Equipe de Recherche en Ingénierie des Connaissances

Bilan – Projet

Décembre 2001

Université Lumière Lyon 2

5, avenue Pierre Mendès-France

69676 Bron

Tél. : (33) [0]4 78 77 23 76 - Fax. : (33) [0]4 78 77 23 75

e-mail. zighed@univ-lyon2.fr

Web http://eric.univ-lyon2.fr

TABLE DES MATIERES

1 BILAN SYNTHETIQUE 5

1.1 POSITIONNEMENT SCIENTIFIQUE DE ERIC 5

1.2 BILAN CHIFFRE 1998-2001 6

1.2.1 Production scientifique 6

1.2.2 Bilan financier 7

1.2.3 Liste des Membres permanents 8

1.2.4 Ressources Informatiques 10

2 DOSSIER SCIENTIFIQUE 12

2.1 BILAN SCIENTIFIQUE DES 4 ANNEES PRECEDENTES 12

2.1.1 Déclaration générale de politique scientifique 12

2.1.2 Description des travaux du pôle Extraction de connaissances 15

2.1.3 Description des travaux du pôle Images 31

2.1.4 Description des travaux du pôle Connexionisme 43

2.1.5 Utilisation des crédits des 4 années précédentes 50

2.2 BILAN QUANTITATIF SUR LES QUATRE DERNIERES ANNEES 51

2.2.1 Les publications et communications majeures 1998-2001 51

2.2.2 Communications avec actes 53

2.2.3 Conférences invitées dans les congrès internationaux 59

2.2.4 Autres publications 59

2.2.5 Les activités internationales 62

2.2.6 Les contrats de recherche 68

2.2.7 Les brevets licenciés 75

2.2.8 La valorisation et le partenariat industriel 75

2.2.9 L'information scientifique et la vulgarisation scientifique *77*

2.3 DECLARATION DE POLITIQUE SCIENTIFIQUE POUR LA PERIODE 2003-2006 81

2.3.1 Prospective – Pôle apprentissage *81*

2.3.2 Prospective – Pôle Images *83*

2.3.3 Prospective – Pôle Bases de données décisionnelles *87*

ANNEXE : SUJETS DE THESES EN COURS 90

1 BILAN SYNTHETIQUE

1.1 POSITIONNEMENT SCIENTIFIQUE DE ERIC

Créé en 1995, le laboratoire ERIC a pour objectifs scientifiques le développement de méthodes et d'outils informatiques pour l'ingénierie des connaissances destinés, plus particulièrement, à l'Extraction automatique des Connaissances à partir des Données (ECD).

Les données traitées sont multiformes (textes en langage naturel, images, données numériques quantitatives ou qualitatives?) et peuvent se situer sur des supports structurés comme les bases de données ou non structurés comme le web. Elles peuvent provenir de sources et de domaines diverses : médecine, archéologie, économie, chimie, enquêtes satisfaction client, etc.

Les problèmes auxquels nous sommes confrontés en ECD couvrent un large spectre qui s'étend du stockage dans les entrepôts de données à l'organisation et à l'accès à ce large volume d'informations, en passant par les méthodes de fouille de données (généralement appelées techniques de Data mining), jusqu'aux méthodes de validation des résultats et de mise en forme des connaissances produites.

Le laboratoire ERIC mène ses recherches dans ce cadre fédérateur selon trois formes :

- Recherches Théorique : Il s'agit de mener des réflexions sur les concepts et les méthodes employées en ECD. Par exemple, la caractérisation d'un bon espace de représentation en apprentissage, l'étude des propriétés des mesures de similarité sur des images, la validation et/ou l'agrégation des modèles de prédiction, etc.

- Développement de prototype Logiciel : Il s'agit de développer les outils informatiques issus des travaux théoriques soit dans un but expérimental, s'il faut comparer des méthodes sur des benchmarks, soit dans un but d'application à des cas pratiques. Nous venons de lancer un projet ambitieux, baptisé « Smart », dont l'objectif est de proposer un environnement logiciel permettant d'extraire des connaissances à partir des données. Cette plate-forme devra supporter tous types de données, fonctionner sur tous types de système, être accessible comme serveur d'applications en ECD, être évolutive et capable de recevoir de nouveaux programmes, etc.

- Recherches Appliquées : Il s'agit de mettre en oeuvre nos idées et d'utiliser, sur des cas concrets, les logiciels que nous avons développés. Ce volet nous a assuré un contact permanent avec les milieux professionnels : médecine, banques, assurances, industrie, institutions gouvernementales, etc.

1.2 BILAN CHIFFRE 1998-2001

1.2.1 Production scientifique

Le tableau ci-dessous résume le bilan quantitatif cumulé sur la période 1998-2001.

		1998	1999	2000	2001	Total
Publications	Ouvrages, chapitres dans ouvrages	6	3	2	2	**13**
	Revues internationales	4	4	1	2	**11**
	Revues nationales	3	2	0	5	**10**
	Conférences internationales	9	13	13	10	**45**
	Conférences nationales	1	9	5	4	**19**
	Total	**23**	**31**	**21**	**23**	**98**
Thèses soutenues		2	0	2	3	**7**[1]
Habilitation					2[2]	**2**

[1] 5 soutenances supplémentaires sont prévues pour 2002.

soutenues						
Formation par la recherche	Effectifs DEA ECD	-	14	17	29	**60**

Tableau 1-1 Résultats académiques 1998-2001

1.2.2 Bilan financier

Le tableau ci-dessous résume en termes de contrats de rechercher sur la même période 1998-2001.

Partenaires	**Période**	**Durée**
France Télecom	01-02	2 ans
VédiorBis	01-03	3 ans
Clinique mutualiste «La Roseraie»	98-99	2 ans
BOTANIC	98	6 mois
CSTI	98	2 ans
RICOH France	98-00	3 ans
ELF ANTAR 1	99	6 mois
Observatoire National du Tourisme	99	6 mois
Santé et HPC, Région Rhône-Alpes	97-99	3 ans
CORA	99	12 mois
INRIA	99	2 ans
Région Rhône-Alpes	99-01	3 ans
CCF	99-01	6 mois
Crédit Agricole du Centre-Est	99-01	18 mois
APRIL-Assurance	99-01	12 mois
Messageries Lyonnaises de Presse	00	6 mois
Ville de Bron	00	6 mois
Institut FOURNIER et Associés	00	6 mois
PKDD	00	12 mois
ADEMO	99-00	2 ans
Diagnos	00-01	14 mois

[2] Soutenance prévue : le 4 janvier 2002.

Tableau 1-2 Contrats de recherche 1998-2001

Ressources financières 1988-2001	
Contrats de recherche	3MF
Contrats de formation	30 KF
BQR	36 KF
Manifestations diverses	600 KF

Tableau 1-3 Ressources financières

1.2.3 Liste des Membres permanents

Personnel administratif et techniques (ITA/IATOS) **Total = 1.5**

- Gabrièle Valérie — IATOS (mi-temps)
- Varaine Astrid — Fonds propres

Professeurs **Total = 4 ⇒ 3**

- Miguet Serge — PR1 — directeur adjoint de l'ERIC
- Nicoloyannis Nicolas — PR2
- Paugam-Moisy Hélène — PR2 — a quitté le laboratoire en novembre 2000
- Zighed Abdelkader — PR1 — directeur de l'ERIC

Maître de conférences **Total = 11**

- Bentayeb Fadila — MCF2
- Boussaid Omar — MCF1
- Chauchat Jean-Hugues — MCF0
- Darmont Jérôme — MCF2
- Lallich Stéphane — MCF1
- Rabaséda Sabine — MCF2
- Rakotomalala Ricco — MCF2
- Sarrut David — MCF2
- Tougne Laure — MCF2

- Viallefont Anne MCF2
- Viallaneix Jacques MCF2

ATER **Total = 5**

- Bentayeb Fadila 1999-2000
- Bonhomme Christine 1999-2000
- Quignard Matthieu 1999-2000
- Puzenat Didier 1998-1999
- Jalam Radwan 2000-2002

Doctorants **Total = 15**

- Tweed Tiffany 2001 (S. Miguet) mi-temps Entreprise
- Clech Jérémy 2001 (D. Zighed) mi-temps Entreprise
- Coeurjolly David 2001 (L. Tougne, S. Miguet) MENRT
- Clippe Sébastien 2000 (D. Sarrut, S. Miguet) mi-temps assist. chef de clinique
- Jouve Pierre 2000 (N. Nicoloyannis) REGION
- Scuturicci Marian 1998 (S. Miguet) CNOUS
- Scuturicci Mihaella 1998 (Pinon J.M., S. Miguet) CNOUS
- Baumes Laurent 1999 (N. Nicoloyannis) BDI CNRS
- Agier Céline 1999 (R. Rakotomalala, D. Zighed) MNERT
- Teytaud Olivier 1999 (H. Moisy) MENRT
- Muhlenbach Fabrice 1998 (D. Zighed) MENRT
- Jalam Radwan 1998 (JH. Chauchat, D. Zighed) ATER
- Di Palma Serge 1997 (D. Zighed) Entreprise
- Erray Walid 2001 (D.A. Zighed) France-Telecom
- Legrand Gaëlle 2001 (N. Nicoloyannis) MNERT

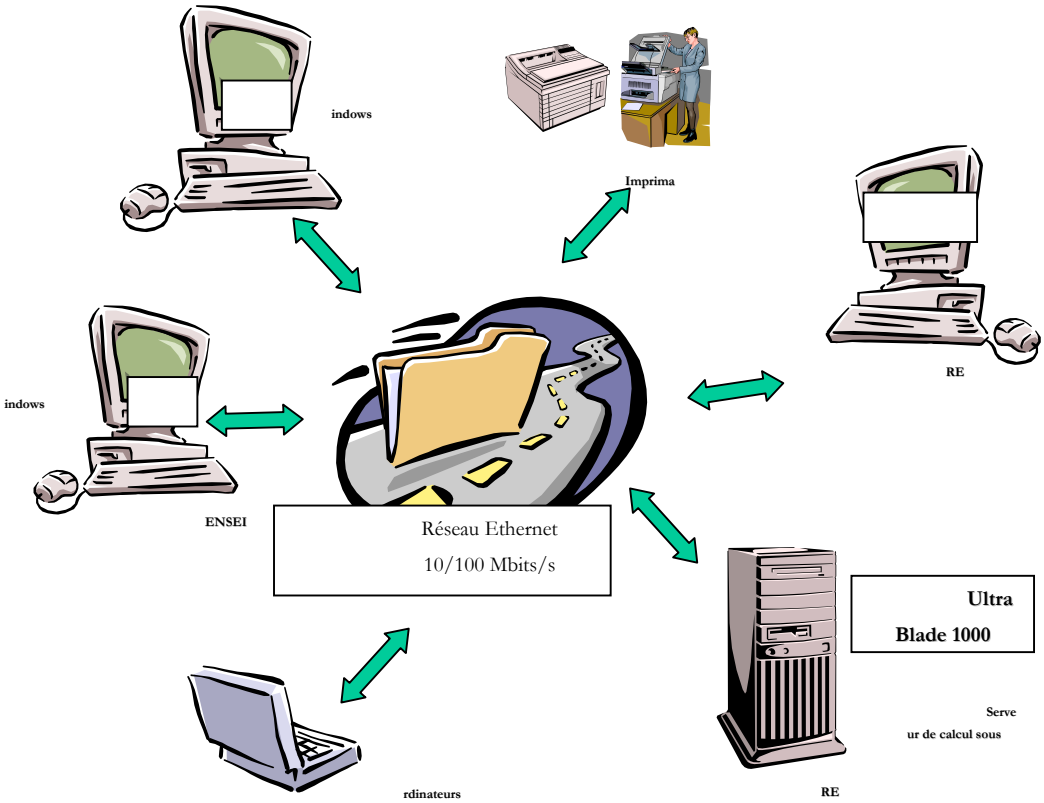
Thèses soutenues du 1/01/99 au 30/06/2001 **Total = 7**

- Scuturicci Marian 2001 (S. Miguet) soutenance le 4 décembre
- Olivier Teytaud 2001 (H. Moisy) soutenance fin décembre
- Gavin Gérald 2001 (D. Zighed) ⇒ MCF Lyon 1

- Elisseef André 2000 (H. Moisy) ⇒ secteur privé - USA
- Sarrut David 2000 (S. Miguet) ⇒ MCF Lyon 2
- Feschet Fabien 1998 (S. Miguet) ⇒ MCF Lyon 1
- Sylvain Contassot 1998 (S. Miguet) ⇒ MCF Monbelliard

1.2.4 Ressources Informatiques

Le schéma suivant illustre les ressources informatiques du laboratoire ERIC.



2 DOSSIER SCIENTIFIQUE

2.1 BILAN SCIENTIFIQUE DES 4 ANNEES PRECEDENTES

2.1.1 Déclaration générale de politique scientifique

2.1.1.1 Place de ERIC à Lyon 2

Il n'est plus à démontrer que la modélisation informatique, et de manière générale l'informatique, est devenue plus qu'un outil mais un cadre méthodologique pour traiter les problèmes qui se posent au psychologue quand il cherche à comprendre les mécanismes cognitifs, au sociologue quand il analyse le comportement d'un groupe social, au géographe quand il souhaite reconstruire des reliefs en image de synthèse, à l'archéologue quand il veut identifier et dater des vestiges du passé, etc. Cette prise de conscience est quasi générale en recherche comme en enseignement dans les sciences humaines et sociales.

La création à Lyon 2 du DEUG et depuis 2000 de la licence «Mathématiques Informatique et Statistiques Appliquées aux Sciences Humaines et Sociales» (MISASHS), du second et du troisième cycles de «Sciences Cognitives», de l'IUP Informatique, Statistique et Econométrie Appliquée, du DEA Extraction des Connaissances à partir des Données et l'existence de la filière d'enseignement «Mathématique Appliquées aux Sciences Sociales» (MASS), la présence de trois DESS d'informatique (Organisation et Protection des Systèmes d'Information dans l'Entreprise (OPSIE), Ingénierie Informatique pour l'aide à la Décision et l'Evaluation Economique (IIDEE) et Statistique et Informatique Socio-Economique (SISE)), la

mise sur pied, à la rentrée 1998-99, du département Statistique et Traitement Informatique des Données de l'IUT Lumière, le développement de la filière Infographie (deux diplômes universitaires de 2ème et 3ème cycles et un DESS) sont une parfaite illustration de cette forte implication des collègues informaticiens et mathématiciens dans l'environnement spécifique des Sciences Humaines et Sociales que représente Lyon 2.

Les récentes créations de postes : deux professeurs et six maîtres de conférences depuis la rentrée 1998-99 témoignent une fois de plus de la dynamique que connaît l'Université Lumière Lyon 2 en informatique. Les besoins en recherche comme en enseignement généreront encore inévitablement de nouvelles créations de postes. ERIC devra donc gérer de manière cohérente et harmonieuse sa croissance. Pour cela, nous souhaitons :

- ✓ maintenir autant que possible une unité de lieu enseignement-recherche ;
- ✓ renforcer les thématiques sur lesquelles ERIC est reconnue ;
- ✓ maintenir une activité de recherche à trois niveaux : travaux à caractère théoriques, développement d'outils informatiques associés, recherche de terrains d'application en particulier dans les domaines des Sciences Humaines et Sociales ;
- ✓ renforcer la politique éditoriale et d'animation scientifique de l'équipe ;
- ✓ développer nos relations avec la communauté scientifique locale, nationale et internationale ;
- ✓ valoriser nos recherches sur le plan industriel.

2.1.1.2 Positionnement scientifique de ERIC

Créé en 1995, le laboratoire ERIC a pour objectifs scientifiques le développement de méthodes et d'outils informatiques pour l'ingénierie des connaissances destinés, plus particulièrement, à l'Extraction automatique des Connaissances à partir des Données (ECD).

Les données traitées sont multiformes (textes en langage naturel, images, données numériques quantitatives ou qualitatives…) et peuvent se situer sur des supports structurés comme les bases de données ou non structurés comme le web. Elles peuvent provenir de sources et de domaines diverses : médecine, archéologie, économie, chimie, enquêtes satisfaction client, etc.

Les problèmes auxquels nous sommes confrontés en ECD couvrent un large spectre qui s'étend du stockage dans les entrepôts de données à l'organisation et à l'accès à ce large volume d'informations, en passant par les méthodes de fouille de données (généralement appelées techniques de *data mining*), jusqu'aux méthodes de validation des résultats et de mise en forme des connaissances produites.

Le laboratoire ERIC mène ses recherches dans ce cadre fédérateur selon trois formes :

- **recherches théoriques** : il s'agit de mener des réflexions sur les concepts et les méthodes employées en ECD. Par exemple, la caractérisation d'un bon espace de représentation en apprentissage, l'étude des propriétés des mesures de similarité sur des images, la validation et / ou l'agrégation des modèles de prédiction, etc. ;
- **développement de prototypes logiciels** : il s'agit de développer les outils informatiques issus des travaux théoriques soit dans un but expérimental, s'il faut comparer des méthodes sur des benchmarks, soit dans un but d'application à des cas pratiques. Nous venons de lancer un projet ambitieux, baptisé « Smart », dont l'objectif est de proposer un environnement logiciel permettant d'extraire des connaissances à partir des données. Cette plate-forme devra supporter tout type de données, fonctionner sur tout type de système, être accessible comme serveur d'applications en ECD, être évolutive et capable de recevoir de nouveaux programmes, etc. ;
- **recherches appliquées** : il s'agit de mettre en oeuvre nos idées et d'utiliser, sur des cas concrets, les logiciels que nous avons développés. Ce volet nous a assuré un contact permanent avec les milieux professionnels : médecine, banques, assurances, industrie, institutions gouvernementales, etc.

2.1.1.3 Commentaire sur le bilan scientifique

Le laboratoire a fédéré son potentiel de compétences autour de trois pôles de recherche essentiels : extraction des connaissances, images, et connexionisme.

Le bilan scientifique de ces quatre dernières années que nous détaillons plus loin appelle quelques commentaires et observations.

La thématique de recherche du laboratoire a évolué de l'apprentissage numérique symbolique vers l'extraction automatique des connaissances à partir de grandes bases de données hétérogènes (numériques, textuelles et images). Cette évolution s'est traduite par un triple élargissement de nos réflexions :

- vers les problèmes qui se situent d'une part, en amont de l'apprentissage (représentation des données, construction d'attributs, caractérisation de la séparabilité des classes, échantillonnage optimal,…) et, d'autre part, en aval de l'apprentissage à travers les problèmes liés à la validation des modèles, leur simplification, leur agrégation,…
- vers de nouveaux supports de données : les images et le texte. Les données images et les données textuelles (en langage naturel) sont particulièrement abondantes. Leur pouvoir

sémantique fait qu'elles sont l'un des supports d'information que l'homme manipule le plus pour analyser, comprendre, décider. Développer une expertise pour extraire des connaissances à partir de ces types de données nous paraît être un enjeu stratégique.

- vers une recherche sur les systèmes complexes « intelligents » en collaboration avec les sciences cognitives. Cette thématique s'est développée pendant deux ans (1998-2000) mais a due être interrompue. Notre collègue, le Professeur Hélène Paugam-Moisy responsable de ce pôle, a, pour des raisons d'efficacité, réuni toutes ses recherches au sein de l'Institut des Sciences Cognitives de Lyon auquel elle était déjà rattachée à temps partiel.

Quantitativement, le bilan des quatre années d'ERIC, est appréciable :

- cinq chercheurs ont pu réaliser ou achever leur doctorat au sein de ERIC.
- ERIC continue d'attirer de jeunes doctorants titulaires d'allocation de recherche ministérielle et de bourses de l'industrie, 13 en cours.
- La qualité et la diversité des publications de ERIC montre que l'équipe est active et présente à tous les niveaux : publications internationales dans des revues, publications d'ouvrages et de monographies, publication nationales, diffusion de logiciels, organisations de conférences majeures, contacts avec des universités étrangères,…
- L'expertise développée par les chercheurs de ERIC a atteint le monde scientifique et industriel comme en témoignent les nombreux contrats.

2.1.2 Description des travaux du pôle Extraction de connaissances

2.1.2.1 Aspects méthodologiques

2.1.2.1.1 Extraction des connaissances et Apprentissage

Dans un cadre méthodologique général, qui dépasse la simple construction d'une fonction de classement, nous travaillons sur des sujets complémentaires qui partent de la définition d'un test de séparabilité des classes dans R^p, condition nécessaire avant de lancer tout algorithme d'apprentissage, jusqu'à la validation statistique permettant de s'assurer que les « formes » mises en évidence sur un échantillon correspondent réellement à des structures présentes dans la population.

a. Modèles géométriques et Apprentissage

(**Participants** : P. Jouve, S. Lallich, F. Muhlenbach, N. Nicoloyannis, D.A. Zighed)

Dans les travaux qui ont été regroupés dans les thèses de Samia Amghar[3] et Marc Sebban[4], nous avons introduit le concept de « réseaux de cooptation ». Celui-ci est formé de l'ensemble des points de l'échantillon d'apprentissage reliés entre eux par une propriété de voisinage. Parmi les modèles de voisinage que nous avons étudié, on peut citer les graphes de Delaunay, les graphes de Gabriel, les graphes des voisins relatifs et l'arbre de recouvrement minimal. Tous ces modèles permettent d'engendrer un graphe de voisinage connexe. Ce cadre nous a permis de déboucher sur deux résultats significatifs :

- Une nouvelle approche de la séparabilité des classes. Dans ce contexte nous avons mis au point un test statistique non paramétrique de séparabilité des classes. Le principe, relativement simple, consiste à étudier la statistique du nombre d'arêtes qu'il faut supprimer du graphe de voisinage afin d'obtenir des composantes connexes formées par des points qui appartiennent à la même classe. Ce test dépasse les limites inhérentes aux tests paramétriques telles que le lambda de Wilks qui repose sur des hypothèses rarement vérifiées (distribution multi-normale, égalité des paramètres d'échelle…). Des résultats établis dans le domaine de l'analyse statistique spatiale[5] nous ont permis d'unifier nos travaux et surtout d'ouvrir de nouvelles pistes en cours d'exploration. Nous travaillons sur une nouvelle formulation du test et de nouvelles extension vers des espaces non métriques.[mettre la référence de l'article de la SFC ?]
- De nouvelles règles de classement. L'approche par les graphes de voisinage nous a permis d'élaborer de nouvelles fonctions de classement proches des méthodes d'estimations locales mais s'affranchissant du paramétrage toujours difficile dans les stratégies telles que les k-plus-proches-voisins (nombre de voisins) ou les fenêtres de Parzen (taille de la fenêtre). Nous avons introduit une première stratégie baptisée « vote universel » qui a été ensuite affinée par le vote probabiliste. Ce dernier aborde la question sous un angle probabiliste : il propose une estimation de la probabilité que deux individus voisins (reliés par une arête) restent voisin à l'arrivée d'un nouvel individu. Ces résultats, produits dans le cadre des deux thèses citées ci-dessus, sont à nouveau repris dans le projet «Mammo» qui sera exposé plus loin.

