de Carvalho, université Fédérale de Pernambuco, Brésil, septembre

Antonio Ciampi, université Mac Gill, Montréal, Canada, janvier-juin

Richard Emilion, université de Dakar, Sénégal, septembre-octobre

Chercheurs doctorants

Mounir Asseraf, bourse du gouvernement marocain
Marie Chavent, boursière MESR
Noël Conruyt, boursier au Muséum de Paris
Younès Hillali, bourse du gouvernement marocain
Vincent Leprince, ingénieur expert, contrat CISIA-INRIA
Edgard Mfoumoune, boursier MESR
Emmanuel Périnel, boursier MESR
Véronique Stephan, boursière MESR
Djamal Ziani, bourse du gouvernement algérien

Stagiaires

Stéphane Baudin, université Paris 13, mars-octobre
Philippe Boursier, université Paris 6, mars-octobre
Ahlame Chouakria, université Paris 9 Dauphine, mars-octobre
Thi Phuong Hoang, université Paris 13, mars-octobre
Thi Thanh Tong, université Paris 9 Dauphine, mars-octobre

2 Présentation du projet

A partir de données expérimentales ou observées, et à partir de connaissances expertes, on cherche à fournir une vue concise et structurée de l'ensemble des objets à traiter, des représentations facilement interprétables par les utilisateurs, des outils d'aide à la découverte de régularités, de règles, de lois et de modèles sous-jacents aux données.

En général cette recherche s'effectue à l'aide d'un ajustement si possible optimal des données à des structures qui peuvent être des partitions, des hiérarchies, des arbres, des treillis, des espaces euclidiens, etc. Le domaine ne se définit pas par une théorie unique qui guiderait le tout mais plutôt par un objectif principal qui est d'extraire des connaissances utiles à partir des données. Ainsi quand le problème se pose sous la forme *données=structure+erreur*, l'ajustement des données est réalisé grâce à

des techniques mathématiques pouvant être très différentes et plus ou moins sophistiquées, selon le type de structure cherchée et selon la nature des données.

Toutes ces recherches s'appuient sur des applications industrielles sollicitées par des organismes tels THOMSON-CSF-RCM, RENAULT, EPSHOM, EDF–DR, ELF, PHILIPS-LCR, etc., avec lesquels nous sommes en relation suivie, par exemple, sous forme de contrats (participation à l'accord cadre avec THOMSON, contrat avec EPSHOM). On s'appuie aussi sur des données médicales (Institut Gustave Roussy) et biologiques (Génome humain, Muséum national d'histoire naturelle de Paris).

Notre équipe participe à la production de systèmes expérimentaux tels MODULAD, SICLA et APYROS. Ces deux derniers étant diffusés par CISIA qui nous tient au courant de l'évolution des besoins.

Les recherches se sont poursuivies cette année selon des directions fortement connectées : on est d'abord amené à représenter des données et des connaissances numériques et (ou) symboliques pour les analyser de façon à obtenir par exemple des classes. Il faut ensuite décrire ces classes et trouver des règles permettant de les distinguer à l'aide d'arbres de décision et réseaux neuronaux. On est ensuite amené à les ordonner, valider et interpréter.

Concernant les diverses actions menées par le projet les faits marquants ont été cette année :

- EUROSTAT : préparation à DOSIS (Development of Statistical Information Services) par la création d'un consortium en vue de l'appel d'offre prévu fin 1994 sur le thème : Prototype de logiciel d'analyse de données symboliques.
- ANVAR-CISIA : seconde phase de ce contrat afin d'aboutir à un système opérationnel de classification pyramidale muni d'outils d'aide à l'interprétation et de validation.
- Conférence internationale d'analyse des données ordinales et symboliques. Lancement de l'appel aux communications de cette conférence organisée les Professeurs Janowitz (USA), Mc Morris (USA), Wille (Allemagne), Barthélémy (France), et dont E. Diday est chairman.

- EPSHOM : Démarrage de ce contrat dont l'objectif est de représenter de façon la plus synthétique possible la situation bathycélérimétrique des océans.
- Informatique et génome : on a entamé la seconde phase de ce contrat en développant une méthode optimale de classification maximalement prédictive pour l'étude des grands génomes.

Les coopérations à l'intérieur de l'INRIA se sont poursuivies cette année par l'encadrement commun de thèses : celle de J.-C. Aude avec J.-J. Codani du projet ICSLA, celles de M. Chavent et V. Stephan en relation avec l'action de J. Lévy-Véhel, celle de V. Leprince avec I.-C. Lerman (REPCO à l'IRISA).

En ce qui concerne l'action génome et le PRC IA, dans lequel participent J. Nicolas (REPCO) et G. Bisson (SHERPA) à Grenoble, une rédaction synthétique est en cours de préparation. Enfin des liens potentiels existent avec tous les projets qui s'intéressent "au raisonnement à partir de cas" tels Psycho-Ergo, Syco ou Secoia.

3 Actions de recherche

3.1 Analyse des données symboliques : vers un prototype de logiciel

Participant : Edwin Diday

Dès qu'il s'agit de décrire des faits définis en intension comme des situations de pannes, des scénarios d'accidents, des espèces de plantes ou d'insectes, des types de radar ou de maladies, on se trouve confronté à des données difficilement intégrables dans le carcan habituel des tableaux de données de l'analyse des données classique. En effet, ces données représentent une expérience d'un niveau de connaissance plus élevé que celle qui est fournie classiquement par de simples observations exprimables par un tableau individus x variables. Dans un tel tableau chaque case contient une valeur unique. Les méthodes d'analyse des données symboliques se dégagent du cadre étroit de tels tableaux en élargissant le champ d'application, au cas où la valeur peut ne pas être unique mais comporter plusieurs valeurs munies de contraintes, de taxonomies et de différents types de connaissances issues du domaine d'application. On modélise de tels individus appelés "objets symboliques" par leur

description exprimant l'intension de la classe qu'ils représentent et une fonction de ressemblance qui permet de calculer leur extension. De tels objets peuvent provenir de sources variées, par exemple, à partir de requêtes d'une base de données, à la suite d'une classification automatique ou directement à partir des connaissances d'un expert. L'analyse des données symboliques s'assigne pour objectif d'étendre la problématique, les méthodes et algorithmes à de tels objets.