---

[3] AMGHAR S. *Nouvelles propositions pour le classement : Le Vote Universel (VU) et le modèle Logit-Probit à travers une formalisation prétopologique.* Thèse de doctorat, Université Claude Bernard - Lyon 1 (1995).

[4] SEBBAN M. *Modèles théoriques en Reconnaissance de Formes et Architecture Hybride pour Machine Perceptive.* Thèse de doctorat, Université Claude Bernard - Lyon 1 (1996).

L'exploitation des approches géométriques se poursuit également pour générer des règles de décisions similaires à celles que donneraient des arbres de décision mais qui en diffèrent par le fait que les variables de segmentation sont généralement des combinaisons de variables originelles.

b. Apprentissage PAC et Agrégation de prédicteurs

(**Participants** : G. Gavin, D.A. Zighed)

Au cours des années 1990 de nouveaux principes d'induction sont apparus : les méthodes d'agrégation à base de vote. Les exemples les plus connus sont certainement le « Bagging » initié par Breiman[6] et le Boosting initié par Freund[7]. Ces méthodes présentent d'excellentes performances sur les problèmes réels. Leur simplicité en font de bons candidats pour résoudre les problèmes de data-mining. Elles reposent en outre sur de solides fondations théoriques inspirées de l'apprentissage Probablement Approximativement Correct (PAC) et plus précisément des travaux de Vapnik[8]. Ce travail qui a aboutit à la thèse de Gérald Gavin **[C5]** a porté sur deux aspects.

Le premier **[F5, F15, F16, F39, F42]** consiste en une étude approfondie de l'apprentissage PAC et de la théorie de Vapnik. Cette théorie permet de contrôler l'erreur réelle à partir de l'erreur empirique pour tout algorithme d'apprentissage sans faire aucune hypothèse sur le problème à traiter. Le théorème de Vapnik et Chervonenkis est un résultat issu de cette théorie. Il propose de majorer le nombre d'exemples requis pour que l'erreur empirique soit un bon estimateur de l'erreur réelle au vu de deux critères : l'un de précision et l'autre de fiabilité. Des minorations sur ce nombre d'exemples existent et montrent l'optimalité à une constante multiplicative près de la borne supérieure fournie par le théorème de Vapnik et Chervonenkis. Nous avons obtenu un résultat permettant d'améliorer cette constante. Ce résultat ne permet cependant pas d'expliquer le comportement des algorithmes d'agrégation à base de vote. De nouveaux résultats, utilisant une nouvelle notion, appelée « marge » parait beaucoup plus adaptée.

---

[5] CLIFF A.D., ORD J.K. *Spatial processes, models and applications*. Pion limited, Londres (1981).

[6] BREIMAN L. *Bagging predictors*. Machine Learning, 2(24):123-140 (1996).

[7] FREUND Y., SHAPIRE R. *A short introduction to boosting*. Journal of the Japanese Society for Artificial Intelligence, 14(5):771-780 (1999).

FREUND Y., SCHAPIRE R. *Experiments with a new boosting algorithm*. Proc. 13th International Conference on Machine Learning, pp. 148-146, Morgan Kaufmann, (1996).

FREUND Y., SCHAPIRE R. *A decision-theoric generalization of on-line learning and an application to boosting*. Journal of Computer and System Sciences, 55(1):119-139 (1997).

FREUND Y., SCHAPIRE R. *Discussion of the paper "arcing classiers" by Breiman*. Technical report, ATT labs (1997).

[8] VAPNIK V., CHERVONENKIS A. *Theory of Pattern Recognition*. Nauka, Moscow (1974).

Cette notion de « marge » a inspiré de nouveaux algorithmes d'agrégation à base de vote mettant en lumière l'interaction forte entre la théorie et la pratique.

Le second volet de cette recherche est justement l'étude des algorithmes à base de vote. Nous avons proposé d'adapter ces méthodes à deux problématiques du *data mining* : le traitement des grandes bases de données, et le traitement des valeurs manquantes. Cette deuxième partie reste encore à poursuivre.

c. Discrétisation bi-dimensionnelle et arbres de segmentation généralisés

(**Participants** : W. Erray, N. Nicoloyannis, G. Ritschard, D.A. Zighed)

L'étude de l'association entre variables catégorielles repose en général sur une analyse du tableau croisé des variables concernées. Pour saisir l'intensité de l'association on se réfère à des mesures d'association telles que le v de Cramer, le τ de Goodman et Kruskal, le coefficient d'incertitude de Theil, le λ de Kruskal,… La question abordée dans cette recherche est celle de l'effet de l'agrégation de lignes ou colonnes d'un tableau croisé sur les mesures d'association.

La maximisation du degré d'association trouve sa motivation dans la discrétisation d'attributs continus qui est un problème majeur en apprentissage automatique et en recherche de règles d'associations. Dans le cas de l'apprentissage supervisé, les solutions optimales par rapport à la discrimination recherchée, répertoriées par exemple dans **[A3],** procèdent individuellement pour chaque variable. Dans une optique prédictive, des solutions bidimensionnelles, où l'on discrétise conjointement deux variables, devraient s'avérer plus efficaces. Breiman[9] et al. ont étudié le cas de la dichotomisation conjointe de deux variables. Notre propos est de généraliser ce cas à un nombre quelconque de catégories **[B9, E10].**

Hormis les cas triviaux avec un nombre initial réduit de catégories (trois par exemple), rechercher la solution optimale par recombinaison systématique ne peut être envisagé à cause du nombre exponentiel de cas à considérer. Il s'agit alors de formuler une procédure qui puisse être exploitée algorithmiquement pour trouver les regroupements conjoints des catégories lignes et colonnes qui (quasi-)maximisent un critère donné d'association. Notons qu'il conviendra ici de distinguer les cas des variables nominales de celui des variables ordinales pour lesquelles seuls des regroupements de catégories adjacentes ont un sens.

L'optimisation considérée n'est pas sans rappeler la question des partitions des lignes et des colonnes qui maximisent le χ^2 de Pearson, abordée par Benzécri[10], et traitée en particulier par

[9] BREIMAN L., FRIEDMAN J., OLSHEN R., STONE C. *Classification and Regression Trees*. Wadsworth International, California (1984).

[10] BENZECRI J.-P. *Analyse des données. Tome 2 : analyse des correspondances*. Dunod, Paris (1993).

des techniques de classification automatique par Celeux[11] et al. Ces auteurs se placent cependant dans un contexte où, contrairement à celui retenu ici, seules des partitions avec un nombre de classes fixé a priori sont pertinentes.

Dans cette recherche engagée, il y a seulement 2 ans, en collaboration avec le Professeur G. Ritschard de l'Université de Genève, nous avons obtenu des résultats fort intéressants **[F32, G14, F36, B9]**. Ce travail, encore en cours, connaît un développement assez spectaculaire dans le cadre d'un mémoire de DEA ECD 2001 (Walid Erray), où ce principe de la décomposition optimale d'un tableau est utilisé pour construire des arbres d'induction. A chaque nœud de l'arbre nous recherchons la meilleure décomposition du tableau de contingence croisant la variable endogène à une variable exogène. Ceci est particulièrement intéressant car les variables endogènes peuvent être continues. Des comparaisons sont en cours avec les arbres de régression proposés dans le livre «CART»[12].

d. Extraction de règles d'associations : approches mono et polythétique

(**Participants** : C. Agier, P. Jouve, G. Legrand F. Muhlenbach, N. Nicoloyannis, R. Rakotomlala D.A. Zighed)

L'extraction de connaissances sous forme de règles d'associations est apparue dans les années 90. Provenant de la recherche sur les bases de données, les premières méthodes proposées, malgré leur extrême simplicité et leurs limites, connurent un grand succès et elles sont maintenant implémentées dans de nombreux logiciels de *data mining*. Une de leurs principales limites est la complexité algorithmique. En effet, ces méthodes imposent un mode de représentation sous forme d'attributs dont les valeurs sont binaires et la recherche des règles d'association repose sur un parcours quasi exhaustif du treillis des parties. Les conséquences de ce mode opératoire sont souvent peu exploitables par les utilisateurs, car les algorithmes proposés jusque là comme celui d'Agrawall[13] génèrent de nombreuses règles, parfois plus nombreuses que le cardinal de l'ensemble d'apprentissage lui même. L'usager obtient donc une masse importante d'information, difficilement interprétable, et dont la validité, de surcroît, ne peut être que douteuse.

---

[11] CELEUX G., DIDAY E., GOVAERT G., LECHEVALLIER Y., RALAMBONDRAINY H. *Classification automatique des données*. Informatique, Dunod, Paris (1998).

[12] BREIMAN L., FRIEDMAN J., OLSHEN R., STONE C. *Classification and Regression Trees*. Wadsworth International, California (1984).

[13] AGRAWALL R.,MANNILA H., SRIKANT R., TOIVONEN H., VERKAMO A.I. *Fast discovery of association rules*, In "Advances in knowledge discovery and data mining", Fayyad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. (Eds). pages 307-328. AAAI/MIT Press (1996).

Un effort de formalisation a été entrepris par de nombreux chercheurs. Deux voies sont explorées au laboratoire ERIC.

La première, impliquant D.A.Zighed, R. Rakotomalala, C. Agier et F. Muhlenbach, vise à mettre en place des procédés de simplification de bases de règles en s'appuyant sur les techniques d'analyse de données. Une règle dans une base pouvant être considérée comme un tuple, en regroupant des tuples « similaires », on peut faire émerger des groupes qui correspondraient à des concepts plus généraux. Un premier résultat a été obtenu avec l'algorithme « data squeezer » **[E6].** Ce premier travail reste partiel puisqu'il suppose que les conclusions des règles sont fixées, disons qu'il se place dans un contexte d'apprentissage supervisé.

La seconde voie, explorée par N. Nicoloyannis, P. Jouve, G. Legrand, considère la base de données comme un tableau de données assimilable à une relation binaire entre deux ensembles : celui des individus et celui des attributs. L'objectif est alors de décomposer cette grande relation binaire complexe en un ensemble minimal de relations binaires dites maximales. Une relation binaire est dite maximale si elle n'est contenue dans aucune autre relation. Visuellement, il s'agit de réorganiser le tableau binaire en une cascade de blocs rectangulaires de 1. Chacun des blocs peut ainsi conduire à une association, c'est-à-dire, à un ensemble d'attributs prenant la même valeur 1. La recherche de la décomposition optimale demeure un problème difficile. Nous travaillons sur des heuristiques conduisant à des solutions « quasi-optimales ». Pour cela, nous faisons appel au formalisme de la prétopologie[14] qui semble bien adapté.

e. Echantillonnage et Extraction de Connaissances à partir de Données

(**Participants** : O. Boussaid, J.H. Chauchat, R. Rakotomalala)

En extraction des connaissances, l'exploration exhaustive de la base de données devient quasi-impossible à cause des tailles gigantesques qu'atteignent ces bases. L'accroissement de la puissance de calcul n'est pas forcément la meilleure réponse. Il est clair que les données stockées dans une base renferment une information redondante et que l'exploration exhaustive de toute la base n'est pas indispensable pour accéder au contenu informationnel. La statistique intègre le caractère aléatoire des données et donne des résultats (intervalle de confiance, test d'hypothèse) accompagnés d'une probabilité d'erreur que l'on peut contrôler **[F3].**

Nous avons travaillé à l'application des procédures d'échantillonnage dans la construction des arbres de décision et graphes d'induction. L'objectif de ce travail est de chercher à réduire les temps de calcul quand on dispose d'une très grande base contenant de nombreux attributs continus, pour lesquels on doit rechercher la discrétisation optimale à chaque nœud de graphe.

[14] BELMANDT Z.T. *Manuel de prétopologie et ses applications*. Hermès, Paris, (1993).

Pour réduire l'imprécision apportée par l'échantillonnage, nous avons utilisé la notion de puissance d'un test statistique **[B6].** Les conclusions issues de cette recherche sur échantillonnage et graphes d'induction sont les suivantes :

- il faut tirer des échantillons de même taille dans chacun des sous-ensembles de la base définis par les classes à apprendre (« stratification équilibrée ») ;
- il faut tirer un nouvel échantillon équilibré sur chaque nœud ;
- pour réduire le temps de calcul de l'échantillonnage, il faut construire l'ensemble du graphe à partir d'un « échantillon maître » équilibré, représentatif de la base ;
- cette démarche est particulièrement utile (réduction du temps de calcul et bonne stabilité des résultats) quand les classes à apprendre sont très déséquilibrées.

A l'occasion d'une collaboration avec le service marketing d'une grande banque de dépôts, nous avons introduit un coefficient de qualité de ciblage qui permet de comparer divers algorithmes de classification supervisée quand la classe d'intérêt est rare : diagnostic de maladies rares, prévention de pannes, détection de fraudes, ciblage marketing de petits sous-groupes à partir d'une base de clients, etc. **[F34, E9].**

Nous avons expliqué pourquoi la taille des graphes d'induction et des arbres de décision croît avec la taille de la base d'apprentissage, malgré l'utilisation de tests, ou 'pruning', qui ont pour fonction d'empêcher la complexification inutile (« non significative ») du modèle construit. Cette question, posée en 1997 par Oates et Jensen[15], était restée sans réponse à ce jour. Notre réponse **[D10]** utilise le fait que la puissance des tests tend vers un, quand la taille de l'échantillon tend vers l'infini, même si l'apport de la nouvelle dichotomie est très faible.

Nous travaillons actuellement à l'adaptation des procédures de validation croisées, par l'intégration du mode d'échantillonnage (souvent implicite) qui a été utilisé pour constituer la base l'apprentissage : données en grappes, et/ou en strates. Nous avons des résultats sur des données simulées[16] et des applications sur des bases réelles sont en cours.

f. Mesures d'association

(**Participants** : Stéphane Lallich, Ricco Rakotomalala)

Par delà la discrétisation des données numériques, le développement des bases de données dans des domaines très variés a donné un rôle encore accru aux variables catégorielles et aux mesures

[15] OATES T., JENSEN D. *The effects of training set size on decision tree complexity,* Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics, pp. 254-262, Morgan Kaufmann (1997).

[16] CHAUCHAT J.-H., RAKOTOMALALA R. *Validation in machine learning with clustered data set.* Rapport technique du laboratoire ERIC, (2001).

d'association entre variables catégorielles qui fondent la fouille de ces données. Nous avons entrepris un travail à la fois théorique et appliqué sur les mesures d'association entre variables catégorielles.

Nous sommes partis de la notion de diversité développée en statistique par Patil et Tallie[17], analogue à celle d'entropie utilisée en apprentissage[18,19]. A partir de ce concept de diversité, nous avons élaboré **[G5]** une présentation synthétique des mesures d'association brute entre variables catégorielles fondée sur le principe de la réduction proportionnelle de diversité, à l'image des coefficients « proportional reduction in error » (P.R.E) de Goodman et Kruskal[20]. Cette formulation synthétique a été implémentée sur SIPINA[21] ; appliquée en induction par arbres, elle nous a permis d'étudier le comportement des mesures d'association usuelles dans différentes situations de données bruitées **[F7, F14]** et de proposer une gamme de nouvelles mesures d'association fondées sur les entropies de rangs [G4] qui a donné des résultats satisfaisants. Ce travail a été étendu à d'autres expressions de l'association, qui normalisent plutôt le gain d'entropie généralisée par l'entropie de la variable explicative (gain ratio de Quinlan[22]) ou par une moyenne des deux entropies explicative et à expliquer (coefficient de Kvalseth[23]), généralisant ainsi la présentation de Wehenkel[17].

En présence de variables multiples, et au-delà de l'association brute entre deux variables catégorielles X et Y, il devient indispensable d'apprécier comment une troisième variable Z intervient dans l'association. Il s'agit de mesurer l'association entre X et Y en éliminant l'effet de Z (association partielle). A cet effet, on peut évaluer l'association brute de X et Y à Z fixé et synthétiser les résultats obtenus suivant les différentes valeurs de Z par une moyenne pondérée. C'est ainsi que nous avons étendu notre présentation des mesures d'association brutes à celle des mesures d'association partielles entre variables catégorielles **[G5, G15],** suivant la pondération proposée par Olszak et Ritschard[24], pour aboutir à une expression générale de l'association partielle.

---

[17] PATIL G.P., TAILLIE C. *Diversity as a concept and its measurement*, Journal of the American Statistical Association, 77(379):548-567 (1982).

[18] DAROCZY A. *Generalized information function*, Information and Control, 16:36-51 (1970).

[19] WEHENKEL L. *On uncertainty measures used for decision tree induction*, In Proceedings of Info. Proc. and Management of Uncertainty, pp. 413-418 (1996).

[20] GOODMAN L.A., KRUSKAL W.H. *Measures of association for cross-classifications - I*, Journal of the American Statistical Association 49:732-764 (1954).

[21] SIPINA est une plate-forme logiciel développée au laboratoire, et disponible sur Internet à l'adresse : http://eric.univ-lyon2.fr/

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

[23] KVALSETH, T.O. *Entropy and correlation : some comments*, IEEE Transactions SMC, 17(3):517-519 (1987).

[24] OLSZAK M., RITSCHARD G. *The behaviour of ordinal and partial associations measures*, The Statistician 44:195-211 (1995).

g. Méthodes rapides de sélection de variables en apprentissage supervisé.

(**Participants** : Stéphane Lallich, Ricco Rakotomalala)

Parallèlement à nos travaux ayant trait aux propriétés des mesures d'association, nous avons élaboré une méthode rapide de sélection de variables catégorielles multivaluées en situation d'apprentissage supervisé.

Dans un premier temps, nous avons montré l'intérêt du coefficient de corrélation de Pearson pour la sélection rapide d'attributs booléens en apprentissage supervisé. En effet, le coefficient de corrélation partielle qui en découle vérifie deux propriétés importantes :

- il s'exprime à partir des seuls coefficients de corrélation 2 à 2 suivant une formule de passage.

- il est nul en cas d'indépendance de Y et X à Z fixé.

Pour tester cette stratégie de sélection dont la complexité est en $O(p^2n)$, (n exemples, p attributs booléens) nous l'avons associée au modèle bayésien naïf pour l'appliquer à 10 bases de données qui font référence, obtenant d'excellents résultats **[G13].** La sélection de variables induit une forte réduction des attributs, sans détériorer les performances en généralisation, améliorant même celles-ci dans le cas où la taille de la base est faible par rapport au nombre d'attributs.

Pour résoudre le cas des attributs multivalués **[F33],** nous proposons de travailler dans l'espace des paires d'exemples décrits par les indicatrices booléennes de co-étiquetage associées aux différentes variables. Pour une paire d'exemples (i, i'), l'indicatrice de co-étiquetage associée à X vaut 1 lorsque i et i' ont la même étiquette sur X, 0 sinon. On construit les paires avec ordre et répétition pour que l'indépendance de deux variables implique l'indépendance des indicatrices de co-étiquetage associées. Pour une matrice de n exemples décrits par p variables multivaluées, la matrice des indicatrices de co-étiquetage est alors une matrice booléenne de dimensions n^2 et p à laquelle on peut appliquer la stratégie de sélection définie dans le cas booléen. Qui plus est, il n'est pas nécessaire de construire cette table, car on peut exprimer directement le coefficient de corrélation de points entre deux indicatrices de co-étiquetage à partir de leur table de contingence suivant des formules analogues à celles développées par Marcotorchino[25]. C'est ainsi que la complexité de notre méthode est toujours en $O(p^2n)$.

Les résultats obtenus sur un ensemble hétérogène de bases issues du serveur UCI Irvine sont satisfaisants : sur les bases volontairement bruitées, tous les attributs aléatoires ont été éliminés ; chaque fois que l'on connaît les bons attributs, ceux-ci sont parmi les premiers

[25] MARCOTORCHINO, J.-F. *Utilisation des comparaisons par paires en statistique des contingences, Parties I, II, III*, Etudes F-069, F-071, F-081, Centre Scientifique IBM Paris, (1984, 1984, 1985).

introduits ; enfin, la sélection induit systématiquement une réduction drastique du nombre d'attributs, alors même qu'il s'agit de bases souvent déjà épurées. Le taux d'erreur en généralisation est au moins conservé sinon amélioré.

Cette méthode a été implémentée sur SIPINA et nous avons commencé à explorer les perspectives ouvertes par cette méthode de sélection rapide : construction d'arbres contraints, «look ahead», «post pruning», variables synthétiques.

h. Validation de connaissances extraites

(**Participants** : Stéphane Lallich, Fabrice Muhlenbach, Olivier Teytaud, Djamel Zighed)

Un problème aigu du *Data mining* concerne la validation des connaissances qui sont extraites des bases de données. En effet, la puissance des outils et des procédures d'optimisation oblige à s'interroger sur la portée réelle des résultats que l'on peut mettre en évidence. Le problème est d'autant plus ardu que le modèle du *Data mining* est différent de celui de la statistique. Alors qu'en Statistique, on teste une hypothèse spécifiée au préalable indissociable d'une théorie du phénomène étudié, en *Data mining*, on explore les données sans a priori, ce qui amènerait à multiplier les tests tout en travaillant en situation optimisée. Cette spécificité donne tout leur intérêt aux outils issus de la théorie du « statistical learning ». Nos recherches se sont ainsi développées dans trois directions.

- Nous avons entrepris d'utiliser les apports de la statistique spatiale pour construire un critère de validité en apprentissage supervisé en utilisant des modèles géométriques de voisinages.
- Dans un article récent, Nock et Sebban[26] avaient proposé d'utiliser les propriétés du boosting pour élaborer PSBOOST, un algorithme qui réduit la taille de l'ensemble d'apprentissage lorsque l'on utilise des méthodes liées aux plus proches voisins en apprentissage supervisé à 2. L'idée à la base de cet algorithme est que chaque individu est un classifieur de ses plus proches voisins. On peut donc appliquer la démarche du boosting qui consiste à sélectionner à chaque étape l'individu qui classe le mieux ses plus proches voisins, en modifiant au fur et à mesure les poids des individus de manière à renforcer le poids de ceux qui sont mal classés à l'étape précédente. Un critère d'arrêt pour PSBOOST est capital, au contraire du boosting usuel, puisque le but de l'algorithme est de réduire la taille de l'ensemble d'apprentissage et non pas de produire un classifieur pondéré. Nous avons proposé un critère d'arrêt qui valide un individu

[26] NOCK R., SEBBAN M. *Advances in Adaptive Prototype Weighting and Selection*, International Journal on Artificial Intelligence Tools (IJAIT), 10(1-2):137-155 (2001).

candidat uniquement si dans son voisinage réciproque la répartition entre les deux modalités de la classe est significativement différente du résultat d'un pile ou face équilibré (H_0). Nous avons alors identifié la loi sous H_0 de la quantité qui est optimisée à chaque étape par PSBOOST et proposé un critère d'arrêt, par simulation ou approximation normale. Expérimenté sur différentes bases du serveur Urvine, ce critère conduit à une réduction moyenne de 45 % de la taille des ensembles d'apprentissage tout en maintenant les performances en généralisation **[F44].**

- Nous nous sommes intéressés à la validité des règles d'association extraites des bases de données. De nombreuses mesures ont été proposées pour évaluer la validité de ces règles, par delà les seuls critères de *support* et de *confiance*. Cette évaluation est rendue difficile par le très grand nombre de règles à examiner : si l'on traite les différentes règles séparément, les risques s'accumulent et on a toutes les chances de valider des règles absurdes. Nous avons donc entrepris une étude détaillée des intervalles de confiance associés aux évaluations de ces mesures, en nous appuyant sur les outils de la théorie de l'apprentissage statistique, plus particulièrement sur la dimension de Vapnick [G16]. Nous traitons aussi bien les règles vérifiées dans tous les cas que les règles vérifiées partiellement et nous proposons des bornes uniformes pour toutes les règles considérées, bornes qui sont non asymptotiques. Kivinen et Mannila[27] avaient engagé un travail du même ordre limité au seul cas particulier de la détection des formules «suffisamment» fausses et appelé à un travail similaire incluant les apports de la théorie de la VC dimension.

2.1.2.1.2 Structuration et Accès aux données

Les recherches menées sur ce thème s'inscrivent dans le cadre général de l'extraction de connaissances et traitent plus précisément de deux sujets essentiels : d'une part l'organisation et l'analyse des données multidimensionnelles et d'autre part l'organisation et la performance des grandes bases de données. De plus, un intérêt particulier est porté à la possibilité d'utiliser les outils bases de données dans le processus d'extraction de connaissances, tant dans l'organisation et la préparation des données, que pour la construction de règles.

a. Data Warehousing, OLAP

(**participants** : F. Bentayeb, O. Boussaid)

[27] KIVINEN J., MANNILA H. *The power of sampling in knowledge discovery*, In proc. of PODS'94, pp.77-85 (1994).

Les entrepôts de données et les bases de données multidimensionnelles en général permettent un support efficace des processus OLAP (*On-Line Analytical Processing*) qui améliorent la prise de décision dans les entreprises. Plusieurs approches existent parmi lesquelles nous citons l'approche relationnelle ROLAP (Red Brick) et l'approche multidimensionnelle MOLAP tel (Express Server). Plusieurs travaux de recherche sont apparus concernant la modélisation multidimensionnelle[28,29,30]. Ces différentes propositions sont des modèles logiques qui nécessitent des choix de modélisation inhérents à l'environnement choisi (ROLAP, MOLAP, OOLAP). Dans le domaine de la modélisation, nous avons proposé un modèle multidimensionnel de données basé sur le concept de voisinage noté VOLAP **[E8]**. Ce modèle est indépendant de tout système existant, le formalisme utilisé est simple et les opérateurs OLAP y sont redéfinis[31]. Par ailleurs nous avons proposé une architecture physique pour le modèle VOLAP. Le choix de structure repose sur l'utilisation de listes chaînées à liens multiples notées MLL (Multi-Links Lists). L'architecture proposée se compose de différentes listes de données comprenant les coordonnées et les mesures des points (représentant des faits), les dimensions et leurs attributs, et des listes de méta-données comprenant les hiérarchies et leurs paramètres. L'organisation à l'aide de liens multiples fonctionne comme un mécanisme d'indexation et facilite ainsi la navigation dans les données pour la construction des cubes de données. Elle favorise d'autre part la manipulation des données afin d'appliquer les opérateurs OLAP. Une étude de performance a été réalisée en comparant la structure MLL[32] avec un tableau multidimensionnel sans recourir à des algorithmes de compression et d'optimisation. Les résultats de la structure MLL sont meilleurs en termes soit d'espace mémoire soit de temps de réponse lors de la construction des cubes de données. Il en est de même pour les temps de réponse des requêtes de la structure MLL qui sont plus performants. Une plate-forme de programmes pour l'analyse multidimensionnelle basée sur cette structure MLL est en cours de développement.

b. Organisation et performances des grandes bases de données

(**Participant :** J. Darmont)

---

[28] GRAY J., BOSWORTH A., LAYMAN A., PIRADESH H. *Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab and Sub-Totals.* Data Mining & knowledge Discovery 1(1):29-54 (1997).

[29] AGRAWAL R., GUPTA A., SARAWAGI A. *Modeling Multidimensional Databases.* Proc. 13th Int. Conf. Data Engineering, ICDE, pp. 232-243, IEEE Press, 7-11 April 1997.

[30] CHANDURI S., DAYAL U. *An Overview of Data Warehousing and OLAP Technology.* ACM SIGMOD Records 26(1):65-74 (1997).

[31] BENTAYEB F., BOUSSAID O., NICOLOYANNIS N. *Neighborhood Based Multidimensional Data.* ERIC Technical Report N° 08, December 2000.

[32] BENTAYEB F, BOUSSAID O., ORLANDO L. *Multi-Links Lists as Data Structure in MOLAP Environment*. Article soumis à DEXA 2001

Nous avons finalisé lors de l'année 1999-2000 des travaux initiés en dehors du laboratoire ERIC. Ces travaux, menés en collaboration avec les laboratoires LIMOS (Université de Clermont-Ferrand II) et OUDB (Université d'Oklahoma, USA) concernent les performances des systèmes de gestion de bases de données (SGBD) et des SGBD orientés objets en particulier. Ils ont donné lieu à deux types de contributions : (1) des méthodes d'optimisation des performances comme l'algorithme de regroupement d'objets StatClust **[F23]** ; et (2) des méthodes d'évaluation des performances comme le banc d'essais OCB[33] ou le modèle de simulation VOODB[34]

2.1.2.2 Recherches appliquées

2.1.2.2.1 Extraction de connaissances à partir de données textuelles

a. Représentation et sélection d'attributs dans les corpus de textes.

(**Participants** : S. Di Palma, D.A. Zighed)

Nous nous sommes intéressés aux techniques de représentation des textes pour les problèmes d'apprentissage supervisé, privilégiant celles qui imposent de construire les descripteurs à partir du seul contenu typographique des documents à classer. Nous avons plus particulièrement étudié la technique dite des « n-grammes (en français) » consistant à utiliser comme descripteurs potentiels l'ensemble des séquences de n signes à l'intérieur des textes constituant l'échantillon d'apprentissage.

La dimensionnalité (nombre de variables) très grande de ces problèmes nous a naturellement conduit à étudier les différentes techniques de sélection d'attributs concrètement utilisées en catégorisation de textes. Nous avons proposé une ré-interprétation de plusieurs critères de sélection, associée à une méthodologie plus générale de définition d'un critère de sélection d'attributs prenant en compte les spécificités du problème d'apprentissage et de l'algorithme de catégorisation de textes que l'utilisateur souhaite utiliser.

---

[33] DARMONT J., SCHNEIDER M. *Benchmarking OODBs with a Generic Tool.* Journal of Database Management 11(3):16-27 (2000).

[34] DARMONT J., SCHNEIDER M. *VOODB: A Generic Discrete-Event Random Simulation Model to Evaluate the Performances of OODBs.* 25th International Conference on Very Large Databases (VLDB '99), Edinburgh, Scotland, UK, September 1999, pp. 254-265.

Nous travaillons actuellement à la caractérisation des conditions d'équivalence (du point de vue de l'ordre induit sur les attributs) des différents critères de sélection utilisés en catégorisation des textes. Cette caractérisation permettrait de fournir une explication théorique au constat empirique d'équivalence des différents critères que de nombreuses études ont diagnostiqué suite à des tests sur des bases de textes exemples.

Parallèlement à ces techniques de sélection, nous avons également étudié l'application au domaine du texte des techniques de construction d'attributs fondées sur la recherche de combinaisons (le plus souvent linéaires) des attributs initiaux. Cette étude a abouti à la définition d'un cadre cohérent permettant de situer les différentes pratiques dans ce domaine (essentiellement occupé par les techniques dites de « Latent Semantic Indexing »). Elle a également permis la mise au point d'une technique originale de construction d'attributs pour l'apprentissage supervisé à partir de textes dans une optique d'indexation. Cette technique repose sur l'utilisation de méthodes traditionnelles d'analyse conjointe de plusieurs tableaux (analyse canonique ou analyse de co-inertie) appliquées simultanément à une matrice de fréquences de termes et à une matrice booléenne d'indexation.

b. Catégorisation de textes multilingues

(**Participants** : J.H. Chauchat, R. Jalam)

Ces travaux portent sur la conception et la réalisation d'un agent informatique intelligent qui apprend à sélectionner les documents pertinents parmi les pages WEB multilingues selon les thèmes qui intéressent l'utilisateur. Le jugement porté par l'utilisateur sur la pertinence des documents est utilisé par l'agent en auto-apprentissage pour affiner le profil de ce qui est recherché. Le flot de documents est constitué de pages WEB multilingues. Une interface permet à un utilisateur de définir un profil pour chaque langue sur la base de différents paramètres (requêtes, mots clés, documents pertinents, date...).

c. Extraction et gestion de connaissances à partir de données textuelles distribuées

(**Participants** : J. Clech, D.A. Zighed)

Dans ce travail de recherche qui vient de débuter, nous travaillons sur la mise en œuvre des méthodes d'apprentissage automatique pour réaliser des moteurs de recherche «intelligents» capables de classifier dynamiquement des documents issus du web ou d'un intranet. Pour ce faire, ces moteurs devront comporter des outils de représentation de données textuelles comme ceux basés sur les n-grammes que nous avons développés. Ces moteurs devront être capables d'apprendre de façon incrémentale. Ces travaux devront conduire à la réalisation d'un prototype

qui sera testé dans des cas concrets comme la recherche d'informations dans de grandes bases de données médicales. Enfin, ce travail devra s'orienter dans le cadre de la problématique de partage et d'appropriation du cycle de Knowledge Management.

2.1.2.2.2 Extraction de connaissances à partir de données images

(**Participants** : D. Coeurjolly, M. Hammami, S. Miguet, L. Tougne, T. Tweed, D.A. Zighed)
L'extraction de connaissances à partir d'images (Image Mining) est en forte croissance. Cette demande vient essentiellement du domaine de l'indexation automatique. Notre objectif a été de réfléchir sur la manière de développer des outils spécifiques pour l'image mining au même titre qu'ont été développées des techniques pour le text mining. Le contrat que nous avons avec la société Diagnos (cf Projet Mammo) nous a donné le cadre de validation pratique pour ce projet.

Notre premier travail s'est porté sur la transformation d'une image en un vecteur attributs-valeurs car la majorité des techniques de *data mining* travaillent sur ce type de données. Nous avons mis au point une série de primitives pour extraire les paramètres de texture, de forme, de représentation des distributions de niveaux de gris. Cette démarche qui a débuté en 2001 et qui va s'intensifier entre dans le cadre d'un sujet plus large, « la feature construction », que nous considérons comme stratégique pour le développement du multimédia mining. Outre le projet Mammo décrit plus loin, nous sommes en train de tester les techniques d'image mining dans le domaine du filtrage, par exemple pour déterminer si un site web contient ou non des photos pornographiques. Dans son mémoire de DEA, Mohamed Hammami, a pu extraire un modèle prédictif permettant d'identifier si un pixel d'une image appartient ou non à une zone de peau. Les résultats ont été particulièrement encourageants puisqu'ils sont supérieurs à ceux d'une équipe de recherche COMPAQ, qui travaille sur le même sujet depuis plus longtemps.

2.1.2.3 Projets contractuels

2.1.2.3.1 Projet Mammo

(**Participants** : Les trois équipes : image, *data mining* et bases de données décisionnelles, sont impliquées)
La société DIAGNOS Inc, corporation canadienne spécialisée dans le développement de technologies d'extractions de connaissances, a signé un contrat de recherche avec le laboratoire

ERIC pour le développement d'un outil d'extraction de connaissances dans le domaine du dépistage du cancer du sein.

L'objectif est la mise au point d'un logiciel capable d'apprendre à identifier des mammographies cancéreuses. Cette nouvelle technologie pourrait avoir plusieurs applications parmi lesquelles : aide au diagnostic, aide au dépistage de masse, évaluation de pratiques médicales par rapport à un référentiel ou pour intégrer un réseau.

- ✓ **Les objectifs scientifiques du projet Mammo.**

Mammo est un système d'aide au diagnostic pour la détection des cancers de sein. L'objectif de ce projet est d'identifier les paramètres radiologiques, cliniques et ou biologiques et leurs interactions en vue de mettre au point un automate capable de reconnaître la nature cancéreuse ou non d'une mammographie.

Le projet Mammo comporte deux phases :

- La première vise à mettre au point un moteur d'inférence pour l'aide au diagnostic automatique des mammographies. La mise au point de ce moteur passe par une étape d'apprentissage qui consiste à travailler sur un corpus de dossier (radios, examens cliniques,…) de patients pour lesquelles le diagnostic est déjà parfaitement connu.
- La seconde phase consiste à mettre au point un environnement informatique qui permette l'exploitation de ce modèle et son amélioration en situation réelle. Il s'agit d'un logiciel, couplé ou non à un système d'acquisition de radios, qui permette l'accès à la base de données des patients.

La démarche que nous proposons se distingue de celles utilisées jusque là pour l'aide au diagnostic des cancers du sein. En effet, elle ne cherche pas des modèles de visualisation permettant de mieux localiser les tumeurs sur une radio mais d'exploiter toute l'information disponible sur le patient pour émettre un diagnostic avec la plus grande fiabilité possible. Cette information peut être contenue dans les radiographies, les observations cliniques, les paramètres biologiques,… Généralement, c'est une combinaison de ces sources qui est utilisée par les praticiens. Grâce aux techniques de *data mining*, nous pouvons mettre toutes ces données, de sources hétérogènes, sur le même plan et les prendre en considération simultanément comme le font les médecins qui pratiquent naturellement cette fusion de données pour proposer un diagnostic.

2.1.2.3.2 Projet «coeur-cerveau»

La région Rhône-Alpes a confié à cinq laboratoires de recherche un projet sur trois ans concernant la constitution et l'exploitation d'une base de connaissances «coeur-cerveau». Les objectifs généraux de ce projet sont :

- l'amélioration de la prise en charge de patients cérébro-lésés et de patients ischémiques.
- la constitution d'une banque de données anatomo-fonctionnelle aidant à la compréhension des mécanismes cérébraux et cardiaques chez l'homme.
- la création du premier centre distribué national de recueil de données médicales anatomiques et fonctionnelles de patients cérébro-lésés d'une part et de patients atteints d'affections du myocarde d'autre part.

Les laboratoires partenaires concernés par ce projet sont :

- UMR 5823 LASS, Université Claude Bernard Lyon 1
- UMR 5515 CREATIS, Université Claude Bernard Lyon 1
- UMR 5015 ISC, Université Claude Bernard Lyon 1
- TIMC, Université Joseph Fourier Grenoble

2.1.3 Description des travaux du pôle Images

2.1.3.1 Présentation scientifique

Nous résumons ici sous forme synthétique nos principaux résultats, d'une part en présentant les axes thématiques généraux de nos recherches, d'autre part les projets appliqués dans le cadre desquels ces résultats ont été obtenus. Nous obtenons ainsi une matrice à deux entrées, dont les colonnes représentent les actions (souvent contractuelles) ayant motivé ces recherches, et dont les lignes constituent nos principaux champs d'investigation. Une « pastille » à l'intersection d'une ligne et d'une colonne indique qu'un axe scientifique est mis au service d'un projet contractuel.

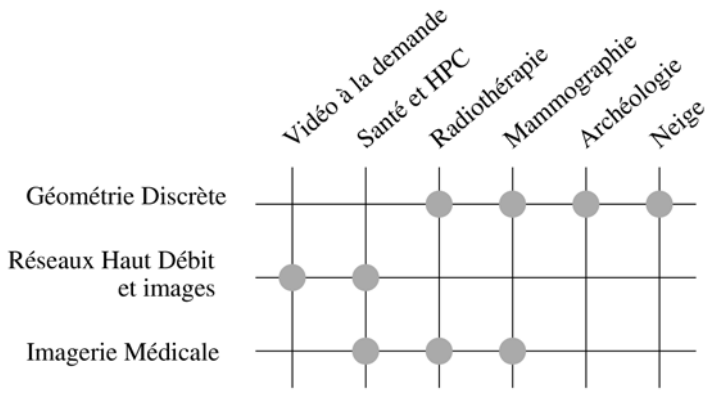


Figure 1 : matrices projets/thèmes

La suite de cette partie est organisée de la manière suivante : nous décrivons en section 2.1.3.2 les trois champs thématiques qui sont développés au sein du pôle images, ainsi que les principaux résultats que nous avons obtenus dans chacun de ces domaines. Nous détaillons ensuite (paragraphe 2.1.3.3) les principales actions que nous avons développées, en décrivant les partenaires impliqués et le cas échéant, les organismes financeurs et la nature des contrats obtenus.

2.1.3.2 Axes de recherche et résultats

2.1.3.2.1 Géométrie Discrète

Tous les systèmes d'acquisition de données images fournissent des données discrètes qu'il faut manipuler et traiter. Une première approche possible consiste à transposer ces données dans le domaine du continu pour ensuite appliquer des outils mathématiquement bien définis dans ce cadre (transformations géométriques, visualisation, calcul de courbure, etc.). L'un des problèmes posés par une telle approche est que le passage des données discrètes aux données continues s'effectue généralement à l'aide d'interpolations qui induisent des erreurs dans les données. La géométrie discrète est une voie alternative qui consiste à préserver aux données leur caractère discret (arithmétique entière) en définissant des opérateurs agissant directement sur ces données. Une partie du pôle Image inscrit ses recherches dans ce cadre.

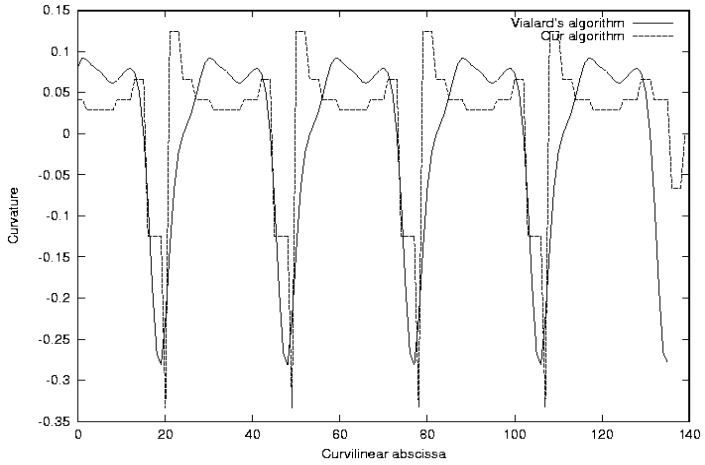


Figure 2 : arcs de cercles discrets et estimation de la courbure en chaque point

Au cours du projet «Archéologie» décrit ci-après, nous avons étudié les notions de droite et de courbure discrètes[35,36]. En particulier, nous avons mis en œuvre un algorithme permettant de calculer en temps linéaire la tangente en chaque point d'une courbe discrète **[F13].** Un tel algorithme permet, en appliquant un filtre gaussien, d'obtenir la courbure en tout point de la courbe. Cependant, l'application d'un tel filtre nous fait sortir du cadre de la géométrie discrète et nécessite l'utilisation d'un seuil qui n'est pas toujours facile à fixer car fortement dépendant des données.