Les thèses en cours de E. Périnel, M. Chavent, V. Stephan, V. Le Prince, E. Mfoumoune et D. Ziani vont dans ce sens. Suivant que les données d'entrées ou de sorties sont d'ordre symbolique ou numérique, on conçoit différentes approches. Par exemple, si en entrée on dispose d'un ensemble d'objets symboliques et que l'on désire obtenir en sortie un positionnement graphique des objets les uns par rapport aux autres (par une technique de positionnement multidimensionnel ou *multidimensionnal scaling*), on utilise une mesure de similarité entre objets. Diverses solutions sont proposées pour calculer cette similarité (voir, par exemple, ci-dessous F.A.T. De Carvalho, V. Stephan ou I.-C. Lerman à Rennes). On peut aussi s'inspirer de techniques issues de l'IA comme le "raisonnement à partir de cas" pour obtenir des mesures d'appariement. (Voir, par exemple, Beaubouchais (ANAIS), G. Bisson (Projet SHERPA) à Grenoble ou B. Trousse (Projet SECOIA) à Sophia)).

L'essor nécessaire des méthodes d'analyse des données symboliques résulte de plusieurs facteurs.

- Le besoin, dans les applications réelles, de pouvoir analyser de telles données (en effet, les données réelles sont toujours complexes, et les méthodes d'analyse classique nécessitent de leur appliquer des transformations impliquant une forte perte d'information).
- Le développement de bases de connaissances et des méthodes d'apprentissage symbolique automatique.
- Le développement d'outils informatiques orientés objets facilitant le stockage et la manipulation de structures de données complexes.

Cette année au plan théorique plusieurs résultats nouveaux ont été obtenus concernant entre autres la stabilité de certains opérateurs de type t–norme entre objets symboliques, et le lien avec la théorie des capacités de Choquet a été approfondi avec R. Emilion. Ce type de résultat est nécessaire afin de pouvoir construire des mesures sur des ensembles d'objets symboliques.

Les applications de l'analyse des données symboliques sont multiples et variées par exemple dans le domaine industriel (THOMSON CSF-RCM, Renault) celui des transports (INRETS) ou dans le domaine médical (OMS). Certains programmes commencent à être implantés dans des logiciels standards (SPAD au CISIA, SAS à l'EDF-DR, par exemple). Cependant cette nouvelle approche ne reste accessible qu'à un public restreint par manque d'un outil "grand public" permettant sa mise en œuvre. On a donc travaillé à la conception d'un logiciel prototype intégré d'analyse des données symboliques (pour plus de détails voir §4.1 "actions industrielles" : préparation à l'appel d'offre européen DOSIS).

3.2 Partitionnement récursif et identification

3.2.1 Modèles statistiques et réseaux de neurones

Participants : Antonio Ciampi, Yves Lechevallier

Les réseaux de neurones formels peuvent être considérés comme des modèles statistiques d'une très grande flexibilité, cette flexibilité étant liée au choix de l'architecture, des connexions et des poids. Toutefois, malgré leur flexibilité et leur universalité, les réseaux de neurones ne peuvent se soustraire aux limites intrinsèques de toute modélisation statistique, et plus particulièrement aux limites de l'estimation non-paramétrique, notamment au dilemme biais/variance.

Afin de rendre lisible, en termes de structures, le réseau par des modèles statistiques, nous devons organiser son architecture. Chaque structure détermine un bloc à l'intérieur du réseau, de telle sorte que celui-ci prend la forme d'un réseau de réseaux. Nous proposons deux modèles ; un réseau associé à un modèle additif et un réseau associé à un arbre de régression.

- Un réseau réalise le modèle additif à partir de l'équation :

$$log\left(\frac{p_k}{p_k}\right) = \sum_{j=1}^{m} \beta_j^{(k)} z_j$$

où $z = (z_1, \ldots, z_j, \ldots, z_m)$ est le vecteur de représentation de l'individu, p_k est la probabilité que le vecteur z appartienne à la classe k, $\beta_j^{(k)}$ sont les coefficients de l'équation estimés à partir des données en maximisant la fonction de vraisemblance.

Ce réseau est un réseau à deux couches de connexions ou trois couches de neurones dont une cachée. La première couche représente l'entrée des données. Elle comprend autant de neurones que la dimension du vecteur de représentation de l'individu. La seconde couche est la couche cachée et elle est constituée d'un bloc de neurones par variable. La troisième couche est la couche de sortie, elle est constituée de $k-1$ neurones représentant chacun une classe a priori.

- Un réseau construit à partir d'un arbre de régression.

Cette structure d'arbre permet de mettre en évidence les interactions entre variables surtout s'il s'agit d'interactions complexes. Ce réseau à cinq couches est construit de la même manière qu'un réseau construit à partir d'un arbre de classification.

3.2.2 Sériation et Classification Pyramidale

Participant : Patrice Bertrand

Les méthodes de sériation proposent généralement plusieurs ordres pour classer un ensemble d'objets. Nous avons étudié la détermination du nombre d'ordres qui sont optimaux (et par conséquent équivalents) selon un critère de sériation donné.