Aussi, dans le cadre du stage de DEA de David Cœurjolly, nous avons mis en œuvre un algorithme de calcul de courbure purement discret d'une courbe 2D ou 3D. L'idée consiste à calculer en chaque point de la courbe le cercle osculateur discret **[F41].**

[35] DEBLED-RENESSON I., RÉVEILLÈS J.P. *A linear algorithm for segmentation of digital curves*, Parallel Image Analysis: Theory and Applications, 73-100 (1993).

[36] A. VIALARD : *Geometrical Parameters Extraction from Discrete Paths*, In Miguet S., Montanvert A., Ubéda S. "Discrete Geometry for Computer Imagery", pages 24-35, Lecture notes in Computer Science, Springer Verlag, Berlin (1996).

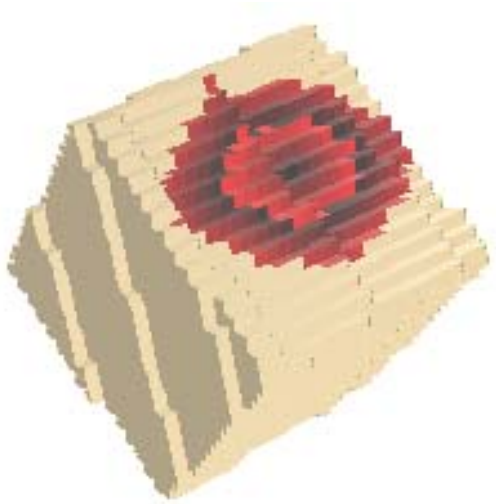


Figure 3 : propagation géodésique à la surface d'un cube discret

Des recherches se poursuivent dans ce cadre et visent notamment le calcul de courbure d'objets discrets 3D. Pour ce faire, un algorithme de calcul de géodésiques discrètes a été mis en œuvre. Ces recherches seront appliquées aux échantillons de neige de Météo-France et serviront dans le cadre du futur projet avec les archéologues de la Maison de l'Orient Méditerranéen.

Une autre voie débutée par Laure Tougne au cours de sa thèse au Laboratoire de L'Informatique du Parallélisme de l'ENS Lyon concerne l'étude des cercles discrets. L'exploitation de ces résultats a donné lieu à l'article **[D8].** Deux extensions de ce travail ont ensuite été mises en œuvre. La première concerne l'étude des phénomènes de moirés que l'on peut voir lorsque l'on dessine sur un écran d'ordinateur des cercles concentriques **[F8].** Le but de la seconde était d'étendre les travaux en dimension 3 et par conséquent d'étudier la construction de sphères sur automates cellulaires de dimension 3 **[F24, D11]**.

2.1.3.2.2 Réseaux Haut Débit et Calculs Hautes Performances

a. Une nouvelle stratégie pour la vidéo à la demande utilisant HTTP.

Alors que la machine développée par CSTI (voir plus bas, projet Vidéo à la demande) utilisait des protocoles réseau propriétaires, nous avons étudié une utilisation particulière du protocole HTTP, pour diffuser ce type de média. Une stratégie originale pilotée par le client a été mise au point. Elle permet la consultation simultanée de près de 60 flux MPEG-1 (qualité télévision) depuis un même serveur **[F19].** De nouvelles mesures, absentes des benchmarks standards

utilisés pour les serveurs web, ont été introduites pour estimer la qualité de service d'une machine spécialisée dans la vidéo à la demande **[E4].**

b. Une plate-forme d'imagerie médicale à accès distant

Dans le cadre du projet « Santé et HPC », nous avons conçu et réalisé une interface Java nommée ARAMIS (*A Remote Access Medical Imaging System*) **[F9, F17]**, permettant à tout utilisateur d'un réseau hospitalier, depuis sa machine habituelle et sans installation de logiciel préalable, d'accéder à partir de son navigateur Web à nos outils parallèles de traitement d'images **[D3, D5, D9, F20]** (voir également[37,38,39]). Nous optimisons ainsi les temps de traitement et minimisons le volume d'information à transférer sur le réseau dans chacune des étapes du calcul (voir figure 4). Dans une deuxième étape, nous avons étudié une première approche de *requête par le contenu*, utilisant à la fois des opérateurs relationnels classiques et des opérateurs d'analyse d'images. La pertinence de ces requêtes a été évaluée par le biais d'une collaboration avec P. Clarysse (laboratoire CREATIS, INSA-UCBL de Lyon). Il s'agit de déterminer automatiquement dans une séquence temporelle d'images cardiaques, l'image de la contraction maximale du cœur (télésystole). Pour cela, nous utilisons une télésystole de référence et la comparons aux autres images à l'aide de mesures de similarité sans segmentation (techniques décrites section suivante).

[37] MIGUET S. *Un environnement parallèle pour l'imagerie 3D*, Habilitation à Diriger des Recherches, Université Lyon 1 (1995).

[38] MIGUET S., NICOD J.M. *An optimal parallel iso-surface extraction algorithm*. In Fourth International Workshop on Parallel Image Analysis (IWPIA'95), pages 65-78. LIP-ENS Lyon (France), December 1995.

[39] LI J.J., MIGUET S. *Parallel volume rendering of medical images*. In Q. Stout, editor, EWPC'92: From Theory to sound Practice, pages 332-343, Barcelone (1992).

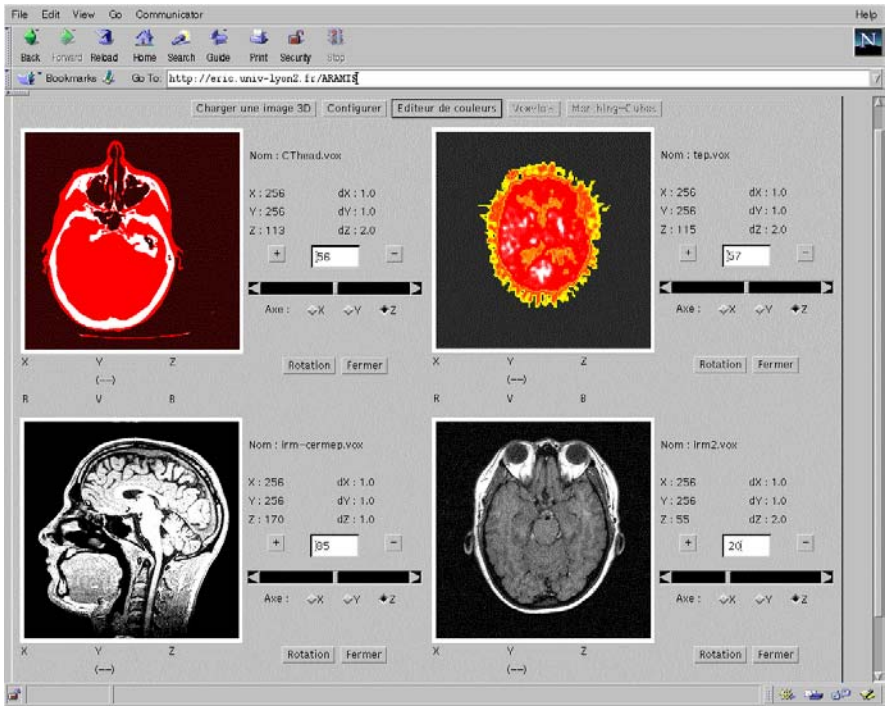


Figure 4 : « ARAMIS » : accès à des bases de données médicales distantes

Cette dernière expérience, couplée au logiciel ARAMIS, a fait l'objet d'une démonstration en direct lors de la présentation du bilan du projet « Santé et HPC » devant les membres du comité d'évaluation de la région, et a également été présentée lors des journées *ASTI* (*Sciences et Technologies de l'Information*).

2.1.3.2.3 Imagerie médicale

Les hôpitaux produisent aujourd'hui des volumes considérables de données images relevant de diverses modalités d'acquisition (CT, IRM, TEP, US ...). Dans ce contexte, nos recherches regroupent divers travaux d'analyse, de traitement et d'extraction de connaissances à partir de ces images, l'objectif étant de fournir aux médecins une aide au diagnostic et à la planification thérapeutique. Nous étudions ainsi le problème de la *comparaison* d'images, par exemple, l'analyse d'une image par rapport à une ou plusieurs images de référence. Cela est étudié à travers deux aspects : d'une part les mesures de similarités « iconiques » et d'autre part l'extraction puis la comparaison de vecteurs d'attributs. L'originalité de ces approches réside dans le parti pris de ne pas chercher à segmenter l'image au préalable, cette étape de segmentation étant extrêmement délicate dans certains cas (images portales).

Les développements effectués ont été valorisés par des publications régulières :

- *Transformations géométriques d'images* : la mise en correspondance passe nécessairement par l'évaluation de positions relatives des images. Ainsi, nous avons développé une approche permettant d'accélérer le calcul des mesures de similarité à travers les transformations rigides d'images **[F19].** Les procédures d'interpolation impliquées ont également fait l'objet d'une étude **[F18].**
- *Décimation de surface* : certains algorithmes de reconstruction surfacique tel les « Marching-cubes » générant un nombre trop important d'éléments de surface (triangles), nous nous sommes intéressés à différentes techniques permettant de simplifier cette description de surface (réduire le nombre de triangles), tout en contrôlant la perte de précision **[G2].**
- *Recalage multimodal et mesures de similarité* : nous avons étudié la mise en correspondance d'images multi-modales, basée sur des mesures de similarité statistiques globales et sans segmentation **[C2, C4, F22].**
- *Radiothérapie* : il s'agit d'une technique de traitement du cancer qui consiste à délivrer à l'aide d'accélérateurs linéaires une dose de rayons X de haute énergie à la zone tumorale tout en épargnant au maximum les tissus sains environnants. Le traitement étant fractionné, la position du patient doit être parfaite lors de chaque séance. Dans ce cadre, nous cherchons à estimer la position du patient en comparant différents types d'images : des images portales, un volume scanner du patient et des images simulées nommées DRR (« Digitally Reconstructed Radiographs »). Des premiers résultats ont été obtenus, donnant lieu à différentes publications internationales **[F31, F40]** et communications nationales **[G12, G17].** D'autres méthodes sont en cours d'évaluation.
- *Mammographie* : toujours sans segmenter les images, notre approche consiste à comparer des radiographies de mammographies à d'autres radiographies présentant des lésions cancéreuses afin d'en déterminer par similarité leur nature cancéreuse ou non. Cela passe par l'extraction d'un grand nombre de caractéristiques de texture et d'histogrammes, qui sont ensuite analysées en collaboration avec le pôle apprentissage. Des premiers résultats ont été publiés **[G18].**

2.1.3.3 Projets et collaborations

2.1.3.3.1 Projet « Vidéo à la demande »

(**Participants** : Serge Miguet, Marian Scuturici)

Ce projet est lié à une collaboration entre les laboratoires LISI de l'INSA de Lyon, ERIC de Lyon 2, et la société CSTI (filiale de la Compagnie des Signaux). CSTI commercialise depuis 1997 une architecture spécialisée dans la vidéo à la demande, et souhaitait développer une recherche technologique liée à ses activités. Deux bourses de thèse ont été financées par la société pour initier des recherches sur la transmission sur des réseaux haut débits et l'archivage de données multimédia. Les deux doctorants ont été co-encadrés par le LISI et par ERIC. Une pompe vidéo Antara développée par CSTI a pu être installée à Lyon 2 dans le cadre de cette collaboration

2.1.3.3.2 Projet « Santé et HPC »

(**Participants** : Serge Miguet, David Sarrut)

Nous avons participé durant trois ans (1997-2000) à un projet multi-thématique nommé Santé et HPC[40]. Ce projet coordonné par J. Demongeot (laboratoire TIMC, Université Joseph Fourier, Grenoble) était une réponse dans le cadre de la thématique prioritaire « optoélectronique et vision » à l'appel d'offres 1997 de la région Rhône-Alpes. Nous faisions partie de la thématique vision quantitative et diagnostic médical coordonnée par I. Magnin et P. Clarysse (laboratoire CREATIS, INSA-UCBL de Lyon), mais le projet dans son ensemble regroupait une dizaine d'équipes de recherche et de partenaires socio-économiques de la région. Notre tâche dans ce projet était de démontrer que les avancées récentes dans le domaine des réseaux haut débit, du calcul parallèle et de l'imagerie médicale pouvaient être rassemblées pour concevoir des systèmes bon marchés utilisables en routine au sein des hôpitaux, ou en enseignement, dans des facultés de médecine.

2.1.3.3.3 Projet « Radiothérapie »

(**Participants** : Sébastien Clippe, Fabien Feschet, Serge Miguet, David Sarrut)

Les travaux de recherche du projet « radiothérapie » concernent l'étude de techniques d'imagerie permettant d'estimer puis de corriger le positionnement de patients en radiothérapie conformationnelle. Ils s'effectuent en collaboration avec le Centre de lutte contre le cancer Léon Bérard de Lyon (CLB), notamment avec plusieurs membres du département de radiothérapie : C. Carrie (chef du département), C. Ginestet (chef du département de radiophysique), C. Malet et F. Lafay (physiciens). Cette équipe possède une très grande expérience en radiothérapie

[40] HPC peut être décliné sous deux formes : «Haute Puissance d'exécution de Calcul» ou «Haut Pouvoir d'analyse et de gestion de la Complexité»

conformationnelle[41,42] sur le plan national et international. Elle dispose de deux accélérateurs linéaires munis de collimateurs multi-lames (le premier acquis en 1994 et le second en 1997). Le Centre Léon Bérard dispose également de systèmes modernes d'acquisition des images portales (EPID, voir Figure) depuis 1994 et d'un logiciel de génération de DRR. Ces travaux sont soutenus par deux projets financés. Le premier est un BQR[43] inter-établissement entre les universités de Lyon 2 et Lyon 1 (le coordinateur est B. Shariat du LIGIM). Le deuxième est un projet de la région Rhône-Alpes nommé « Adémo » : Acquisition & Décision conduites par le Modèle. Dans ce cadre, nous faisons partis du thème modélisation et suivi spatio-temporel pour le diagnostic et la thérapie, dont la coordinatrice est J. Troccaz (GMCAO, UJF, Grenoble).

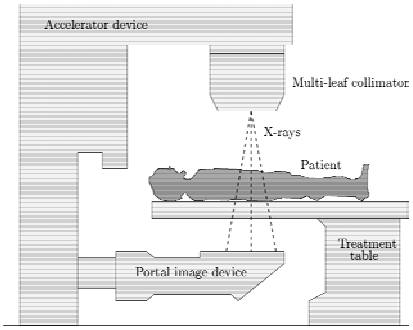


Figure 5 : imagerie portale (Electronic Portal Imaging Device, EPID)

2.1.3.3.4 Projet « Mammographie »

(**Participants** : Sébastien Clippe, David Cœurjolly, Serge Miguet, Tiffany Tweed)

Le projet MAMMO, présenté dans sa globalité plus haut, est repris dans cette section pour la dimension imagerie médicale importante qu'il comporte.

Le pôle d'imagerie médicale travaille sur un corpus composé de plusieurs milliers d'images, correspondant à des examens sains et cancéreux. En cas de cancer, le type et la localisation de la lésion sont précisément décrits par un spécialiste. Par des techniques d'analyse d'images, nous extrayons un très grand nombre de vecteurs attributs, à la fois sur des régions cancéreuses et normales. Nous effectuons ensuite une sélection des zones d'intérêt à partir de ces

[41] MALET C., GINESTET C., HALL K., LAFAY F., SUNYACH M.P., CARRIE C. *A study of dose delivery in small segments.* Int. J. Radiat. Oncol. Biol. Phys., 48(2):535-9 (2000).

[42] GINESTET C., MALET C., COHEN A., LAFAY F., CARRIE C. *Impact of tissues heterogeneities on minitor units calculation and ICRU dose point: analysis of 30 cases of prostate cancer treated with 3D planning system.* Int J. Radiat. Oncol. Biol. Phys., 48(2):529-34 (2000).

[43] Bonus Qualité Recherche

attributs. Ces paramètres sont juxtaposés aux paramètres cliniques et biologiques de chaque patient. Le pôle Apprentissage traite par la suite cet ensemble de paramètres et obtient la liste des variables pertinentes pour distinguer les cas cancéreux des cas normaux ou douteux. Notre but final est que le système d'aide au diagnostic ne déclare jamais bénin un cas diagnostiqué comme cancéreux ou douteux par un expert.

Figure 6 : Vues latérales des seins gauche et droit avec détourage de la tumeur.

2.1.3.3.5 Projet «Archéologie»

(**Participants** : David Cœurjolly, Serge Miguet, Laure Tougne)

Ce projet est une collaboration entre E.R.I.C. et la Maison de l'Orient Méditerranéen. Il a pour but d'extraire des caractéristiques et de classifier des profils de stèles funéraires numérisés (voir figure ci-contre). Nous possédons, outre ces profils numérisés, des textes écrits résumant pour chacun de ces profils certaines de leurs «propriétés» (datation, lieu,...). L'objectif du projet est de rechercher un lien entre les «motifs dominants» extraits des images et les propriétés contenues dans les textes. Un tel système doit permettre d'apporter une assistance aux archéologues dans l'étude de la datation et de la provenance d'une stèle dont on ne connaîtrait que les caractéristiques géométriques. Une nouvelle collaboration entre le laboratoire et la Maison de l'Orient Méditerranéen est en train de se mettre en place et devrait donner lieu à un contrat européen. Il s'agirait là de mettre en place un outil permettant d'aider les archéologues dans leur étude des ornements sculptés figurant sur des monuments funéraires. Un tel outil serait pour

notre équipe l'occasion de mettre en œuvre, sur des données réelles, les algorithmes 3D discrets basés sur la courbure, actuellement en cours de développement.

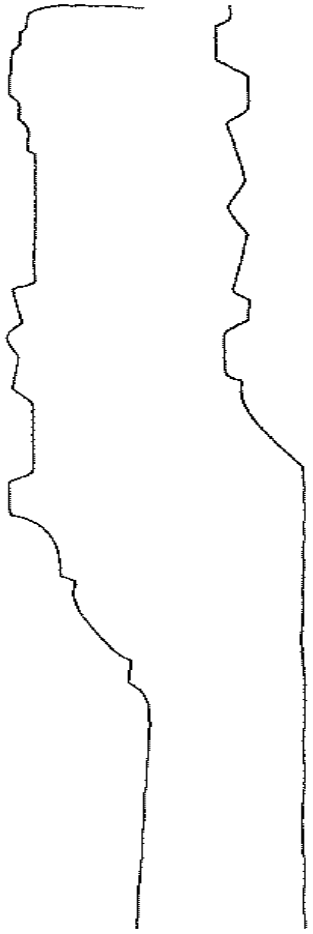


Figure 7 : profils numérisés de stèles funéraires

2.1.3.3.6 Projet «Neige»

(**Participants** : David Cœurjolly, Laure Tougne)

Le projet NEIGE est une collaboration avec le Centre d'Etudes de la Neige de Météo-France de Grenoble. L'un des principaux axes de recherche du CEN consiste en l'analyse des propriétés mécaniques et thermodynamiques d'un échantillon de neige afin de mieux modéliser le comportement d'un manteau neigeux (avalanches). Pour étudier les microstructures d'un échantillon de manière non invasive, des techniques de tomographies X par absorption sont

utilisées permettant ainsi d'obtenir un volume binaire 3D[44]. Ces objets discrets sont ensuite caractérisés par des mesures de courbure ou de champ de normales et trouvent donc une nouvelle application à nos recherches.

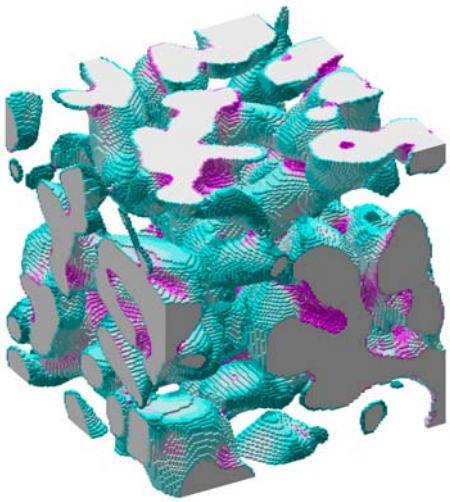


Figure 8 : échantillon de neige et calcul de courbure

2.1.3.4 Perspectives

Le projet *radiothérapie* se poursuit par le prolongement logique du DEA de Sébastien Clippe en une thèse de doctorat. Une nouvelle méthode de repositionnement a déjà été développée et deux articles sont en cours de soumission.

Déformation du poumon : nous allons également débuter une étude sur la déformation d'un poumon lors de la respiration du patient. Ce travail est la poursuite des développements sur le recalage d'images puisqu'il concerne le recalage non rigide (déformable) des organes. Vlad Boldéa, stagiaire effectuera cette année son stage de DEA au laboratoire, nous aide actuellement pour l'étape bibliographique de ce projet. L'application de cette étude est liée à la validation en radiothérapie d'un système de blocage de la respiration, permettant d'immobiliser la tumeur lors d'un traitement.

[44] BRZOSKA J.B., LESAFFRE B., COLÉOU C., XU K., PIERITZ R.A. *Computation of 3D curvature on a wet snow sample*, European Physical Journal Applied Physics, 7:45-57 (1999).

Mammographies : les travaux effectués dans le cadre précédent pourraient également être très utiles dans le projet mammographies : ces méthodes serviraient à mettre en correspondance spatiale les deux mammographies (sein droit et sein gauche) afin d'en étudier les différences, comme le font intuitivement les médecins.

CPER : dans le cadre du *Contrat de Plan Etat Région*, 8 millions de francs ont été inscrits sur les thèmes liés à l'Ingénierie de la santé. De nombreux projets en Rhône-Alpes entrant dans ce cadre bénéficieront de financements complémentaires. Ainsi, les projets *mammo* et *radiothérapie* ont été présentés au comité scientifique en charge de l'évaluation, et sont susceptibles d'obtenir de nouveaux crédits.

2.1.4 Description des travaux du pôle Connexionisme

2.1.4.1 Présentation scientifique

L'émergence du Pôle Connexionnisme dans le laboratoire ERIC avait pour objectif de développer divers aspects de recherche concernant les réseaux de neurones : études théoriques visant à mieux comprendre la complexité des modèles connexionnistes et à mieux cerner leurs capacités d'apprentissage et de généralisation ; études pratiques conduisant au développement d'algorithmes rapides et performants, en particulier pour le traitement de grandes bases de données, ou de données incomplètes ; élaboration de réseaux connexionnistes hétérogènes et modulaires pour la modélisation de processus cognitifs. Les premiers travaux, développés à ERIC, sur ce thème **[F11,G6,G7],** ont été poursuivis à l'Institut des Sciences Cognitives suite au rattachement à 100% de Hélène Paugam-Moisy à l'ISC.

2.1.4.2 Axes de recherche et résultats obtenus

2.1.4.2.1 Réseaux de neurones et SVM

a. Complexité des réseaux multicouches

(**Participants** : André Elisseeff, Hélène Paugam-Moisy)

Nous nous préoccupons de la taille des réseaux multicouches et de la complexité de leur apprentissage. Pour les réseaux d'unités à seuil, Hélène Paugam-Moisy a obtenu des résultats prolongeant une étude menée en collaboration avec Claire Kenyon qui est maintenant professeur à l'Université Paris-Sud/Orsay **[G8].** Les résultats obtenus portent sur le nombre de couches cachées nécessaires à la réalisation exacte des dichotomies polyédriques[45].

Pour les réseaux d'unités à fonctions de transfert continues (sigmoïdes, gaussiennes, etc...), André Elisseeff et Hélène Paugam-Moisy ont achevé et publié **[D7]** un travail sur la détermination, par un algorithme d'apprentissage en temps polynomial, du nombre de cellules cachées et des poids d'un réseau sur une base d'exemples fixée. Ils ont également obtenu des bornes sur les nombres de cellules sur plusieurs couches cachées, les résultats théoriques venant confirmer une étude expérimentale dont les conclusions étaient a priori contraires aux idées reçues **[G3].**

b. Support Vector Machines

(**Participants** : André Elisseeff, Hélène Paugam-Moisy, Olivier Teytaud)

Nous nous intéressons depuis plusieurs années **[B7]** aux Support Vector Machines (SVM), ces nouveaux systèmes d'apprentissage définis par Vapnik[46] et qui s'avèrent très performants pour de multiples applications (en classification et en régression) mais dont les limites et les conditions de mise en oeuvre sont encore mal cernées. Au laboratoire ERIC, nous avons développé l'étude de plusieurs aspects théoriques et pratiques de ces modèles **[F38].** Olivier Teytaud a établi, en collaboration avec Jérôme Droniou, mathématicien en thèse à l'Université de Marseille, plusieurs résultats concernant le comportement asymptotique du nombre de [en italique]*support vectors* dans les SVM et le fait que celles-ci, sous certaines conditions, réalisent des séparations moins

---

[45] KENYON C., PAUGAM-MOISY H. *Advances for exact resolution of polyhedral dichotomies by multilayer neurel networks*, Rapport technique ERIC 99-06, Université Lyon 2, juin 1999.

[46] VAPNIK V.M. *The nature of statistical Learning*, Springer Verlag, Berlin (1995).

VAPNIK V.M. *Statistical Learning Theory*, Wiley, NY (1998).

performantes que les méthodes bayésiennes **[G3].** André Elisseeff et lui ont recherché et mis en oeuvre de nombreux algorithmes d'optimisation efficaces afin d'accélérer l'apprentissage des SVM et de permettre à ce modèle connexionniste d'être appliqué à de grandes bases de données (nombre d'exemples élevé et/ou grande dimension de l'espace d'entrée). Dans ce contexte, l'adaptation de l'algorithme de Franck et Wolfe[47] à l'optimisation d'un algorithme d'apprentissage dans les SVM **[G11]** a valu à André Elisseeff le prix du meilleur article «Jeune-Chercheur» à la Conférence CAp'2000 sur l'Apprentissage.

c. Discrimination multiclasse

(**Participants** : André Elisseeff, Hélène Paugam-Moisy)

Ce thème a été essentiellement développé par André Elisseeff, avec Hélène Paugam-Moisy et en collaboration avec Yann Guermeur, maître de conférences, qui poursuit ses recherches au LORIA, à Nancy. Il s'agit d'aborder directement le problème de la classification multiclasse, sans passer par la résolution de plusieurs problèmes «une classe contre les autres» **[G9, F27, F26].** Une nouvelle notion de marge multiclasse a ainsi pu être introduite et sa mise en oeuvre pour définir et programmer directement des SVM multiclasses représente une partie essentielle de la thèse d'André Elisseeff **[C3].**

2.1.4.2.2 Théorie de l'apprentissage

a. Théorie des bornes

(**Participants** : André Elisseeff, Didier Puzenat, Olivier Teytaud)

Plusieurs autres publications concernent, plus généralement, la théorie statistique de l'apprentissage. En particulier, des calculs de bornes d'erreur réalistes et proches de celles mesurées par cross-validation **[F12],** sont reliés à ce thème d'étude. Les nombres de couverture, nombres d'entropie et l'étude de la marge d'erreur sont les notions pivots de cette recherche **[B8]**[48]. André Elisseeff et Didier Puzenat ont collaboré avec Gérald Gavin, rattaché au pôle Apprentissage-Extraction des Connaissances, pour étudier et borner les performances des combinaisons linéaires de fonctions de prédiction **[F16, F15].**

---

[47] FRANCK M., WOLFE P. *An algorithm for quadratic programming*, Naval Research Logistics Quaterly 3:95-110 (1956).

[48] D'ALCHE-BUC F., PAUGAM-MOISY H. *Analyse probabiliste de l'apprentissage et extensions de la notion de VC-dimension*, Atelier du GDR-13 - Plate-forme AFIA, Palaiseau, France (1999).

b. Réseaux de régularisation

(**Participants** : André Elisseeff)

En collaboration avec Stéphane Canu, professeur à l'INSA de Rouen, André Elisseeff s'est également intéressé à la théorie de la régularisation. Ils ont montré comment cette approche pouvait justifier les bonnes performances obtenues par les réseaux multicouches à fonctions sigmoïdes ainsi que par les méthodes d'apprentissage à noyaux[49].

c. Stabilité des algorithmes d'apprentissage

(**Participants** : André Elisseeff)

Cette voie de recherche, s'incluant, comme la précédente, dans le cadre de la théorie statistique de l'apprentissage, a été explorée par André Elisseeff et figure dans sa thèse **[C3].** Des résultats plus récents ont été obtenus à l'occasion de sa collaboration avec Olivier Bousquet, étudiant en thèse à l'Ecole Polytechnique, sous la direction de Michèle Sebag. Ces travaux ont donné lieu à une communication à NIPS*2000 **[F25].**

d. Paradigmes d'apprentissage

(**Participants** : Hélène Paugam-Moisy, Olivier Teytaud)

Il s'agit de mener une étude comparée de différents paradigmes d'apprentissage : du paradigme bayésien au risque structurel en passant par le maximum de vraisemblance. Des méthodes basées sur les classes de Donsker permettent d'obtenir des résultats positifs en VC-dimension infinie, donc dans un cadre souvent plus réaliste que la VC-dim finie. Ce travail **[F38]** se poursuit à l'ISC.

2.1.4.2.3 Recherches appliquées

a. Text Mining

(**Participants** : Olivier Teytaud)

Ce thème de recherche, qui s'inscrit directement dans les domaines d'étude du laboratoire ERIC, a tout à attendre des méthodes d'apprentissage. C'est donc tout naturellement qu'Olivier Teytaud a appliqué les SVM à un problème qui intéressait Radwan Jalam. Les résultats obtenus au cours de cette collaboration avec le pôle Apprentissage-Extraction des Connaissances a donné lieu à deux publications dans des conférences **[F43, E7].**

---

[49] CANU S., ELISSEEFF A. *Why Sigmoid-shaped functions are good for learning ?*, Learning Workshop, Snowbird, USA, Mars 1999.

b. Prédiction et stabilisation d'oscillateurs

(**Participants** : Olivier Teytaud)

En physique, des méthodes de prédiction sont appliquées dans le cadre des systèmes chaotiques. Les objectifs sont, d'une part, de déterminer des tailles de fenêtrage (correspondant à l'ordre du système dynamique), et d'autre part, de manière plus originale, de tester les relations entre les différentes grandeurs en jeu (ex. amplitude et fréquence dans le système de Chua). Une application éventuelle serait de vérifier si les fluctuations prédictibles sont d'amplitude négligeable devant les fluctuations correspondant à un bruit blanc. S'il s'avère que les fluctuations sont prédictibles, même dans le cadre d'oscillateurs stables, un objectif à long terme serait la stabilisation de tels oscillateurs. Ce travail, mené en collaboration avec des Physiciens de l'ENS Lyon, a donné lieu à une publication dans une conférence **[F30]** et se poursuit à l'ISC.

2.1.4.3 Projets contractuels

2.1.4.3.1 European Working Group NeuroCOLT

Suite à notre participation active au Working Group Européen NeuroCOLT lorsque nous étions à l'ENS Lyon (1994-1997, responsable : Michel Cosnard), nous avons poursuivi nos activités dans ce groupe de travail européen, et nous avons étendu le nœud lyonnais à l'Université Lyon 2, pour le second projet, NeuroCOLT 2 (1998-2001, responsable : Pascal Koiran). A ce titre, nous avons publié un rapport de recherche[50] sur le site web de NeuroCOLT (consultation internationale) et nous avons participé aux rencontres annuelles de Londres (avril 1999) et Graz (mai 2000). Hélène Paugam-Moisy a été invitée à présenter les activités de son groupe de recherche au Workshop « New Perspectives in Neural Nets Theory » organisé par Wolfgang Maass, à Graz en mai 2000[51]. Ces activités ont été financées par le budget européen géré par Pascal Koiran, à l'ENS Lyon.

---

[50] ELISEEFF A., PAUGAM-MOISY H., GUERMEUR Y. *Margin error and generalization capabilities of multi-class discriminant systems*, NeuroCOLT 2 Tech. Report, NC-TR-99-51, Décembre 1999.

[51] PAUGAM-MOISY H. *Multiclass discrimination, SVM and Multimodal Learning.* NeuroCOLT Workshop "New perspectives in the theory of neural nets", Graz, Autriche, Mai 2000.

2.1.4.3.2 ELF-France, laboratoire CRES

Nous avons travaillé sur un contrat avec la société ELF entre octobre et décembre 1999. Ce contrat (CR 11 669) portait sur des comparaisons de méthodes d'apprentissage, en classification et en régression. Il a impliqué André Elisseeff et Olivier Teytaud et nous a permis une collaboration avec Ricco Rakotomalala, du pôle Apprentissage-Extraction des Connaissances. Ce travail[52] a été reconnu comme de grand intérêt pour la société ELF qui a invité Olivier Teytaud à participer au workshop des thésards ELF-TOTAL, en octobre 2000[53]. Nous avons obtenu un nouveau contrat, d'octobre à décembre 2000 (CR 13 165), portant essentiellement sur le traitement du bruit et de l'apprentissage sur des sous-zones de la base d'exemples. Ce deuxième contrat a été exécuté par Olivier Teytaud et Hélène Paugam-Moisy[54].

2.1.4.3.3 Association ECRIN, GT «Prévision des pics de pollution»

André Elisseeff et Olivier Teytaud ont mis en oeuvre et comparé (en performance et en temps) diverses méthodes d'apprentissage (SVM, K-plus proches voisins, etc...) pour participer à l'étude d'un benchmark proposé par l'Association ECRIN (collaboration laboratoires publics / entreprises) portant sur le thème «Prévision des pics de pollution par méthodes d'apprentissage». Des méthodes permettant de traiter les données incomplètes ont également été étudiées et mises en oeuvre. A l'issue de ce travail, un article a été présenté aux Journées Inter-Réseaux, à Bordeaux, en octobre 2000[55]. Cette étude préliminaire et bénévole, mais néanmoins conséquente, pourrait déboucher prochainement sur des contrats avec des entreprises intéressées par cette thématique.

2.1.4.3.4 Projet financé par le Programme Cognitique

Répondant à l'appel d'offre lancé à l'été 1999, nous avons obtenu un projet financé par le Programme national « Cognitique », sur le thème «Etude de codages catégoriels et métriques de

---

[52] ELISSEEFF A., PAUGAM-MOISY H., RAKOTOMALALA R., TEYTAUD O. *Sélection de Support Vector Machines efficaces*, Rapport sur contrat CR 11 669, Université Lyon 2 et ELF ANTAR France (2000).

[53] TEYTAUD O., PAUGAM-MOISY H., CHEBRE M., DI CRESCENZO E., HARTMANN F. *Apprentissage à partir de données et Support Vector Machines*, Journées Scientifiques ELF et TOTAL, Cannes, France (2000).

[54] PAUGAM-MOISY H., TEYTAUD O. *Apprentissage robuste*, Rapport sur contrat CR 13 165, Université Lyon2 et ELF ANTAR France (2001).

[55] ELISSEEFF A., PAUGAM-MOISY H., TEYTAUD O. *Tentative de prédiction de pollution sans réduction manuelle du nombre de variables - Support Vector Machines.* Journée Inter-Réseaux 2000, Bordeaux, France (2000).

l'information visiospatiale et de leurs conséquences sur les performances des êtres humains et des modèles». Ce projet regroupe sept équipes de recherche, réparties sur plusieurs régions de France et au Luxembourg, et son responsable est Frédéric Alexandre, chercheur à l'INRIA Lorraine (équipe Cortex du laboratoire LORIA, Nancy). Il est pour nous l'occasion d'une collaboration de recherche avec Olivier Koenig et son laboratoire «Etude des Mécanismes Cognitifs», à l'Université Lyon 2, collaboration qui vient compléter notre large investissement commun dans les formations de 2ème et 3ème cycles en Sciences Cognitives. Notre recherche consiste tout particulièrement à développer des modèles connexionnistes, et à en étudier les complexités respectives, pour la simulation de tâches catégorielles et de tâches métriques. Cette étude va de pair avec des expériences en Psychologie Cognitive réalisées sur des sujets humains et des études par IRMf événementielle, afin d'apporter des arguments fonctionnels aux hypothèses émises par les Psychologues et aux observations qu'ils réalisent. Nos modèles et nos conclusions seront amenés à être comparés avec les études menées en parallèle par les équipes de Nancy, de Paris et du Luxembourg. Un stagiaire du DEA de Sciences Cognitives, Jean-Philippe Magué, et un stagiaire de la Maîtrise de Sciences Cognitives, Julien Besle, poursuivent ce projet à l'Institut des Sciences Cognitives.

2.1.5 Utilisation des crédits des 4 années précédentes

Dotation ministérielle annuelle : 80 000 F TTC

Dont :

- TVA 19.6% 13 110 F
- BQR 10 034 F

Ventilation fonctionnement	1998	1999	2000	2001
Fournitures administratives	7 525 F	9 480 F	10 520 F	8 397 F
Mission, déplacement, accueil extérieurs à l'établissement	11 220 F	15 526 F	18 479 F	19 255 F
Frais de colloques	900 F	1 300 F	1 200 F	/
Réceptions	3 255 F	6 541 F	5 985 F	6 327 F
Ouvrages, abonnements	1 227 F	1 394 F	1 685 F	2 754 F
Prestations internes (téléphone, affranchissement, reprographie)	13 527 F	15 837 F	16 700 F	17 929 F
Fournitures et matériels de recherche (mise à jour logiciels, consommables informatique)	4 474 F	6 778 F	2 287 F	2 194 F
Travaux d'entretien et de sécurité	14 728 F			

2.2 BILAN QUANTITATIF SUR LES QUATRE DERNIERES ANNEES

2.2.1 Les publications et communications majeures 1998-2001

2.2.1.1 Ouvrages

P1 **[A1]** MIGUET S., MONTANVERT A., WANG P.S.P. World Scientific. *Parallel Image Analysis : tools and models.* Machine Perception 31, World Scientific, Londres **(1998)**.

P2 **[A2]** RITSCHARD G., BERSCHTOLD A., DUC F., ZIGHED D.A. *(Eds) Apprentissage : Des principes naturels aux modèles artificiels.* Hermès, Paris **(1998)**.

P3 **[A3]** ZIGHED D.A., RAKOTOMALALA R. *Graphes d'induction - apprentissage et data mining.* Hermès, Paris **(2000)**.

P4 **[A4]** ZIGHED D.A., KOMOROWSKI J., ZYTKOW J. *(Eds) Principles of Data mining and Knowledge Discovery.* Lecture Notes in Artificial Intelligence, Springer Verlag, Berlin **(2000)**.

2.2.1.2 Chapitres dans un ouvrage

P5 **[B1]** CHAUCHAT J. H., RISSON A. *BERTIN's Graphics and Multidimensional Data Analysis.* In Visualization of Categorical Data, Blasius J., Greenacre M. (Eds) pages 37-45, Academic Press, San Diego, CA, USA **(1998)**.

P6 **[B2]** SEBBAN M., RABASÉDA S., ZIGHED D.A. *Features selection for a gait identification model.* In Apprentissage : des principes naturels aux méthodes artificielles. Ritschard G., Berchtold A., Duc F., Zighed D.A. (Eds) pages 139-150, Hermès, Paris **(1998)**.

P7 **[B3]** RAKOTOMALALA R., ZIGHED D.A. *Mesures P.R.E dans les graphes d'induction : une approche statistique de l'arbitrage généralité-précision*, In Apprentissage : des principes naturels aux méthodes artificielles. Ritschard G., Berchtold A., Duc F., Zighed D.A. (Eds.), pages 257-270, Hermès, Paris **(1998)**.

P8 **[B4]** ZIGHED D.A., RABASEDA S., RAKOTOMALALA R., FESCHET F. *Discretization Methods in Supervised Learning*. In Encyclopedia of computer science and technology, vol 40. Kent A., Williams J.G. (Eds) pages 35-50 Marcel Dekker, New York, NY **(1998).**

P9 **[B5]** RAKOTOMALALA R., ZIGHED D.A., FESCHET F. *Caractérisation des règles de production dans un processus d'induction*, In Apprentissage automatique, Sebban M., Venturini G. (Eds) pages 159-174 Hermès, Paris **(1999)**.

P10 **[B6]** CHAUCHAT J.H., RAKOTOMALALA R., *Sampling Strategy for Building Decision Trees from Very Large Databases Comprising Many Continuous Attributes*, In Instance Selection and Construction - A *Data mining* Perspective. Liu H., Motoda H. (eds.) pages 179-188, Kluwer Academic Publishers, Londres **(2001)**.

P11 **[B7]** GUERMEUR Y. , PAUGAM-MOISY H. *Théorie de l'apprentissage de Vapnik et SVM, Support Vector Machines*. In Apprentissage automatique, Sebban M., Venturini G. (Eds), pages 109-138, Hermès, Paris **(1999)**.

P12 **[B8]** SMOLA A., ELISSEEFF A., SCHOLKOPF B., WILLIAMSON R.C. *Entropy Numbers for Convex Combinations and Multilayer Networks*. In Smola A., Bartlett, Scholkopf B. and Schuurmans (Eds), Advances in Large Margin Classifiers, MIT Press, Denver, USA **(1999).**

P13 **[B9]** RITSCHARD G., ZIGHED D.A., NICOLOYANNIS N. *Maximiser l'association par agrégation dans un tableau croisé*. In La fouille dans les données par la méthode d'analyse statistique implicative. Gras R., Bailleul M. (Eds), pages 219-233, Ecole Polytechnique de l'Université de Nantes, IRIN et IUFM Caen, n° ISBN : 2-9516505-0-7 **(2001)**.

2.2.1.3 Articles dans des revues internationales

P14 **[D1]** ZIGHED D.A., RABASÉDA S., RAKOTOMALALA R. *FUSINTER : a discretization method of continuous attributes*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 6(3):307-326 **(1998)**.

P15 **[D2]** NICOLOYANNIS N., TERRENOIRE M., TOUNISSOUX D. *An Optimization model for aggregating preferences : a simulated annealing approach*. Health and System Science 2(1-2):33-44 **(1998).**

P16 **[D3]** FESCHET F., MIGUET S., PERROTON L. *ParList: a Parallel Data Structure for Dynamic Load Balancing*. JPDC, 51:114-135 **(1998)**.

P17 **[D4]** BERGERON S., LALLICH S., LE BAS C., *Location of innovative activities and technological structure in the french economy, 1985-1990. Some evidences from US patenting*, Research Policy 26(7-8):733-751 **(1998).**

P18 **[D5]** CONTASSOT-VIVIER S., MIGUET S. *A Load-Balanced Algorithm For Parallel Digital Image Warping.* International Journal of Pattern Recognition and Artificial Intelligence. 13(4):445-463 **(1999)**.