Lorsque les données se présentent sous la forme d'une dissimilarité définie sur un ensemble fini Ω, une sériation consiste à trouver un ordre total entre les éléments de Ω qui "respecte au mieux" la dissimilarité. Le cas d'une dissimilarité dont les valeurs augmentent selon un ordre défini sur l'ensemble des objets, peut se traduire mathématiquement, dans un cadre non nécessairement euclidien, par la notion de dissimilarité Robinsonienne. Plus précisément, une dissimilarité d est dite *Robinsonienne* s'il existe au moins un ordre $\ll$ défini sur Ω pour lequel $x \ll y \ll z$ entraine $d(x,z) \geq max\{d(x,y), d(y,z)\}$.

Lorsque cette propriété est vérifiée, on dit aussi que l'ordre $\ll$ est un ordre *compatible* avec la dissimilarité d. Par conséquent, effectuer une sériation revient à déterminer une dissimilarité Robinsonienne dont l'écart à la dissimilarité initiale est minimal. De plus, dénombrer les ordres optimaux déterminés par la sériation, revient à compter tous les ordres (totaux) compatibles avec la dissimilarité Robinsonienne trouvée par la sériation.

L'existence d'une bijection entre l'ensemble des pyramides indicées au sens large et l'ensemble des dissimilarités Robinsoniennes (cf. les travaux de E. Diday), permet d'affirmer que le nombre d'ordres compatibles avec une dissimilarité Robinsonienne d est égal au nombre d'ordres (totaux) compatibles avec la famille d'intervalles formée des classes de la pyramide indicée associée à d. Le problème de la détermination de ce nombre d'ordres s'apparente donc à celui de la reconnaissance d'un graphe d'intervalles (cf. N. Korte et R. H. Mohring 1989[1]) pour une approche utilisant les arbres de composition). La solution que nous proposons, utilise les propriétés de *congruence* des partitions induites par les composantes fortement hiérarchiques de chaque classe. Cette approche permet d'étendre, du cas hiérarchique au cas pyramidal, la formule donnant le nombre d'ordres totaux permettant de visualiser sans croisement une famille d'intervalles.

3.2.3 Classification pyramidale : interprétation

Participants : Patrice Bertrand, Jacques Lebbe, Yves Lechevallier, Vincent Leprince

Dans le cadre du contrat ANVAR, avec le CISIA, nous développons un outil d'aide à l'interprétation des classifications pyramidales. La première approche consiste à adapter les techniques usuelles de l'interprétation numérique au cadre symbolique. Par exemple, on a étendu des critères basés sur l'inertie au cas des variables dont les valeurs sont des intervalles.

Deux autres types d'approches ont également été définies. La première est l'extention des travaux de I.-C. Lerman (IRISA) au cas des pyramides. Le deuxième est la mise au point d'une heuristique qui, par un parcours descendant de la pyramide, détermine les variables dont l'ordonnancement des modalités est le plus compatible avec la répartition des individus dans les classes.

[1] Korte, N. and Mohring, R. H., An incremental linear-time algorithm for recognizing interval graphs. SIAM J. Comput. Vol. 18,1, (1989), pp. 68–81.

3.2.4 Classification incrémentale pyramidale

Participants : Patrice Bertrand, Yves Lechevallier, Edgard Mfoumoune

L'incrémentalité pyramidale consiste à insérer dans une pyramide tout nouvel objet présenté séquentiellement.

Nous avons étudié les problèmes relatifs aux questions fondamentales suivantes : "comment insérer un objet dans une pyramide ?" et, sous l'hypothèse de présentation aléatoire d'objets, "la pyramide change-t-elle lorsque l'ordre de présentation des objets change ?".

Nous utilisons des critères locaux et globaux permettant de réduire la sensibilité de l'adaptateur incrémental à l'ordre de présentation aléatoire des objets. Nous proposons une méthode d'incrémentation des pyramides de complexité linéaire par rapport au nombre d'objets connus. Nous étudions les effets et les facteurs d'instabilité de l'adaptateur, et les méthodes de réorganisation d'une pyramide conceptuelle lorsque la détérioration de l'ordre pyramidal induit est jugée importante.

3.3 Fonctions fondamentales

3.3.1 Extension de la distance de Kolmogorov-Smirnov

Participants : Mounir Asseraf, Yves Lechevallier

Le calcul de la distance de Kolmogorov-Smirnov est basé sur un ordre total ou partiel défini sur les valeurs possibles prises par une variable aléatoire. Cette notion d'ordre n'a pas de sens pour les variables qualitatives. Nous avons introduit une notion ensembliste, la relation d'inclusion, qui va jouer le rôle de la relation d'ordre. Mais le calcul de ce critère a une complexité importante ; cette année nous avons développé des stratégies réduisant cette complexité.

3.3.2 Objets probabilistes et meilleure adéquation

Participants : Marie Chavent, Edwin Diday

Dans le cadre de l'analyse des données symboliques, nous étudions comment calculer l'objet symbolique ayant la meilleure adéquation possible à un ensemble d'objets symboliques particuliers, dits probabilistes.

Pour cela, il faut optimiser un critère mesurant cette adéquation, soit au sens d'une distance, soit au sens d'une similarité entre objets probabilistes, ces mesures dépendant elles-mêmes de la fonction d'agrégation choisie.

Les solutions seront alors des sortes de "prototypes", pouvant être utilisés au sein d'algorithmes de partitionnement ou hiérarchiques d'objets probabilistes.

Pour représenter des classes d'objets symboliques, des objets de fusion pourraient également être calculés grâce à différents types d'unions et d'intersections basées sur la définition de t–normes et t–conormes.

3.3.3 Représentation des connaissances sous forme d'assertions et définition des méthodes de base associées

Participants : Marc Csernel, Edwin Diday, Véronique Stéphan

Les connaissances traitées sont décrites sous la forme d'une conjonction de propriétés dites assertions auxquelles nous associons des mesures de comparaison telles que similarités, dissimilarités et distances.

Nous étudions spécifiquement le cadre où les connaissances traitées sont extraites d'une base de données relationnelle à partir de requêtes. Le résultat d'une requête fournit un ensemble de classes, où chaque classe correspond à l'extension d'une assertion. La création de ces assertions consiste à définir la description qui caractérise chacune des extensions ainsi qu'une fonction d'interprétation qui mesure le degré d'appariement d'un individu à l'assertion traitée.

Dans une seconde étape, nous proposons différents types d'indices de similarité et de dissimilarité entre assertions qui vont dépendre du choix des variables de description et du choix de la sémantique de représentation. Les indices de comparaison entre deux assertions sont calculés à l'aide d'une fonction d'agrégation de l'ensemble des mesures de comparaison sur chaque variable de description. Dans le cas où le calcul de l'extension d'une assertion est possible, nous définissons des mesures de comparaison non plus entre deux intensions mais entre les extensions des deux assertions. Comme précédemment, nous calculons un indice de comparaison sur chaque variable de description. Dans le cas de variables qualitatives, nous pouvons alors adapter à notre problème la famille de mesures classiques calculées à partir de tableaux de contingence.

3.3.4 Arbre de discrimination binaire en Analyse des Données Symboliques

Participants : Jacques Lebbe, Antonio Ciampi, Emmanuel Périnel

Sur la base des travaux développés par A. Ciampi (Partitionnement Récursif et Analyse des Données Symboliques), nous nous sommes intéressés plus particulièrement à la phase de dichotomie d'un nœud pour des données entâchées d'une incertitude de nature probabiliste. A chaque pas, la recherche de la coupure optimale est effectuée à l'aide d'un critère basé sur la notion de vraisemblance : la meilleure coupure est celle générant deux nouveaux nœuds associés à un sous ensemble de l'ensemble d'apprentissage pour lesquels le niveau de vraisemblance associé à la variable à prédire est maximum. Cette phase est envisagée sous deux hypothèses différentes :

- la variable à expliquer est la probabilité d'appartenance d'un individu à chacune des classes mais les valeurs observées sur les prédicteurs sont certaines. Dans ce cas, l'estimation des paramètres définissant les lois de probabilité de la variable à prédire (et par suite, la meilleure coupure) est effectuée de manière locale et indépendante dans chaque sous-population ;
- les prédicteurs sont observés avec incertitude, l'appartenance aux classes étant connue ou non avec certitude. Ici, la recherche de la meilleure coupure peut se plonger dans le cadre général des problèmes de mélange de lois de probabilités, ce qui permet d'envisager plusieurs optiques de résolution (en particulier l'approche estimation proposée par l'algorithme EM).

Cette méthode est actuellement appliquée dans le domaine médical au problème général de la caractérisation du système neuro-endocrinien (aspects prédictifs et descriptifs) dans le cadre d'une collaboration avec le Dr Caillou (service d'histopathologie A de l'Institut Gustave Roussy, Villejuif). Les données concernent l'utilisation de marqueurs de différenciation neuroendocrinienne sur différents types cellulaires et traduisent deux types d'incertitude : des jugements personnels exprimés par le médecin ainsi que des descriptions en variation traduisant l'idée de diversité au sein d'un ensemble d'individus. Une première approche sur un jeu de données simulées a été réalisée puis présentée dans le cadre du congrès IPMU'94.

3.3.5 Réduction du nombre de descripteurs pour la génération automatique de règles

Participants : Régine Vignes, Djamal Ziani

La sélection de variables est un problème traditionnellement abordé en analyse de données classique, cependant très peu de travaux ont été réalisés en analyse de données symboliques, où la sélection ne se fait pas sur les individus simples mais sur des objets symboliques représentant des classes d'individus ou des concepts.

L'algorithme Minset-Plus que nous proposons permet de trouver l'ensemble minimum de variables discriminant les objets en entrée.

Le calcul de la discrimination s'effectue en deux étapes :

- représentation sur un espace à n dimensions (où n est le nombre de variables) de tous les objets symboliques, et ce grâce à leur potentiel de description qui est égal au produit cartésien des valeurs prises par les variables ;
- le calcul de la contribution de chaque variable dans la discrimination des objets, s'effectue en projetant tous les potentiels de description des objets sur l'axe de chaque variable.

Afin d'affiner le calcul de la discrimination et de contrôler la cohérence des variables sélectionnées, l'algorithme traite les dépendances logiques entre variables. Pour ce faire, il utilise un moteur d'inférence d'ordre 0 en chaînage avant qui lui permet de parcourir et de mettre à jour le graphe de dépendances.

3.3.6 Calcul de l'extension des objets probabilistes

Participants : Edwin Diday, Younès Hillali

Dans le cadre des objets probabilistes de l'analyse des données symboliques, nous étudions comment calculer le degré de ressemblance d'un individu par rapport à une classe d'individus décrite par un objet probabiliste. Pour cela nous avons fait appel à la notion de copule (*copulas*) introduite par Sklar et Schweiser[2] qui nous permettra de définir la meilleure fonction d'agrégation donnant le plus d'information sachant

[2]Schweiser, B. and Sklar, A. Probabilistic Metric Spaces, New-York (1983)

que les lois marginales des variables aléatoires décrivant les individus sont connues.