P19 **[D6]** SAGOT C., BRUN J.J., GROSSI J.L., CHAUCHAT J.H., PONGE J.F. *Earthworms distribution and humus forms evolution in the sylvigenetic cycle of a subnatural mountain spruce forest*, Eur. J. Soil. Biol. 35(4):163-169 **(1999)**.

P20 **[D7]** ELISSEEFF A., PAUGAM-MOISY H. *JNN, a Randomized Algorithm for Training Multilayer Networks in Polynomial Time*. NeuroComputing, 29:3-24 **(1999).**

P21 **[D8]** DELORME M., MAZOYER J., TOUGNE L. *Discrete parabolas and circles on 2D cellular automata,* Theoretical Computer Science 218, 2:347-417 **(1999)**.

P22 **[D9]** MIGUET S., PIERSON J.M. *Quality and Complexity Bounds of Load Balancing Algorithms for Parallel Image Processing*, International Journal of Pattern Recognition and Artificial Intelligence, 14(4):463-476 **(2000).**

P23 **[D10]** SEBBAN M., NOCK R., CHAUCHAT J.H., RAKOTOMALALA R. *Impact of Learning Set Quality and Size on Detection Tree Performances : a Comparative Study,* International Journal of Computers, Systems and Signals, (IJCSS), 1(1):85-105 **(2001)**.

P24 **[D11]** FESCHET F., TOUGNE L. *Discrete Waves on Cellular Automata*, International Journal on Pattern Recognition and Artificial Intelligence, (à paraître) **(2001)**.

2.2.2 Communications avec actes

C1 **[F1]** CONTASSOT-VIVIER S. *Optimization and Parallelization of a Geographical Stereo Vision Code.* The Sixth International Conference in Central Europe on Computer Graphics and Visualization '98. pages 65-72 **(1998)**.

C2 **[F2]** CÔTÉ M., ZIGHED D.A., TROUDI N. *SPAM Miner: the data-mining and the agents technology to fight the illicit messages.* Fourth International Conference on Computer Science and Informatics. Proceedings of JCIS '98, Li K., Pan Y., Wang P.P.(eds) vol. IV, pages 21-24, Duke University, Durham, NC **(1998) .**

C3 **[F3]** CHAUCHAT J.H., BOUSSAID O., AMOURA L. *Optimization sampling in a large database for induction trees*. Fourth International Conference on Computer Science and Informatics. Proceedings of JCIS '98, Li K., Pan Y., Wang P.P.(eds) vol. IV, Pages 28-31, Duke University, Durham, NC **(1998).**

C4 **[F4]** NICOLOYANNIS N., TERRENOIRE M., TOUNISSOUX D. *Pertinence for a classification.* IFCS-98, International Conference Data Science, Classification and Related Methods, Rome, Italy **(1998).**

C5 **[F5]** GAVIN G., DI PALMA S., ZIGHED D.A. *Generalisation properties of convex linear combinations of classifiers belonging to a finite VC-dimension family*. Fourth International Conference on Computer Science and Informatics. Proceedings of JCIS '98, Li K., Pan Y., Wang P.P.(eds) vol. IV, Pages 32-34, Duke University, Durham, NC **(1998)**.

C6 **[F6]** RAKOTOMALALA R., ZIGHED D.A., FESCHET F. *Empirical evaluation of rule characterization in rule induction process*. In Proceedings of the Fourteenth European Meeting on Cybernetics and Systems Research. Robert Trappl (Ed.) pages 799-804, Vienna, April 98 **(1998)**.

C7 **[F7]** RAKOTOMALALA R., LALLICH S. *Handling noise with generalized entropy of type beta in induction graphs algorithms*, in Proceedings of the Fourth International Conference on Computer Science and Informatics - JCIS'98, Li K., Pan Y., Wang P.P.(eds) vol. IV, pages 25-27, Duke University, Durham, NC **(1998)**.

C8 **[F8]** TOUGNE L. *The moirés of circles*. In Proceedings of the IEEE Southwest Symposium of Image Analysis and Interpretation (SSIAI'98). pages 115-120. Tucson, Arizona, USA **(1998)**.

C9 **[F9]** SARRUT D. *ARAMIS: an «on line» Parallel Platform for Medical Imaging*. In H.R. Arabnia (Ed.), Proc. of the International Conference on Parallel and Distributed Processing Technique and Applications, Las Vegas, USA, July 98, pages 509-516, CSREA Press, Athens, GA, USA (**1998)**.

C10 **[F10]** LALLICH S. *ZigZag, a new clustering algorithm to analyze categorical variable cross classifications tables,* in Proceedings of the 3th European Conference on Principles and Practice of *Data mining* and Knowledge Discovery, Zytkow J. M., Rauch J. (Eds) PKDD'99, Prague, pages 398-405, LNAI, Springer-Verlag, Berlin **(1999).**

C11 **[F11]** BERTOLINI C., PAUGAM-MOISY H., PUZENAT D. *Priming an Artificial Associative Memory*. In J. Mira and J.V. Sanchez-Andres (eds), Proc. Of International Workshop on Artificial Neural Networks, WANN'99, Alicante, SPAIN**,** LNCS, 1606:348-356, Springer Verlag, Berlin **(1999).**

C12 **[F12]** GUERMEUR Y., ELISSEEFF A., PAUGAM-MOISY H. *Estimating the sample complexity of a multiclass discriminant model.* In Proc. of 9th International Conference on Artificial Neural Networks, ICANN'99, Edimbourg, UK, pages 310-315, IEEE Computer Society Pres, Los Alamitos, CA, USA **(1999)**.

C13 **[F13]** FESCHET F., TOUGNE L. *Optimal time computation of the tangent of a discrete curve : application to the curvature.* In Proceedings of Discrete Geometry for Computer Imagery (DGCI'99), Lecture Notes in Computer Science 1568, pages 31-40. Springer Verlag, Berlin **(1999)**.

C14 **[F14]** RAKOTOMALALA R., LALLICH S., DIPALMA S. *Studying the behaviour of generalized entropy in induction trees using a m-of-n concept,* In Proceedings of the Third European Conference on Knowledge Discovery in Databases, sept. 1999, LNAI vol. 1704 : 510-517, Springer Verlag, Berlin **(1999)**.

C14 **[F14bis]** SEBBAN M., ZIGHED D.A., DI PALMA *Selection and Statistical Validation of Features and Prototypes*, In Proceedings of the Third European Conference on Knowledge Discovery in Databases, sept. 1999, LNAI vol. 1704 : 184-192, Springer Verlag, Berlin **(1999)**.

C15 **[F15]** GAVIN G., PUZENAT D., ZIGHED D.A. *How to use linear combinations of functions to get classifiers with low decision making error.* In proceedings of « International Conference on Artificial Intelligence » (ICAI'99), Las Vegas, juin 99, vol(1):226-232 **(1999)**.

C16 **[F16]** GAVIN G., ELISSEEFF A. *Confidence bounds for the generalization performances of linear combination of functions.* Ninth International Conference on Artificial Neural Networks (ICANN), University of Edinburgh, UK, Sept. 1999 **(1999)**.

C17 **[F17]** SARRUT D., MIGUET S. ARAMIS: *A Remote Access Medical Imaging System.* In 3rd International Symposium on Computing in Object-Oriented Parallel Environments, San Francisco, CA, pages 55-60. Lecture Notes in Computer Science, Springer-Verlag, Berlin **(1999).**

C18 **[F18]** SARRUT D., FESCHET F. *The Partial Intensity Difference Interpolation.* In H. R. Arabnia (Ed.) International Conference on Imaging Science, Systems and Technology, Las Vegas, USA, July 99, pages 46-51 CSREA Press, Athens, GA, USA **(1999)**.

C19 **[F19]** SARRUT D., MIGUET S. *Fast 3D Images Transformations for Registration Procedures.* In 10th International Conference on Image Analysis and Processing, pages 446-452, IEEE Computer Society Press, Los Alamitos, CA, USA **(1999)**.

C20 **[F20]** MIGUET S., SCUTURICI V.M. *Using the Hyper Text Transfer Protocol for Multimedia Streaming*. International Symposium on Intelligent Multimedia and Distance Education, Baden-Baden, Germany, august 99, pages 215-223 **(1999).**

C21 **[F21]** MIGUET S, PIERSON J.M. *Guaranteed Heuristics for 1D rectilinear partitioning,* In Proc. Sixth International Workshop on Parallel Image Processing and Analysis, Madras Christian College, pages 257-266 **(1999).**

C22 **[F22]** SARRUT D., MIGUET S. *Similarity Measures for Image Registration*. In European Workshop on Content-Based Multimedia Indexing, IHMPT-IRIT Toulouse, France, October 99, pages 263-270 **(1999)**.

C23 **[F23]** DARMONT J., FROMANTIN C., RÉGNIER S., GRUENWALD L., SCHNEIDER M. *Dynamic Clustering in Object-Oriented Databases: An Advocacy for Simplicity*, In ECOOP 2000 Symposium on Objects and Databases, Sophia Antipolis, France, June 00; LNCS, Vol. 1944, pages 71-85, Springer Verlag, Berlin **(2000).**

C24 **[F24]** FESCHET F., TOUGNE L. *Discrete Spheres on cellular automata*. In Proceedings of Seventh International Workshop on Combinatorial Image Analysis (IWCIA'00), pages 11-21 **(2000).**

C25 **[F25]** BOUSQUET O., ELISSEEFF A. *Algorithmic Stability and Generalization Performance*. In Proc. of Neural Information Processing Systems, NIPS*2000, (à paraître, mai 2001), MIT Press, Denver, USA **(2000)**.

C26 **[F26]** PAUGAM-MOISY H., ELISSEEFF A., GUERMEUR Y. *Generalization performance of multiclass discriminant models*. In Proc. of International Joint Conference on Neural Networks, IJCNN'2000, Como, Italie, IV: pages 177-182 **(2000)**.

C27 **[F27]** GUERMEUR Y., ELISSEEFF A., PAUGAM-MOISY H. *A new multi-class SVM based on a uniform convergence result.* In Proc. of International Joint Conference on Neural Networks, IJCNN'2000, Como, Italie, IV: pages 183-188 **(2000)**.

C28 **[F28]** BONNEVAY S., BOUNEKKARR A., LAMURE M., NICOLOYANNIS N. *Experimental analysis of color representation spaces. An attempt of synthetization.* Neural Computation'2000, Berlin, Germany **(2000).**

C29 **[F29]** SCUTURICI V.M., SCUTURICI M., *Video on Demand Using HTTP*, Protocols for Multimedia Systems - PROMS2000, Cracow, Poland, October 22-25, pages 228-235 **(2000)**.

C30 **[F30]** FRIEDT J.-M., TEYTAUD O., GILLET D., PLANAT M. *Simultaneous amplitude and frequency noise analysis in Chua's circuit - neural network based prediction and analysis*. In Proc. of VIII Van Der Ziel Symposium on quantum 1/f Noise and Other Low Frequency Fluctuations in Electronic Devices, St Louis **(2000)**.

C31 **[F31]** SARRUT D., CLIPPE S. *Patient positioning in radiotherapy by REGISTRATION of 2D portal to 3D CT images by a contend-based research with similarity measures*. In Computer Assisted Radiology and Surgery, San Francisco, USA, pages 707-712. Elsevier Science, Amsterdam, NL **(2000)**.

C32 **[F32]** CIAMPI A., ZIGHED D. A., CLECH J. *Trees and Induction Graphs for Multivariate Response*, In 4th PKDD Conference, D.A. Zighed, J. Komorowski, J. Zytkow *(Eds) Principles of Data mining and Knowledge Discovery*, pages 359-366, Lecture Notes in Artificial Intelligence, Springer Verlag, Berlin **(2000)**.

C33 **[F33]** LALLICH S., RAKOTOMALALA R., *Fast feature selection using partial correlation for multi-valued attributes*, In Proceedings of the 4th European Conference on Knowledge Discovery in DataBases, PKDD'2000, D.A. Zighed, J. Komorowski, J. Zytkow *(Eds) Principles of Data mining and Knowledge Discovery,* pages 221-231, Lecture Notes in Artificial Intelligence, Springer Verlag, Berlin **(2000)**.

C34 **[F34]** CHAUCHAT J.H., RAKOTOMALALA R. *Sampling Strategy for Targeting Rare Groups from a Bank Customer Database*, Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases. D.A. Zighed, J. Komorowski, J. Zytkow *(Eds)*, *Principles of Data mining and Knowledge Discovery*, pages 181-190, Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin **(2000).**

C35 **[F35]** CHAUCHAT J.H., RAKOTOMALALA R. *Sampling Strategy for Building Decision Trees from Very Large Databases Comprising Many Continuous Attibutes*, Proceedings of the 7th Conference of the International Federation of Classification Societies, Namur, Belgium. Kiers, Rasson, Groenen & Schader (Eds.), pages 199-204, Springer-Verlag, Berlin **(2000).**

C36 **[F36]** RITSCHARD G., NICOLOYANNIS N. *Aggregation and association in cross tables,* 4th PKDD Conference, D.A. Zighed, J. Komorowski, J. Zytkow *(Eds) Principles of Data mining and Knowledge Discovery*. Lecture Notes in Artificial Intelligence, pages 593-598, Springer Verlag, Berlin **(2000)**.

C37 **[F37]** S. MINIAOUI, J. DARMONT, O. BOUSSAID, *Web data modeling for integration in data warehouses*, 2001 International Workshop on Multimedia Data and Document Engineering (MDDE 01), Lyon, France, July 01 **(2001)**.

C38 **[F38]** TEYTAUD O., PAUGAM-MOISY H. *Bounds on the generalization ability of Bayesian inference and Gibbs algorithms.* Accepted in ICANN 2001. **(2001)**.

C39 **[F39]** GAVIN G. *Confidence bounds for one hidden-layer perceptrons*. In proceedings of "International Conference on Artificial Intelligence and Soft Computing» **(2001)** à paraître.

C40 **[F40]** CLIPPE S., SARRUT D. *Patient positioning in radiotherapy*. In 92nd Annual Meeting - American Association for Cancer Research, à paraître **(2001)**.

C41 **[F41]** COEURJOLLY D., MIGUET S., TOUGNE L. *Discrete Curvature based on Osculating Circle Estimation*. 4th international workshop on visual form (IWVF4'2001); à paraître **(2001).**

C42 **[F42]** GAVIN G., TEYTAUD O. *Equivalence in the worst case between the training error estimate and the Leave-One-Out estimate.* Acceptée à IJCNN **(2001)**.

C43 **[F43]** TEYTAUD O., JALAM R. *Kernel based text categorization*. Acceptée à IJCNN **(2001)**.

C44 **[F44]** SEBBAN M., NOCK R., LALLICH S. *Boosting Neighborhood-Based Classifiers, in Proceedings of the Eighteenth International Conference on Machine Learning* – ICML'2001, Williams College, Massachusetts, à paraître **(2001).**

C45 **[F45]** MINIAOUI S., DARMONT J., BOUSSAID O. *Web data modeling for integration in data warehouses*, 2001 International Workshop on Multimedia Data and Document Engineering (MDDE 01), pp. 88-97, Lyon, France, July 2001 **(2001)**.

C46 **[F46]** TEYTAUD O., SARRUT D. *Kernel Based Image Classification.* Accepted in ICANN 2001. **(2001)**.

C47 **[F47]** COEURJOLLY D., GERARD Y., REVEILLES J.P. ET TOUGNE L. *An elementary algorithm for digital arc segmentation*. International Workshop on Combinatorial Image Analysis (IWCIA'2001) **(2001)**

C48 **[F48]** FESCHET F**.** ET TOUGNE L. *An approach for the estimation of the precision of a real object from its digitisation.* International Workshop on Combinatorial Image Analysis (IWCIA'2001) **(2001)**

C49 **[F49]** SCUTURICI V.M., SCUTURICI M., MIGUET S., PINON J.M., *Measuring Web Servers Performance in VoD*, International Conference on Telecommunications (ICT2001), Romania, Bucharest **(2001)**

2.2.3 Conférences invitées dans les congrès internationaux

2.2.4 Autres publications

2.2.4.1 Thèses

[C1] CONTASSOT-VIVIER S. *Calculs parallèles pour le traitement des images satellites*. Thèse de doctorat de l'Ecole Normale Supérieure de Lyon **(1998)**.

[C2] FESCHET F. *Techniques d'imagerie pour l'aide au positionnement en dosimétrie conformationnelle*. Thèse de doctorat de l'INSA de Lyon **(1998).**

[C3] ELISSEEF A. *Etude de la complexité et contrôle de la capacité des systèmes d'apprentissage : SVM multi-classe, réseaux de régulation et réseaux de neurones multicouches*, Thèse de doctorat de l'Ecole Normale Supérieure de Lyon **(2000)**.

[C4] SARRUT D. *Recalage multimodal et plate-forme d'imagerie médicale à accès distant*. Thèse de doctorat de l'Université Lumière Lyon 2 **(2000)**.

[C5] GAVIN G. *Etude du Modèle d'Apprentissage Probablement Approximativement Correct : Application aux méthodes d'agrégation*. Thèse de doctorat de l'Université Lyon 2 **(2001)**.

2.2.4.2 Articles dans des revues nationales

[E1] RAKOTOMALALA R., ZIGHED D.A, FESCHET F. *Caractérisation des règles de production dans un processus d'induction*. Revue Electronique sur l'Apprentissage par les Données 2(1) **(1998)**.

[E2] ZIGHED D.A., SEBBAN M. *Sélection et Validation Statistique de Variables et de Prototypes*. Revue Electronique sur l'Apprentissage par les Données 2(1):1-21 **(1998)**.

[E3] BERGERON S., BOUSSAID O. *Localisation Industrielle des Connaissances, profil technologique des Firmes et des Secteurs : une étude du cas français, 1985-1990*. In Economies et Sociétés, Dynamique technologique et organisation, série W, 4(7):39-75 **(1998)**.

[E4] MIGUET S., SCUTURICI M. *Metrics for Measuring the Web Servers Performance*, Studia Universitatis «Babes-Bolyai», Series Computer Science, Cluj-Napoca, Romania, 1(1):85-96 **(1999).**

[E5] BELLET M., LALLICH S. *Externalités, modèle input-output et économie de l'innovation : quelques enseignements*, Economie Appliquée, 4e trim, pp. 7-33 **(1999).**

[E6] MUHLENBACH F., ZIGHED D.A., D'HONDT S. *Génération de règles par compression*, Extraction des connaissances et apprentissage (ECA), 1(1-2):93-104 **(2001)**.

[E7] JALAM R., TEYTAUD O. *Identification de la langue et catégorisation de textes basées sur les N-grammes*, ECA, 1(1-2):227-238 **(2001)**.

[E8] BENTAYEB F., BOUSSAID O., NICOLOYANNIS N. *Un Cadre Topologique pour OLAP*, ECA, 1(1-2):355-355 **(2001)**.

[E9] CHAUCHAT J.H., RAKOTOMALALA R., PELLETIER CH., CARLOZ M. *TQC: un indice d'évaluation du ciblage d'événements rares; une application en marketing*, ECA 1(1-2):155-159 **(2001)**.

[E10] RITSCHARD G., ZIGHED D.A., NICOLOYANNIS N. *Maximisation de l'association par regroupement de lignes ou de colonnes d'un tableau croisé,* Revue Mathématiques appliquées aux sciences humaines, à paraître **(2001).**

2.2.4.3 Articles dans des conférences nationales avec comité de lecture

[G1] CHAUCHAT J.H. *Estimation de la précision dans les enquêtes complexes*. XXVI° Colloque de l'A.R.A.E, Complexité Intelligence et décision '98 pages 76-79 **(1998)**.

[G2] COEURJOLLY D., SARRUT D., TOUGNE L. *Décimation en imagerie médicale 3d*. In Courbes Surfaces et Algorithmes, Journées du Groupe de Travail «Modélisation Géométrique» Grenoble, France, Septembre 99, pages 52-62 **(1999).**

[G3] DRONIOU J., ELISSEEFF A., PAUGAM-MOISY H., TEYTAUD O. *Contrôle de l'architecture et des représentations internes dans les réseaux de neurones multicouches*. In Actes de la Conférence sur l'Apprentissage, CAP'99, Palaiseau, FRANCE, pages 185-194 **(1999)**.

[G4] LALLICH S., RAKOTOMALALA R. *Les entropies de rangs généralisées en induction par arbres*, in Actes Colloque Société Française de Classification (SFC), Nancy, pages 101-107 **(1999)**.

[G5] LALLICH S. *Concept de diversité et association prédictive*, in Actes XXXIe Journées de Statistique, Société Française de Statistique, Grenoble, 17-21 mai 1999, pages 673-676 **(1999)**.

[G6] BERTOLINI C., PUZENAT D., PAUGAM-MOISY H. *Amorçage d'une mémoire associative artificielle.* In Actes du 3ème Colloque Jeunes Chercheurs en Sciences Cognitives, pages 18-24. Bordeaux, France **(1999)**.

[G7] REYNAUD E., PUZENAT D., PAUGAM-MOISY H. *Vers un modèle de mémoire associative multi-modale.* In Actes du 3ème Colloque Jeunes Chercheurs en Sciences Cognitives, pages 195-203. Bordeaux, FRANCE, **(1999)**.

[G8] KENYON C., PAUGAM-MOISY H. *Résolution du cas des 2-cycles, pour les dichotomies polyédriques dans R2.* In Actes de la Conférence sur l'Apprentissage, CAP'99, pages 195-204. Palaiseau, France **(1999)**.

[G9] ELISSEEFF A., PAUGAM-MOISY H., GUERMEUR Y. *Risque garanti pour les modèles de discrimination multi-classes.* In Actes du Colloque de la Société Française de Classification, SFC'99, pages 111-118. Nancy, France **(1999)**.