Un premier résultat a été obtenu dans le cas bivarié où nous pouvons définir la fonction d'agrégation sachant que les marginales et les corrélations entre les deux variables aléatoires descriptives sont connues. Par la suite il s'agira, entre autres, de généraliser ce résultat au cas de plusieurs variables aléatoires en utilisant les récents travaux de P. Deheuvels sur les copules.

3.3.7 Système d'aide à la description en biologie

Participants : Noël Conruyt, Edwin Diday

Notre démarche sur la robustesse des systèmes d'aide à la description, à la classification et à la détermination des objets biologiques s'appuie sur l'application de la méthode scientifique en biologie (expérimenter et tester) afin d'aider les naturalistes à mieux comprendre leur domaine, à éprouver leurs opinions et à transmettre leurs connaissances. Nous avons conçu des outils informatiques conviviaux permettant de construire une base de descriptions structurées et pré-classées (les exemples), d'apprendre des hypothèses inductives (les classifications), puis de les mettre à l'épreuve par de nouvelles observations (détermination déductive ou identification). La qualité des descriptions est fondamentale pour l'apprentissage. De plus, elles doivent être comparables entre elles, et reposent donc sur un modèle descriptif que l'expert va explicitement représenter et structurer. Pour l'aider, nous avons dégagé certains mécanismes d'observation à partir de monographies publiées dans la littérature. Le modèle correspond aux objets observables du domaine représenté par un arbre de description. Ensuite, un questionnaire est construit automatiquement à partir du modèle descriptif. Le biologiste utilise celui-ci comme un guide d'observation pour acquérir des descriptions observées et constituer une base de cas cohérente par rapport au modèle. Les cas sont alors traités selon deux technologies complémentaires en fonction de l'objectif poursuivi. Pour la classification, une méthode d'apprentissage inductif permet d'engendrer un arbre de décision caractérisant les classes. Pour la détermination, le raisonnement par cas remplace avantageusement l'induction en partant directement des exemples et en indexant dynamiquement les critères en fonction des réponses de l'utilisateur. Néanmoins, la méthode inductive, tout comme

l'utilisation répétitive du questionnaire, permet de détecter des incohérences éventuelles dans la base de cas, ce qui permet la validation du modèle descriptif. La méthode proposée donne donc à l'expert la possibilité de mettre à jour les connaissances en fonction des résultats (classification et identification) et d'améliorer son modèle descriptif de manière itérative afin de constituer un système d'apprentissage de plus en plus robuste.

4 Actions industrielles

4.1 Préparation à l'appel d'offre DOSIS

Participants : Patrice Bertrand, Marc Csernel, Edwin Diday, Jacques Lebbe, Yves Lechevallier

L'essor des méthodes d'analyse des données symboliques nécessite la mise au point d'un logiciel permettant :

- la gestion des structures de données supportées par les méthodes d'analyse des données symboliques ;
- une acquisition facile des données, soit à partir d'un expert, soit à partir du contenu de bases de données existantes ;
- l'application de méthodes d'analyse à ces données, soit directement par des méthodes d'analyse spécifiques aux données symboliques, soit par des méthodes d'analyse classique, après transformation des données et au travers de couplages avec des logiciels d'analyse de données existants.

Dans cette perspective nous préparons une réponse à l'appel d'offre DOSIS avec des partenaires français (Thomson, EDF), anglais, belges, espagnols et italiens.

4.2 Contrat ANVAR-CISIA : APYROS

Participants : Patrice Bertrand, Marc Csernel, Jacques Lebbe, Yves Lechevallier, Vincent Leprince, Edgard Mfoumoune

En collaboration avec le CISIA et avec le soutien de l'ANVAR, nous terminons actuellement la conception et le développement de la première version d'un logiciel dédié à la classification pyramidale. Le développement de ce logiciel s'est effectué dans l'idée de faciliter non seulement la

tâche d'exploration des données, mais aussi le contrôle de la qualité des résultats obtenus.
L'état d'avancement du développement de ce logiciel diffère selon les programmes qui doivent être réalisés. Les aspects qui ont été plus particulièrement développés au sein du projet sont les suivants :

– *Algorithmes de classification.* Deux algorithmes ont été programmés. Le premier, qui est le plus général, permet de réaliser des classifications ascendantes pyramidales à l'aide des indices d'agrégation classiques : lien du saut maximum, lien du saut minimum, lien moyen. Un deuxième algorithme, plus rapide (sa complexité est en $O(n^2 log(n))$ si n désigne le nombre d'objets à classer) n'est applicable que pour des indices d'agrégation qui sont monotones, comme, par exemple, l'indice du diamètre. Dans la version actuelle, ces deux algorithmes ne fonctionnent qu'avec des données initiales se présentant sous la forme d'un tableau de dissimilarités.

– *Evaluation.* Deux mesures d'adéquation sont proposées : une mesure globale quantifiant l'adéquation de la représentation pyramidale à la dissimilarite initiale, et une mesure locale qui consiste à évaluer l'adéquation de la représentation pyramidale lorsque l'on restreint l'ensemble des objets étudiés à ceux qui forment une classe de la représentation.
Nous proposons deux types d'indices pour évaluer ces mesures d'adéquation. On peut en effet évaluer l'accord entre dissimilarités initiale et induite, soit à l'aide des valeurs numériques des dissimilarités, soit à l'aide des valeurs des préordonnances associées à ces dissimilarités.

– *Condensation.* La représentation pyramidale propose souvent un grand nombre de classes. On peut utiliser l'algorithme de condensation pour réduire le nombre de classes proposées. De multiples utilisations de cet algorithme sont en cours d'implémentation ; citons à titre d'exemple :

  – supprimer les paliers inutilement créés lors de la phase de construction,
  – déterminer une sous-représentation pyramidale, telle que l'adéquation à la dissimilarité initiale soit aussi peu que possible diminuée,
  – déterminer une sous-représentation, telle que le nombre de paliers retenus soit inférieur à une valeur fixée par l'utilisateur.

– *Interprétation.* Contrairement aux hiérarchies, il n'existe pas d'outils standard d'aide à l'interprétation des pyramides. Ce sont les outils de

ce type que nous développons. Parmi les différentes formes d'interprétation d'une structure classificatoire, deux grandes catégories s'appliquent aussi bien aux hiérarchies qu'aux pyramides : d'une part, les méthodes d'interprétation visant à donner une description des classes par rapport à un ensemble de variables, et d'autre part, celles qui interprètent les classes par rapport à la dissimilarite initiale, en indiquant notamment la compacité et l'isolation des classes.
Dans le cas des représentations pyramidales, on peut, de plus, interpréter les classes en fonction du ou des ordres déterminés par la classification. Par exemple, à chaque palier (i.e. classe de la représentation pyramidale) on associe une partition en deux classes : la classe égale au palier, et la classe égale au complémentaire de ce palier dans la population totale. A partir de cette partition nous proposons deux types de critères pour interpréter un palier. Le premier est basé sur la comparaison de préordonnances sur les individus appartenant au palier (évaluation locale). Le second compare les préordonnances sur les individus lorsque l'un des deux individus appartient au palier, et l'autre à son complementaire. L'utilité de ce dernier indice est de permettre la détection du cas où le complémentaire d'une classe s'ordonne naturellement par rapport à la classe. Dans ce cas, la classe joue le rôle de point de repère par rapport auquel se mesure une évolution (ou progression) qui a lieu dans l'ensemble des objets qui n'appartiennent pas à la classe.

En conclusion, le développement de la première version du logiciel APYROS entre actuellement dans sa dernière phase. Cette version se limite, en ce qui concerne l'étape de classification, à la seule prise en compte de données se présentant sous la forme de dissimilarités. Outre les améliorations souhaitables qui émergeront après la phase de tests, une version ultérieure du logiciel devrait permettre d'étendre son champ d'application à d'autres types de données : données du type $individus \times variables$, données se présentant sous la forme *d'un ensemble de classes*.

4.3 Evaluation de l'apport de nouvelles techniques de classification pour la caractérisation des données hydrologiques du CMO

Participants : Sami Bochi, Thi Phuong Hoang, Yves Lechevallier

Le CMO (Centre Militaire d'Océanographie) de l'EPSHOM est chargé d'alimenter les logiciels embarqués des bâtiments de la Marine Nationale

en bases de données d'environnement. Ces bases de données visent à décrire l'environnement bathycélérimétrique de l'océan. L'objectif de cette étude est de représenter de façon la plus synthétique possible la situation bathycélérimétrique des océans. L'étude comprend deux parties : la première sera consacrée à la mise en œuvre des cartes topologiques de Kohonen sur ces profils et une comparaison avec la méthode des Nuées Dynamiques. La deuxième partie sera consacrée, à une évaluation de l'application de ces techniques de classification sur les bathys représentés par des points pertinents, à la prise en compte de l'aspect spatio-temporel dans la classification et aux évolutions possibles vers l'analyse symbolique.

Les cartes topologiques de Kohonen ou cartes auto-organisatrices constituent une famille de méthodes de classification itératives. La règle d'apprentissage proposée par Kohonen se fait en deux étapes :

- Sélection de la classe dont le prototype est le plus proche au sens d'une distance euclidienne du profil présenté.
- Modification des prototypes des classes voisines de cette classe. Cette interaction entre ces classes est souvent mesurée par une fonction décroissante du degré de voisinage entre ces classes et la classe sélectionnée. L'équation de mise à jour des prototypes doit intégrer simultanément cette fonction de voisinage et un pas d'apprentissage décroissant.

Il y a deux phases dans le déroulement de cette procédure.

La phase de structuration topologique, les prototypes s'ordonnent les uns par rapport aux autres sur une grille. La phase de convergence, la grille est maintenant fixée et on utilise une mise à jour de prototypes proche de la méthode des Nuées Dynamiques.

Dans le cas d'environnement complexe, la grille ne se stabilise pas et l'algorithme ne converge pas. Nous avons proposé un contrôle de cette convergence à partir de la fonction d'interaction. Ceci a été vérifié sur les bases de données fournies par le CMO. Nous avons réalisé diverses typologies de la zone 0–400 mètres (zone de surface) et de la zone 0–1800 mètres (zone d'eau profonde) et analysé les liaisons entre ces typologies.

De plus comme les bathys n'ont pas tous la même immersion maximale, nous avons proposé une modification de la méthode des cartes topolo-

giques de Kohonen afin de pouvoir réaliser des typologies sur les profils tronqués.

5 Actions nationales et internationales

5.1 Préparation des actes de IFCS'93

E. Diday, Y. Lechevallier et P. Bertrand sont co-éditeurs des actes de la quatrième conférence de la Fédération Internationale des Sociétés de Classification (IFCS-93) qui a eu lieu à Paris du 31 août au 4 septembre 1993. Ces actes ont été publiés chez Springer-Verlag.

5.2 Participation aux manifestations

E. Diday est président de la conférence internationale sur l'analyse de données ordinales et symboliques qui aura lieu à Paris du 20 au 23 juin 1995. P. Bertrand, J. Lebbe et Y. Lechevallier sont membres du Comité Scientifique local de cette conférence.

E. Diday et Y. Lechevallier ont organisé et ont participé aux Journées Franco-Russes qui se sont déroulées à Saint-Petersbourg en Mai 1994.