[G10] BONNEVAY S., BOUNEKKAR A., LAMURE M., NICOLOYANNIS N. *Aide à la prise de décision et «systèmes multi-agents»* Journée GDR - PRC I3 Information Interaction Intelligence, Paris, France **(1999).**

[G11] ELISSEEFF A. *Considérations sur une méthode de linéarisation pour les SVM.* In Actes de la Conférence sur l'Apprentissage, CAP'2000, St-Etienne, France, pages 99-116 **(2000)**.

[G12] MIGUET S., SARRUT D., CLIPPE C. *Positionnement de patient en radiothérapie conformationnelle.* In Avancées et perspectives en matière de GMCAO (Gestes Médico-Chirurgicaux Assistés par Ordinateur), Grenoble, sept 2000, pages 30-33 **(2000)**.

[G13] RAKOTOMALALA R., LALLICH S. *Sélection rapide de variables booléennes en apprentissage supervisé*, Actes de la 2ème Conférence Apprentissage, Saint-Etienne, CAP'2000, Hermes, pages 225-234 **(2000)**.

[G14] RITSCHARD G., ZIGHED D.A., NICOLOYANNIS N. *Regroupement optimal dans un tableau croisé.* Actes des XXXIIe Journées de Statistiques, Fès, MAroc, mai 00, pages 664-669, Société Française de Statistiques, Paris **(2000).**

[G15] LALLICH S. *Diversité, association brute et association partielle,* in Actes XXXIIe Journées de Statistique, Fès, Maroc, mai 2000, pages 503-507, Société Française de Statistique, Paris **(2000)**.

[G16] TEYTAUD O., LALLICH S. Bornes uniformes en extraction de règles d'association. CAp'01, Grenoble, France **(2001).**

[G17] SARRUT D., CLIPPE C. *Positionnement de patient en radiothérapie par recalage 2d/3d sans segmentation.* In Radiothérapie Assistée par l'Image (RAI 2001), Centre Antoine-Lacassagne, Nice, Mai 01 **(2001)**.

[G18] TWEED T., MIGUET S. *Sélection automatique de régions d'intérêt pour la détection de zones cancéreuses dans les mammographies* in : Coopération Analyse d'Image et Modélisation, 14 Juin 2001, Lyon, Université Claude Bernard, Lyon **(2001)**.

[G19] TWEED T., SAADANE A. *L'apport d'un bloc de segmentation d'erreur dans l'évaluation de la qualité d'images codées.* GRETSI 2001, 10-13 Sept. 2001, Toulouse **(2001)**.

[G20] JALAM R., TEYTAUD O. *Identification de la langue et catégorisation de textes basées sur les N-grammes*, ECA, Vol. 1, n°1-2/2001, pages 227-238 **(2001)**.

2.2.5 Les activités internationales

2.2.5.1 Collaborations avec des universités étrangères

2.2.5.1.1 Université de Tunis et Université de Sfax (Tunisie)

Identification du partenaire

- Département d'informatique de l'Université de Tunis et l'Ecole Nationale d'Informatique de Tunis.
  Professeur Ali Jaoua
- Département d'informatique de l'Université de Sfax
  Professeur Abdelmadjid Benhamadou

Collaboration en enseignement

- Depuis 1995 à 1999, D.A. ZIGHED assure, en tant que professeur invité, un cours de 24h dans le DEA d'informatique de l'Université de Tunis, sur l'ingénierie des connaissances.
- Depuis 2001, D.A. ZIGHED assure, en tant que professeur invité, un cours de 20h dans le DEA d'informatique de l'Université de Sfax, sur le *Data mining*

Collaboration en recherche

Dans le cadre de cette collaboration, D.A. ZIGHED co-encadre avec le Professeur A. Jaoua, deux étudiants en thèse à l'Université de Tunis : Mondher Maddourri, et Moncef El-Loumi qui

sont actuellement en 3ème année de thèse.

Le projet de recherche avec l'Université de Sfax qui est en encore à son tout début, portera sur les méthodologies de text mining.

Thématique de recherche

L'équipe tunisienne travaille sur les liens qui existent entre les problèmes de décomposition optimales de relations binaires et l'apprentissage supervisé. Outre les comparaisons avec les graphes d'induction que nous développons, nous nous intéressons plus particulièrement à la notion d'apprentissage optimal. Les premiers résultats de cette collaboration sont en cours de valorisation.

Perspectives

Nous espérons déboucher sur des publications majeures dans le domaine à l'issue de la finalisation des deux thèses tunisiennes.

2.2.5.1.2 Université York à Toronto

Identification du partenaire

Département d'informatique de Glendon College l'Université York à Toronto.

Professeur Eugène Roventa

Collaboration en enseignement

En 1998, D.A. ZIGHED a assuré, en tant que professeur invité, un cours de 40h destiné à des étudiants en fin de second cycle en informatique, sur la fouille de données.

En 1999, le Professeur Eugène Roventa est venu assurer un cours sur la logique floue et ses applications dans le cadre du DESS IIDEE.

Collaboration en recherche

Le Professeur E. Roventa est un spécialiste de la logique floue et de ses applications en informatique. Les graphes d'induction nous paraissent particulièrement adaptés pour une prise en compte des données floues et imprécises dans des problèmes d'apprentissage. Nous avons engagé une collaboration pour étudier cette question de façon très approfondie. Le Professeur E. Roventa a été invité par l'Université Lumière pour un séjour de 3 mois durant lequel nous approfondi nos recherches.

2.2.5.1.3 Université Laval à Québec

Identification du partenaire

Département d'informatique de l'Université Laval à Québec.

Professeurs : Nadir Belkhiter, Guy Mineau, André Gamache, Jean-Marie Beaulieu

Collaboration en enseignement

En 1998, D.A. ZIGHED a assuré durant la session de printemps, en tant que professeur invité, un cours de 40h au niveau de la maîtrise (équivalent du DEA français) d'informatique de l'Université de l'Université Laval, sur l'extraction automatique des connaissances à partir des données.

L'Université Laval à Québec est particulièrement intéressée par notre D.E.A. sur la fouille des données[56]. Elle est intéressée, à la fois, par les aspects scientifiques de ce DEA qui propose une formation par la recherche aux problèmes de stockage des grandes bases de données et l'extraction automatique des connaissances et par les aspects pédagogiques novateurs car ce DEA utilisera abondamment les nouvelles technologies éducatives : Visio-conférences , Internet,…

L'un des modules de base du DEA ECD est assuré par le Professeur Nadir Belkhiter. Il viendra, à partir de septembre 2001, passer une année sabbatique à l'Université Lyon 2 et poursuivre ses recherches au laboratoire ERIC.

Collaboration en recherche

Grâce aux compétences du Professeur Belkhiter dans le domaine des interfaces de communication homme-machine, nous voulons engager une réflexion méthodologique sur les interfaces et le *data mining*. En effet, les utilisateurs de ces techniques sont potentiellement très nombreux mais, ces outils ne seront réellement utilisés que s'il sont facile à appréhender. Cette recherche visera à étudier les modes d'interaction et les techniques de visualisation dans le domaine de l'ECD.

Perspectives

Une convention officielle définissant les échanges entre nos deux équipes est en cours d'approbation par les instances des deux

[56] Il s'agit du DEA ECD national commun aux Universités : Claude Bernard Lyon 1, INSA de Lyon, Lumière Lyon 2, Nantes et Paris 11 Orsay.

établissements.

2.2.5.1.4 Université STHB d'Alger

Identification du partenaire

Institut d'Informatique et institut d'Electronique. Mme Fatima Boumghar ; Mme Baya Ousséna.

Collaboration en recherche

Le pôle Images a des relations privilégiées avec l'Université des Sciences et de la Technologie « Houari Boumediène « (USTHB) d'Alger. Ces relations se sont concrétisées par l'invitation à plusieurs reprises de Mme Fatima Boumghar, chargée de cours à l'Institut d'Électronique et de Mme Baya Ousséna, chargée de cours à l'Institut d'Informatique. Serge Miguet a co-dirigé la thèse d'état de Fatima Boumghar qui a été soutenue en 1998. Plusieurs travaux communs ont pu être réalisés dans le cadre de notre projet d'imagerie médicale [D1].

2.2.5.1.5 Université de Genève

Identification du partenaire

Département d'économétrie de l'Université de Genève.

Professeur Gilbert Ritschard

Collaboration en enseignement

De 1999 à 2000, D.A. ZIGHED a assuré en tant que professeur invité, un cours de 40h au niveau de la licence d'économétrie de l'Université de l'Université de Genève, sur l'extraction automatique des connaissances à partir des données.

En échange le Professeur G. Ritschard intervient depuis 1999 en tant que professeur invité dans un cours en DEA ECD sur les mesures d'association.

Collaboration en recherche

Nous travaillons avec le Professeur G. Ritschard depuis de nombreuses années et nous avons déjà de nombreuses publications communes.

Perspectives

Nos futurs travaux s'orientent vers les mesures d'associations et leur utilisation en *data mining*.

2.2.5.2 Participation à des manifestations scientifiques internationales

Nom de la manifestation	*Participation des membres d'ERIC*
International Conference on Machine Learning, **(ICML)**	. **Zighed D. A**., membre du comité de programme, depuis 2001.
European Conference on Machine Learning, **(ECML)**	. **Zighed D. A**., Membre du comité de programme, depuis 2000.
European Conference on Principles of *Data mining* and Knowledge Discovery **(PKDD)**	. **Zighed D. A**., membre du comité de programme, depuis 1998, co-Président de la conférence PKDD2000 et Président du steering committee depuis septembre 2000. . **Rakotomalala R.** , membre du comité de programme depuis 2000. . **Chauchat J.-H.** membre du comité de programme du workshop « *data mining* for marketing applications ».
International Symposium on Modelling Intelligent Systems **(ISMIS)**	. **Zighed D. A**. Co-Président du comité de programme depuis 2001. .**Nicoloyannis N.**. Membre du comité de programme depuis 2001. . **Boussaid O.** membre du comité local d'organisation
Conférence Apprentissage **(CAP)**	. **Zighed D. A.** Co-Fondateur de la conférence , membre du comité de pilotage et du comité de programme depuis 1999. . **Paugam-Moisy H.** Co-Fondateur de la conférence , membre du comité de pilotage et du comité de programme depuis 1999. . **Rakotomalala R.** , membre du comité de programme.
Extraction et Gestion des connaissances **(EGC)**	. **Zighed D. A.** Co-fondateur de la conférence, membre du comité de pilotage (EGC2001, 1ère édition), membre du comité de programme et co-président

	de la conférence EGC2002 qui se tiendra à Montpellier, . **Nicoloyannis N.,** membre du comité de programme, . **Darmont J.,** membre du comité de programme, . **Chauchat J.-H**., membre du comité de programme, . **Miguet S.,** membre du comité de programme.
Journées Francophones de la Société Française de Classification **(SFC)**	. **Zighed D. A.**, membre du comité de programme depuis 1997
Journées de la Société Française De Statistique **(SFDS)**	. **Zighed D. A.**, membre du comité de programme depuis 1997, Président des journées SFDS2003.
Fourth Joint Conference on information Science, Durham (**CIS**).	. **Zighed D. A.**, membre du comité de programme depuis 1998, organisateur de deux sessions.
14èmes Journées Bases de Données Avancées – Colloque *Data mining* - Hammamet, Tunisie, 27 Oct. 1998.	. **Zighed D. A.**, membre du comité de programme depuis 1998,
CAPS'98 ; 1-3 Juillet 1998 ; Université de Caen, France Deuxième Colloque International sur l'Apprentissage Personne-Système.	. **Zighed D. A.**, membre du comité de programme depuis 1998,
International Workshop on Combinatorial Image Analysis (**IWCIA**)	. **Miguet S.,** membre du comité de programme depuis 1995.
Discrete Geometry for Computer Imagery (**DGCI**)	. **Miguet S.,** membre du comité de programme depuis 1996.
Multimedia Data and Document Engineering (**MDDE**)	. **Boussaid O.,** membre du comité de programme 2001.
Re Technologies for Information System	. **Boussaid O.,** membre du comité de programme 2001. . **Darmont J.,** membre du comité de programme 2001.

IEEE Transaction and Knowledge and data Engineering (2000)	**. Darmont J.,** membre du comité de programme 2001.
Neuronal Information Processing Systems (**NIPS**)	**. Paugam-Moisy H.,** membre du comité scientifique et du comité de programme (99,2000)
International Conference on Artificial Neural Networks (**ICANN**)	**. Paugam-Moisy H.,** membre du comité scientifique et du comité de programme (98,99)
European Symposium on Artificial Neural Networks (**ESANN**)	**. Paugam-Moisy H.,** membre du comité scientifique et du comité de programme, depuis 96

2.2.6 Les contrats de recherche

2.2.6.1 Clinique mutualiste «La Roseraie »

Identification du partenaire Le Centre Hospitalier Mutualiste de Vénissieux : Jean-Luc PONCET, Directeur Général, Jean BURDIN-FRENEA, Responsable Informatique, J.H MOLLET, Médecin.

Objectifs recherchés La clinique collecte et gère des masses de plus en plus importantes d'informations. Ses bases de données sont très peu exploitées alors que leur mise en valeur permet d'extraire des nouvelles connaissances et d'éclairer des décisions. Cette collaboration vise la mise au point d'un processus d'exploitation continue des données de la clinique en vue d'extraire des connaissances relatives à la clientèle, aux soins, à la gestion,…

Montant et durée du contrat 200 KF TTC sur 2 ans (1998-1999)

2.2.6.2 Contrat BOTANIC

Identification du partenaire Les magasins Botanic : E. Bouchet

Objectifs recherchés 1) Les magasins BOTANIC disposent d'un fichier de plus de 70 000 clients où sont enregistrées des informations signalétiques

d'une part, et des informations concernant leurs visites (et achats) dans leurs magasins d'autre part. L'exploration de leurs bases de données permettrait, outre le nettoyage des données, de dresser une typologie de leur clientèle afin de concevoir des opérations de marketing ciblé.

2) Depuis cinq ans, les ventes quotidiennes dans chaque magasin pour 130 familles de produits sont conservées en mémoire. Il nous est demandé :

a) de modéliser ces séries en fonction des mois dans l'année, des jours dans la semaine, du calendrier des jours fériés, des vacances scolaires, des fêtes particulières, et de la météo ;

b) de produire un prototype de traitement automatique de ces séries pour : i) analyser l'activité passée et évaluer les effets des actions commerciales de chaque magasin et de ses concurrents, et ii) prévoir les approvisionnements et les emplois du temps du personnel .

3) De notre côté, ces bases de données réelles nous servent de terrain expérimental pour l'étude des procédés d'échantillonnage et de leur optimisation.

Montant et durée du contrat 43 KF HT pour une durée de 6 mois.

2.2.6.3 Santé et HPC

Identification du partenaire Le projet Santé & HPC est une collaboration entre des laboratoires lyonnais et grenoblois, soutenue par la région Rhône-Alpes dans le cadre de la réponse à un appel d'offres sur le thème «opto-électronique et vision». Nos principaux partenaires au sein du projet sont Jacques Demongeot, Jocelyne Troccaz et Laurent Desbat pour le laboratoire TIMC de Grenoble, Isabelle Magnin pour le laboratoire CREATIS de l'INSA de Lyon, Michel Jourlin pour le LISA de l'école CPE de Lyon et Lionel Brunie pour le LIP de l'ENS de Lyon.

Objectifs recherchés

Notre objectif au sein de ce projet est de définir un système de gestion et d'interrogation de bases de données d'images médicales multimodales. Nous tirons partie des développements en réseaux locaux haut débit afin de proposer un nouveau type de serveurs d'images à hautes performances, basé sur des composants du marché. Ces architectures offrent la possibilité d'une part de stocker et de transférer rapidement des ensembles d'images volumineux entre des sites de production différents, et d'autre part de traiter en parallèle certaines des requêtes les plus demandeuses en temps de calcul. Nous voulons rendre l'accès à ces ressources de stockage et de calcul le plus transparent possible, grâce à la conception d'interfaces d'interrogations écrites en Java, utilisables depuis des postes banalisés sur internet.

Montant et durée du contrat

70 KF HT pour une durée de 3 ans.

2.2.6.4 Adémo

Identification du partenaire

Le projet ADEMO est une collaboration entre des laboratoires lyonnais et grenoblois, soutenue par la région Rhône-Alpes dans le cadre de la réponse à un appel d'offres sur le thème «sciences et technologies de l'information, outils et applications». Nos principaux partenaires au sein du projet sont les laboratoires CREATIS de l'INSA de Lyon, TIMC de l'UJF Grenoble, TSI de l'Université de Saint-Etienne et le LIGIM de l'UCB Lyon 1.

Objectifs recherchés

Les travaux de recherche du projet « adémo » concernent l'étude de techniques d'imagerie permettant d'estimer puis de corriger le positionnement de patients en radiothérapie conformationnelle. Dans ce cadre, nous faisons partis du thème modélisation et suivi spatio-temporel pour le diagnostic et la thérapie, dont la coordinatrice est J. Troccaz (GMCAO, UJF, Grenoble).

Montant et durée du contrat

250 KF pour une durée de 3 ans.

2.2.6.5 Diagnos

Identification du partenaire Diagnos Inc. Société Canadienne basée à Montréal.

Objectifs recherchés Mammo est un système d'aide au diagnostic pour la détection des cancers de sein. L'objectif de ce projet est d'identifier les paramètres radiologiques, cliniques et ou biologiques et leurs interactions en vue de mettre au point un automate capable de reconnaître la nature cancéreuse ou non d'une mammographie.

Montant et durée du contrat 1 MF sur 14 mois

2.2.6.6 ELF ANTAR

Identification du partenaire Elf Antar

(Hélène Paugam-Moisy)

Objectifs recherchés Il s'agit de concevoir une stratégie incrémentale d'ajout de données d'apprentissage à partir d'un jeu d'observation minimal dédié à l'apprentissage afin d'obtenir un modèle optimal du point de vue de ses capacités de généralisation avec ce sous ensemble de données d'apprentissage de dimension minimale.

L'étude a porté sur les points suivants : état de l'art sur la *feature selection*, conception des stratégies de sélection par méthodes à noyaux, codage des algorithmes sous Matlab, tests sur jeu de données réelles, étude de la sensibilité sur les méthodes de sélection.

Montant et durée du contrat 25 000F HT, 3 mois (en 99)

28 800F HT, 3 mois (en 2000)

2.2.6.7 INRIA

Identification du partenaire INRIA Rocquencourt

(Hélène Paugam Moisy)

Objectifs recherchés L'objectif de cette recherche est de mettre au point des tests et des modèles mettant en évidence le codage et la représentation

obtenus selon des traitements catégoriels ou métriques et d'étudier leur conséquences sur les performances des sujets et des modèles. Il s'agit de montrer que le type de codage réalisé a des répercussions sur la performance et d'étudier la nature du lien entre type de codage, type de lien et performance.

Montant et durée du contrat 62 700 F HT 2 ans

2.2.6.8 Région Rhône-Alpes

Identification du partenaire

- UMR 5823 LASS, Université Claude Bernard Lyon 1
- UMR 5515 CREATIS, Université Claude Bernard Lyon 1
- UMR 5015 ISC, Université Claude Bernard Lyon 1
- TIMC, Université Joseph Fourier Grenoble

Objectifs recherchés

Objectifs généraux du projet :

- amélioration de la prise en charge de patients cérébro-lésés et de patients ischémiques.
- constitution d'une banque de données anatomo-fonctionnelle aidant à la compréhension des mécanismes cérébraux et cardiaques chez l'homme.
- Création du premier centre distribué national de recueil de données médicales anatomiques et fonctionnelles de patients cérébro-lésés d'une part et de patients atteints d'affections du myocarde d'autre part.

Montant et durée du contrat

Durée : 3 ans

Une bourse de trois ans, dont bénéficie Pierre-Emmanuel Jouve.

Fonctionnement 180 KF

Investissement 20 KF

2.2.6.9 CCF

Identification du partenaire Crédit Commercial de France

Objectifs recherchés Recherche de méthodes de ciblage de la clientèle tenant compte des bénéfices et coûts attendus. Méthodes d'optimisation du bénéfice global espéré : on a mis au point une représentation graphique de la liaison entre la taille de la cible optimisée et bénéfice global espéré. Les difficultés de l'extraction de connaissance à partir d'une base de données sur laquelle des actions marketing ont déjà été réalisées. Importance des «méta-données» retraçant ces actions passées. Propositions de plans d'expérience en vue de l'amélioration du ciblage. Les résultats sont présentés à PKDD'01, dans le cadre du Workshop «Dataming & Marketing».

Montant et durée du contrat 60.000 F ht sur 6 mois

2.2.6.10 Crédit Agricole du Centre-Est

Identification du partenaire Crédit Agricole du Centre-Est

Objectifs recherchés Objectif de la recherche : Datamining à partir de la Base de données marketing et mise au point d'un système d'enquête de satisfaction des clients. Une méthode d'échantillonnage dans la base en vue de l'apprentissage a été mise au point et testée. Les résultats ont été présentés à PKDD'00. La rénovation du système de suivi de la satisfaction des clients, par segment d'une part, par agence d'autre part, aboutit a un modèle série d'enquêtes répétables.

Montant et durée du contrat 485.000 F ht sur 18 mois

2.2.6.11 Observatoire National du Tourisme

Identification du partenaire Observatoire National du Tourisme

Objectifs recherchés Analyse critique du système de mesure du tourisme étranger en France, pour améliorer les outils de pilotage de la politique du Ministère français du Tourisme. Les deux grandes méthodes d'observation utilisées ont été décortiquées : 1) l'enquête annuelle aux frontières sur un échantillon de points de passage et à un échantillon de périodes, réalisée par l'administration ; 2) les enquêtes régulières auprès des ménages résidants dans les principaux pays du monde, concernant les voyages à l'étranger, réalisées par des organismes privés.