E. Diday et Y. Lechevallier ont participé à l'organisation des Secondes Journées de la Société Francophone de Classification qui se sont déroulées à Tours avec 80 participants.

5.3 La Revue de Modulad

« La Revue de Modulad » est le périodique semestriel du club Modulad. Le premier numéro est paru en juin 1988. Treize numéros sont parus à ce jour. D'autre part, la revue contient des informations générales sur les activités du club (cours prévus, actions entreprises, etc.). L. Tricot (CNAM) et Y. Lechevallier ont édité les numéros 12 et 13 et préparent le numéro 14.

5.4 Participation au PRC (Programmes de recherche coordonnés) méthodes symboliques-numériques

Edwin Diday, Jacques Lebbe, Yves Lechevallier, Régine Vignes, E. Périnel et D. Ziani participent à cette action qui réunit 5 laboratoires :

INRIA (Rocquencourt, Rennes), LAFORIA (Paris 6), LIASC (Telecom à Brest), LIRMM (Montpellier) et LRI (Orsay).

L'activité du projet dans ce thème concerne "le formalisme de base" (E. Diday), la "validation et réseaux de neurones en statistiques" (Y. Lechevallier), "les arbes de décision symboliques-numériques" (E. Périnel). La réunion annuelle du PRC s'est déroulée à Marseille du 5 au 7 octobre ; E. Diday, Y. Lechevallier, E. Périnel et D. Ziani y ont participé.

5.5 Invitations de professeurs étrangers

Le projet a une politique d'invitation "large", nationale et internationale. Il s'efforce d'accueillir chaque année un ou deux collègues pour une durée supérieure à deux mois.

Nous avons aussi eu la visite :

dans le cadre d'accords de coopération, de :

Paula Brito, université d'Aveiro, Portugal, août-septembre.
F.A.T. De Carvalho, université de Pernanbuco, Récife, Brésil, mars et septembre.
Mohamed Djedour, USTHB, Alger, Algérie, octobre-novembre.
Saliha Djemai, USTHB, Alger, Algérie, octobre-novembre.

et pour des courtes durées, de :

Richard Emilion, université de Dakar, Sénégal, septembre-octobre.
Antonio Ciampi, université McGill, Montréal, Canada, juin.
Yosu Yurramendi Mendizabal, université de San Sebastian, Espagne, septembre.
Franco Mola, université de Naples, Italie, juin.
Abdallah Mkhadri, université Cadi Ayyad, Marrakech, Maroc, juillet-août
Roberto Araya, université du Chili, Santiago, Chili, juin.
Rosanna Verde, université de Naples, Italie, novembre.
Rudolf Wille, Technische Hochschule Darmstadt, Darmstadt, Allemagne, mars.
Manabu Ichino, université de Tokyo, Tokyo, Japon, septembre.

6 Diffusion des résultats

6.1 Diffusion de produits

6.1.1 Le Club Modulad

Le club Modulad, dont les coordinateurs sont Y. Lechevallier et L. Tricot, a pour but de réaliser et diffuser une bibliothèque de programmes d'analyse de données modulaires portables, mettant en valeur les travaux réalisés en France dans le domaine.

A l'heure actuelle, l'INRIA diffuse la version 2.4 de la bibliothèque Modulad écrite en FORTRAN 77 normalisé. Cette version, qui comporte 42 programmes, couvre de manière assez complète la plupart des champs de la statistique multidimensionnelle :

- Analyse factorielle, analyse factorielle ternaire
- Classification automatique (hiérarchie et partitionnement)
- Modèle linéaire et régression
- Discrimination, segmentation
- Utilitaires de codage et statistiques élémentaires.

Elle s'accompagne d'une brochure de documentation entièrement normalisée de 550 pages. Une centaine d'exemplaires ont été diffusés.

Par ailleurs, Modulad a poursuivi son activité de formation en préparant pour novembre 1995 une école qui se déroulera à Montpellier sur le thème "Aspects statistiques et réseaux de neurones".

6.1.2 SICLA

Diffusion de SICLA

Les nouveaux compilateurs sous DOS en mode texte permettent de fonctionner en 32 bits. Nous avons adapté SICLA à ce nouvel environnement et le distributeur CISIA a disposé de cette nouvelle version.

Le développement d'un éditeur visualisant les résultats des commandes de SICLA a été réalisé sous WINDOWS et avec un environnement 32 bits. Cet environnement est indispensable car nos fichiers de résultats sont volumineux.

6.2 Actions d'enseignement

6.2.1 Enseignement universitaire

P. Bertrand, E. Diday et Y. Lechevallier ont été chargés du cours "Analyse des données et intelligence artificielle" du DEA MAI de l'université Paris IX-Dauphine. J. Lebbe et R. Vignes ont participé à cet enseignement.

E. Diday et Y. Lechevallier ont été chargés du cours "Analyse des connaissances numériques et symboliques" du DEA "Modélisation et traitement des données et des connaissances" de l'université Paris IX-Dauphine. P. Bertrand a participé à cet enseignement.

Y. Lechevallier a été chargé du cours "Méthodes de classement" du DESS MASE de l'université Paris IX-Dauphine.

6.2.2 Séminaires et formation permanente

P. Bertrand a été invité au Séminaire "Mathématiques Discrètes et Sciences Sociales" du CAMS où il a présenté "Ordres compatibles avec une famille d'intervalles".

Y. Lechevallier a été invité à un séminaire au CNAM sur le thème "Classification et discrimination".

6.2.3 Jurys de thèse

E. Diday et J. Lebbe ont été membres du jury de la thèse de N. Conruyt.

Y. Lechevallier a été membre du jury de thèse de E. Pernot et rapporteur de la thèse de I. Pons.

E. Diday a été membre du jury des thèses de I. Pons, de F. Nivelle et de A. Boughalem.

7 Publications

Livres et monographies

[1] E. Diday, Y. Lechevallier, M. Schader, B. Burtschy, P. Bertrand, *Classification and related methods of data analysis*, Springer-Verlag, 1994.

Articles et chapitres de livre

[2] M. Becue, M. Diallo, D. Grangé, L. Haeusler, Y. Lechevallier, Y. Majjad, M. Ringenbach, V. Peres, F. Sermier, «Enquête sur l'utilisation des logiciels de statistique», *Revue de Statistique Appliquée XLII*, 1994, p. 5–12.

[3] G. Caraux, Y. Lechevallier, «Les méthodes statistiques de classement», *Revue d'Intelligence Artificielle*, 1994, À paraître.

[4] E. Diday, «Probabilist, possibilist and belief objects for pattern recognition by data analysis», *Indian J. pure appl. math. 25*, 1994, p. 51–70.

[5] T. Dochy, M. Danech-Pajouh, Y. Lechevallier, «Prévision à court terme du trafic routier par réseau de neurones», *RTS 42*, 1994, p. 35–44.

Communications à des congrès, colloques, etc.

[6] P. Bertrand, «Structural properties of Pyramidal Clustering», *in : Partitioning Data Sets*, I. J. Cox, P. Hansen, B. Julesz (réd.), DIMACS, Rutgers, USA, 1994.

[7] M. Chavent, «Objets probabilistes et fusion», *in : Secondes rencontres de la Société Francophone de Classification*, Tours, septembre 1994.

[8] M. Chavent, «Probabilist objects and fusion», *in : IPMU 94*, Paris, 1994.

[9] A. Ciampi, E. Diday, J. Lebbe, R. Vignes, «Recursive partition and symbolic data analysis», *in : IFCS-93*, Springer-Verlag, Paris, 1994.

[10] A. Ciampi, Y. Lechevallier, «Neural Networks ans Statistical Models», *in : Rencontres Franco-Russe*, Saint-Petersbourg, 1994.

[11] F. De Carvalho, «An approach based on extension to calculate the proximity between Boolean symbolic objects», *in : Collection of abstracts of the 18th Annual Conference of the Germany Classification Society*, Oldenbourg, Germany, mars 1994.

[12] F. De Carvalho, «Proximity coefficients between Boolean symbolic objects», *in : IFCS-93*, Springer-Verlag, Paris, 1994.

[13] F. De Carvalho, «Un indice de dissimilarité entre objets symboliques booléens basé sur l'extension», *in : Actes des 4èmes Journées sur l'Induction Symbolique / Numérique*, Orsay, France, mars 1994.

[14] F. De Carvalho, «Une approche basée sur l'hypervolume pour calculer la dissimilarité entre objets symboliques booléens», *in : Secondes rencontres de la Société Francophone de Classification*, Tours, septembre 1994.

[15] E. DIDAY, «An Overview on symbolic data analysis», *in : IPMU 94*, Paris, 1994.

[16] J. FERRARIS, M. GETTLER-SUMMA, E. PÉRINEL, «Automatic aid to symbolic cluster interpretation», *in : IFCS-93*, Springer Verlag, Paris, 1994.

[17] V. LEPRINCE, J. LEBBE, P. BERTRAND, «Contributions des paliers d'une pyramide à la reconstitution d'une dissimilarité», *in : Secondes rencontres de la Société Francophone de Classification*, Tours, septembre 1994.

[18] V. LEPRINCE, Y. LECHEVALLIER, «Interpretation of Set covering in Pyramids», *in : IPMU 94*, Paris, 1994.

[19] E. PÉRINEL, «Classification tree for probabilistic objects in symbolic data analysis», *in : IPMU 94*, Paris, 1994.

[20] E. PÉRINEL, «Segmentation en analyse des données symboliques», *in : Secondes rencontres de la Société Francophone de Classification*, Tours, septembre 1994.

[21] D. ZIANI, Z. KHALIL, R. VIGNES, «Recherche de sous ensembles minimaux de variables à partir d'objets symboliques», *in : IPMU 94*, Paris, 1994.

8 Abstract

Starting from large sets of experimental or observed and numerical or symbolic data, our aim is to analyze and synthetize such sets by finding easy to interpret representations, and by providing new tools for discovering regularities, rules, laws and models underlying the data. The aim of the symbolic approach in data Analysis which is one of our main topics is to extend problems, methods and algorithms used on standard data to more complex data, more able to represent knowledge, where the units are called "symbolic objects", in order to distinguish them from objects (described by numerical or categorical variables) treated by standard Data Analysis methods.

For instance, in this way, Exploratory Data Analysis is transformed into Exploratory Knowledge Analysis. In this framework, we have focused this year, on four directions :

- classification and orders,
- decision trees and identification,

- basic choices (distances, fusion operators in the case of probabilist, possibilist or belief objects, etc.).

Our team has been involved in several international and national actions such as ESPRIT 2154 ("Machine Learning Toolbox" with several European teams (GMD), Turing Institute, Aberdeen University, British Aerospace, Alcatel-Alsthom, etc.), ANVAR-CISIA in order to implement a computer system on Pyramidal symbolic and numerical classification, and "Informatic and Genome" with REPCO (IRISA).

Methodological advances are regularly implemented in two systems : MODULAD developed with a national consortium (CNET, INRA, EDF, RENAULT, INRETS, etc.) and SICLA which is our "Interactive system of Classification" distributed by CISIA.

The result concern various application domains : road accidents (INRETS), species of insectes (OMS), power plant (EDF), genoma (GIP-GREG), radars (Thomson CSF-REM), comfort of vehicle (Renault), etc.

Table des matières