Ce rapport d'expertise à été réalisé en collaboration avec l'INSEE.

Montant et durée du contrat 49.000 F ht sur 6 mois

2.2.6.12 APRIL-Assurance

Identification du partenaire APRIL-Assurance

Objectifs recherchés Objectif de la recherche : Segmentation stratégique des correspondants commerciaux, avec mise à jour annuelle des segments. Il s'agit de produire une classification non-supervisée des correspondants, sur laquelle s'appuie la politique commerciale (par exemple : les très gros agents, peu nombreux, bénéficient d'un support commercial et technique qu'on ne peut pas apporter aux très nombreux petits agents). Une fois la segmentation mise au point, elle doit être assimilée par les collaborateurs de l'entreprise, et la structure des segments doit donc rester stable d'année en année, tout en permettant de transférer les individus d'un segment à l'autre et aussi d'affecter les nouveaux individus à l'un des segments.

Montant et durée du contrat 122.000 F ht sur 12 mois

2.2.6.13 France Télécom

Identification du partenaire France Télécom R&D : Jean-Marc PITIE

Objectifs recherchés L'objectif de cette recherche concerne la sélection et la construction d'attributs pour la préparation des données appliquées aux grandes bases de données gestion-clients pour la réalisation du projet Customer Constructive Inductive Learning.

Montant et durée du contrat 656.000 F ht sur 24 mois

2.2.6.14 Fondation Védior Bis

Identification du partenaire Fondation Védior Bis : Jacques PHILIPPE, Président

Objectifs recherchés

Montant et durée du contrat 800.000 F ht sur 36 mois

2.2.7 Les brevets licenciés

2.2.8 La valorisation et le partenariat industriel

Différents projets d'élaboration de logiciel dans le cadre des travaux de thèse ont été mis en place depuis la création du laboratoire. Ces projets ont pour objectif de créer des outils destinés au test et à l'évaluation des développements théoriques réalisés par les chercheurs. Ils sont également mis en œuvre dans les contrats passés par le laboratoire.

2.2.8.1 SIPINA

Le projet SIPINA_W a été initié à l'été 1995, ce logiciel destiné à l'induction par graphes répondait à la préoccupation de produire une plate-forme suffisamment puissante pour les tests effectués dans les thèses respectives de S. Rabaséda et R. Rakotomalala.

Une diffusion mondiale du logiciel a été assurée par sa mise à disposition gratuite sur Internet, sur les pages Web du laboratoire ainsi que sur les pages Web du forum KDD Nuggets (http://www.kdnuggets.com).

Plusieurs centaines d'utilisateurs dans le monde sont recensés à ce jour, une dizaine d'entre eux ayant publiés des thèses ou des articles scientifiques en utilisant les résultats obtenus à l'aide du logiciel. Par ailleurs, SIPINA_W est utilisé en enseignement dans plusieurs universités : l'Université Lumière Lyon 2, l'Université Joseph Fourrier de Grenoble, l'école ESIEA (Ecole Supérieure d'Informatique Electronique Automatique) de Paris… (et Chambéry ?)

Un nouveau projet a débuté cette année sur le même modèle dans le cadre d'une collaboration avec la société Diagnos qui est en train de mettre au point une version professionnelle destinée à la commercialisation. Dans cette optique, de nouveaux outils d'extraction de connaissances à partir de données sont mis en place par notamment des outils d'extraction de connaissances à partir de données textuelles et images.

2.2.8.2 CSTI

Nous avons débuté en avril 1998 une collaboration avec l'INSA de Lyon et la société Compagnie des Signaux, Traitement de l'Information, installée à Bron. Deux boursiers roumains financés par le CNOUS effectuent leur thèse dans le cadre de cette collaboration l'un au laboratoire ERIC sous la direction de Serge Miguet et l'autre au LISI de l'INSA de Lyon, sous la direction de Jean-Marie Pinon. La société CSTI travaille au développement de l'architecture à haute performance ANTARA dédiée aux applications multimédia (vidéo à la demande, compression, réalité virtuelle, etc.). Notre implication dans ce projet porte sur la modélisation et l'optimisation d'un serveur WEB pour cette architecture. Bien que CSTI ait maintenant cessé ses activités sur Lyon, la thèse de Marian Scuturici pourra être soutenue en novembre 2001.

2.2.9 L'information scientifique et la vulgarisation scientifique

2.2.9.1 Colloques organisés par ERIC

2.2.9.1.1 Journées du Groupe de Travail sur l'Apprentissage

Le GTRA est un groupe de travail, intégré à l'AFIA, qui s'est fixé pour objectifs de :

- Créer un cadre fédérateur pour la communauté française qui s'intéresse à l'apprentissage et qui se trouve dispersée, à la fois géographiquement et par disciplines (RDF, IA, Analyse de Données, Informatique, Statistiques, Connexionnisme)
- Favoriser la circulation des informations au sein cette communauté sous forme de résumés de thèses, d'articles, d'actes de congrès, de pointeurs électroniques sur des sources d'informations (sites ftp, www, etc...)
- Favoriser le développement de projets communs permettant de tirer parti des potentialités des différentes équipes.

Il a été créé en 1995 par le professeur D. A. Zighed qui l'a co-animé avec le professeur Hélène Paugam-Moisy jusqu'en juin 2000. L'activité du groupe est structurée autour de six thèmes :

a) Systèmes Hybrides (Responsable : G. Verley, université de Tours)
b) Apprenabilité et Robustesse (Responsable : H. Paugam-Moisy, ENS Lyon)
c) Apprentissage et Interaction Homme-Machine (Responsable : M. Quafafou, Université de Nantes)
d) Approches constructives en Apprentissage (Responsables : F. D'Alche-Buc, Université Paris 6, et R. Rakotomalala, Université Lyon2)
e) Programmation Logique Inductive (Responsable : C. Vrain, Université d'Orléans)
f) Apprentissage distribué (Responsable : P. Beaune, Ecole des Mines de St Etienne)

Trois réunions annuelles ont été organisées sans interruption depuis la création de ce groupe de travail. Le laboratoire ERIC en a accueilli les 2/3.

2.2.9.1.2 PKDD2000

La conférence " Principles of *Data mining* and Knowledge Discovery **(PKDD)"** est une conférence européenne organisée depuis 1997. Après Trodheim (Norvège) en 1997, Nantes en 1998 et Prague en 1999, nous avons organisé l'édition 2000 de cette conférence. Plus de 150 papiers de 34 nationalités ont été soumis. Six workshops et six tutoriaux ont eu lieu du 14 au 18 septembre 2000. Les actes ont été publiés par Springer Verlag dans la collection Lecture Notes in Computer Science et distribués le jour même de la conférence.

2.2.9.1.3 SFdS 2003

La Société Française de Statistique organise annuellement dans une ville francophone les journées de statistique. Pour 2003, le comité scientifique de la SFdS nous a demandé d'accueillir à Lyon cette manifestation qui réunit, à chaque occasion, plus de 600 personnes.

2.2.9.2 Les séminaires d'ERIC

Depuis octobre 1994, des séminaires sont organisés au sein du laboratoire ERIC, avec une fréquence bimensuelle entre le mois d'octobre et le mois de juin. Ces séminaires, ouverts au public, mobilisent des intervenants d'horizons différents. Les thèmes abordés sont essentiellement en relation avec les préoccupations du laboratoire : apprentissage, traitement d'images, bases de données décisionnelles,... Néanmoins, dans certains cas, des exposés plus généraux sur différents sujets relatifs à l'informatique ont lieu.

Dans le tableau ci-dessous sont résumés le nombre des séminaires organisés depuis octobre 1998, avec la répartition entre intervenants extérieurs et intervenants du laboratoire.

	Séminaires organisés
ERIC	45
Extérieurs	69
Total	114

Les objectifs des séminaires sont :

- mettre en relation les membres du laboratoire avec d'autres chercheurs dont certains viennent de l'étranger,

- obtenir un point de vue différent sur des problèmes qui entrent dans le cadre de l'activité de recherche du laboratoire,
- permettre aux chercheurs, notamment les doctorants, de présenter leurs travaux récents,
- mieux connaître des sujets connexes aux préoccupations du laboratoire.

2.2.9.3 Ecoles et Formations

2.2.9.3.1 Ecole KDD

Depuis 1997, ERIC organise tous les ans une Ecole sur l'Extraction automatique des connaissances dans les bases de données. Celles-ci ont réuni autour d'une vingtaine de participants dont une dizaine venant du secteur industriel. Les intervenants[57] sont des spécialistes internationaux venant d'Europe et du Canada.

2.2.9.3.2 CNRS-Formation

Depuis de nombreuses années, le catalogue de CNRS-Formation propose aux entreprises trois stages intensifs organisés à Lyon par ERIC. Le premier porte sur l'analyse des données qualitatives, le second sur l'analyse des données quantitatives et le troisième sur l'extraction automatique des connaissances à partir des données.

2.2.9.4 Publication d'une revue

READ (Revue Electronique sur l'Apprentissage par les Données) est une revue électronique qui a été créée en décembre 1996 à l'initiative des membres du groupe de travail sur l'apprentissage (GTRA), et plus particulièrement de Zighed A. (ERIC, Lyon2), G. Venturini (LI, Tours) et M. Sebban (RAPID, Univ. Guadeloupe).

Cette revue dont la thématique s'est élargie aux domaines de l'extraction des connaissances est passé au format papier. Elle a pour ambition de traiter et de débattre des thèmes liés à l'extraction des connaissances, aussi bien d'un point de vue théorique que d'un point de vue pratique.

[57] Y. Kodratoff (CNRS, Orsay France), A. Zighed (ERIC, Lyon France) ; F. Bergadano (Univ. Turin, Italie), L. Wehenkel (FNRS, Liège, Belgique), S. Mattwin (Univ. Ottawa, Canada), J. Korczack (Univ. Strasbourg, France).

Les contributions dans les domaines suivants (liste non exhaustive) sont plus particulièrement encouragées:

Apprentissage supervisé/non supervisé, Classification, Approximation de fonctions, Apprenabilité, Capacités de généralisation, Extraction de connaissances (KDD), Interactivité, Applications industrielles, Apprentissage par induction, Programmation logique inductive, Apprentissage par renforcement, Apprentissage symbolique, Apprentissage numérique, Réseaux de neurones artificiels, Algorithmes génétiques, Systèmes hybrides.

2.2.9.5 Vulgarisation scientifique

Nous avons participé à des opérations de vulgarisation scientifique, notamment à travers la manifestation « Fête de la Science 2000 «. Nous avons également participé à la rédaction d'un dossier sur l'utilisation de l'Image dans les laboratoire lyonnais dans la revue « Isotopes « du Pôle Universitaire Lyonnais.

Afin d'entretenir une synergie au sein du groupe BDD, des réunions hebdomadaires sont tenues. Elles permettent d'une part, de faire le point sur les recherches en cours et d'échanger des idées à leur propos et, d'autre part, d'assurer une veille grâce à différentes présentations techniques ou concernant des travaux de recherches extérieurs au laboratoire. Les membres du groupe BDD participent également à des groupes de travail (CAFE-BD, Paris, mars 2001). Ces différents contacts nous ont permis d'initier des collaborations avec d'autres équipes (LISI Lyon 1, ENSI Tunis).

Nous disposons par ailleurs d'un site web[58] servant à la fois de vitrine et de base de données bibliographique et documentaire.

[58] http://bdd.univ-lyon2.fr

2.3 DECLARATION DE POLITIQUE SCIENTIFIQUE POUR LA PERIODE 2003-2006

2.3.1 Prospective – Pôle apprentissage

Nos récents travaux sur l'apprentissage et son application dans l'Extraction automatique de Connaissances à partir des Données (ECD), nous ont persuadé que les gains d'efficacité à attendre d'éventuels raffinements des méthodes aujourd'hui établies en apprentissage et plus largement en reconnaissance de formes (Réseaux de neurones, Graphes, Statistiques…) sont relativement faibles. Plusieurs présentations et débats proposés, notamment lors de la dernière réunion du Groupe de TRavail sur l'Apprentissage (GTRA), consacrée justement à l'ECD, ont montré que cette opinion, qui constituait une des conclusions de la thèse de R. Rakotomalala en ce qui concerne les graphes d'induction, était partagée par de nombreux spécialistes du domaine. Les développements théoriques sur les méthodes d'apprentissage ont longtemps mobilisé l'essentiel des efforts réalisés en ECD, ce qui explique le degré élevé de sophistication atteint par ces méthodes aujourd'hui. Les autres étapes du processus d'ECD ont en revanche été souvent négligées, et mériteraient une attention accrue pour unifier et théoriser des connaissances pratiques peu valorisées. Le laboratoire ERIC a donc choisi d'orienter ses recherches vers la résolution de problèmes pratiques, cruciaux pour les utilisateurs professionnels des techniques d'ECD, mais encore peu abordés du point de vue théorique (du moins pour certains).

1. Concernant l'étape préalable à l'application des algorithmes d'apprentissage proprement dits, deux voies de recherches complémentaires sont actuellement explorées et seront approfondies:
   - La première, concerne la mise au point d'indicateurs de séparabilité. Dans bien des cas, en effet, les résultats médiocres tiennent moins à la mauvaise qualité de la méthode employée qu'à une représentation inadéquate des données du problème. Il est donc primordial de définir des indicateurs permettant de savoir, indépendamment de la méthode d'apprentissage employée, si la représentation des données retenue autorise la distinction des classes. Ces indicateurs peuvent de plus permettre de tester la pertinence des techniques de sélection d'attributs et de réduction de la dimensionalité rendues indispensables par la taille croissante des bases de données. Nous poursuivons cette

réflexion pour définir des indicateurs de séparabilité utilisables lorsque l'espace de représentation des données comprend des données hétérogènes : numériques-symboliques.

- La deuxième porte sur la définition de techniques d'échantillonnages prenant en compte les spécificités du problème à traiter et des méthodes d'apprentissage que l'utilisateur souhaite appliquer. Plus précisément, nous approfondirons les travaux portant sur l'amélioration des techniques d'échantillonnage pour l'apprentissage dans les très grandes bases de données. Ce sujet devient crucial car, la taille des bases à explorer augmentant plus vite que la puissance des machines, il est aujourd'hui souvent impossible d'appliquer les algorithmes d'apprentissage à l'ensemble des données disponibles. Les arbres de décision, comme la recherche de t-uples ou de règles, étant assez sensibles aux fluctuations d'échantillonnage, un soin tout particulier doit être apporté au procédé de choix de l'échantillon pour obtenir le résultat le plus stable et le plus proche possible de celui qu'on obtiendrait sur la base entière. Ce sujet est relativement vierge, du moins dans le cas de la classification supervisée, son aspect transversal associant des compétences en base de données et en statistique le positionne comme une voie de développement à long terme au sein du laboratoire.

2. Le second axe porte sur la définition d'une typologie des problèmes d'apprentissage à partir d'informations afférentes aux données, y compris les caractéristiques numériques telles que la taille de l'échantillon, le nombre de classes, le type des attributs prédictifs, les distributions conditionnelle et a priori des données. La mise en relation de cette typologie des problèmes d'apprentissage avec une catégorisation, encore à construire, des algorithmes d'apprentissage est un aspect important de la problématique de l'extraction de connaissances à partir de données. L'objectif de cette démarche est de permettre une sélection assistée des algorithmes à utiliser en fonction des problèmes à traiter. Notre étude actuelle porte sur une formulation généralisée des mesures d'évaluation des partitions en adoptant le point de vue de Aczel-Daroczy[59], en modulant un paramètre unique β, nous pouvons retrouver une grande partie des indicateurs référencées dans la littérature. De fait, en définissant un protocole d'expérimentation faisant varier graduellement le niveau et le type de bruit sur la variable à prédire, toutes choses égales par ailleurs, nos premiers résultats montrent bien une influence concomitante de la valeur du paramètre β et de la caractéristique de bruit associée au données sur les performances de l'apprentissage par graphes. Notons toutefois que l'apparition et la

[59] « On measures of information and their characterizations », J. Aczel , Z. Daroczy, Academic Press, 1975.

force de cette relation dépend de la complexité de la fonction à apprendre. Les premiers résultats sont en cours de publication.

3. Enfin, le troisième axe de recherche concerne l'agrégation de classifieurs. Cette voie, également née du constat d'une certaine stagnation des performances des méthodes traditionnelles d'apprentissage, est aujourd'hui en pleine expansion. Les méthodes auxquelles s'est déjà intéressé Gérald Gavin sont celles qui sélectionnent une sous-famille finie de classifieurs, qu'elles pondèrent et qui classent les exemples en procédant à un vote à la majorité pondérée.

   Initialement, nous avons étudié ces méthodes d'un point de vue pratique dans l'optique d'une intégration à la plate-forme logicielle développée au sein du laboratoire (SIPINA_W). Les efforts actuels portent sur la validation des connaissances que ces méthodes permettent d'obtenir. Des résultant théoriques prometteurs sont en cours de validation, nous sommes sur une justification théorique des heuristiques d'agrégation précédemment évoquées. Ces résultats, basés sur les travaux de Vapnick, permettent de borner l'erreur en généralisation indépendamment de la complexité du classifieur agrégé construit, contredisant ainsi un principe communément admis en apprentissage, celui du «rasoir d'Occam».

Tous ces efforts sont cristallisés à terme par l'élaboration d'une plate-forme complète d'ingénierie des connaissances, dans la lignée du logiciel SIPINA_W bien implanté maintenant au sein de la communauté scientifique. Toujours sous la forme de contrats avec l'industrie, de nouveaux projets de développement, en relation directe avec les thèmes de recherche du laboratoire, seront mis à profit pour compléter le processus complet d'extraction de connaissances.

2.3.2 Prospective – Pôle Images

Les projets récemment démarrés au sein du laboratoire ERIC et décrits dans la partie « bilan du pôle Images », nous permettent d'envisager des développements importants dans les domaines de recherche suivants :

1. fusion multimodale
2. systèmes multimédia à haute performance
3. fondements du traitement numérique de l'image
4. travaux relatifs à l'apprentissage

2.3.2.1 Fusion multimodale

Les nombreux dispositifs d'acquisition d'image utilisés notamment en imagerie médicale (tomographie X, IRM, TEP, etc.), donnent des informations partielles mais complémentaires sur les objets observés. Une importante étude bibliographique a été entreprise pour répertorier et comparer les techniques de mise en correspondance entre images issues de modalités différentes. Plusieurs auteurs ont montré que la théorie de l'information, couplée à des tests statistiques, pouvait servir de base à la construction de puissants critères de similarité entre les images. Ces outils semblent robustes mais demandent encore des temps de calculs importants. Nous nous proposons d'étudier ces approches afin d'en extraire des caractéristiques susceptibles d'accélérer le processus, au niveau de la stratégie de recherche du maximum global de la fonction de similarité, ainsi qu'au niveau de la robustesse de la mesure (différents test statistiques sont en cours d'évaluation). Nous voulons en particulier étudier la pertinence de ces approches dans le contexte de notre projet de radiothérapie conformationnelle, où l'une des étapes fondamentales consiste en une mise en correspondance entre images observées et images calculées.

Une autre voie dans laquelle nous allons nous investir est l'intégration des modèles déformables dans nos outils de fusion de données. Notre projet de positionnement du patient en radiothérapie ne s'applique pour l'instant qu'aux objets rigides. Mais les médecins du centre de lutte contre le cancer Léon Bérard souhaitent également pouvoir prendre en compte des objets déformables et mobiles, par exemple pour l'irradiation de cancers du poumon. Il s'agit alors d'être capables de décrire le mouvement de l'ensemble des points du poumon au long du cycle respiratoire. L'irradiation de la tumeur se ferait alors de manière fractionnée par tranches de 10 secondes, la respiration du patient étant bloquée par exemple à 70% du volume pulmonaire. Ces outils basés sur des modèles déformables pourraient également être utilisés dans le projet *mammo* dans le cadre d'une comparaison des deux seins. En effet, les spécialistes nous ont expliqué que l'un des indices permettant de déceler la présence d'une tumeur sur un cliché était basée sur la localisation d'opacités apparaissant sur l'un des clichés mais non sur l'autre.

2.3.2.2 Systèmes multimédia à hautes performances

Les développements démarrés au sein du projet « santé et HPC » de la région Rhône-Alpes ouvrent la voie à des perspectives de recherche particulièrement intéressantes. Les machines parallèles sont quasiment inutilisées dans le milieu hospitalier, malgré l'énorme potentiel qu'elles représentent en terme de puissance de calcul. Cela est principalement dû au fait que ces machines sont trop souvent d'un accès difficile, ce qui les réserve principalement à des informaticiens

spécialistes. Nous voulons poursuivre nos efforts dans la conception du prototype ARAMIS, permettant de rendre transparente pour des utilisateurs finals, l'utilisation de machines parallèles à travers des réseaux haut débit. Toute une chaîne de traitements d'images médicales 3D a été parallélisée avec succès dans le cadre de l'habilitation à diriger des recherches de Serge Miguet[60]. Nos objectifs dans ce projet consistent à étudier des méthodologies de conception d'applications permettant d'enchaîner un certain nombre de traitements de manière conviviale, depuis le navigateur web d'un poste banalisé. Un prototype a été produit dans le cadre du projet « Santé et HPC ». Nous poursuivons actuellement cette démarche en participant aux réunions du Work Package 10 du projet européen « datagrid », dont le but est d'étudier la distribution d'applications biologiques et médicales sur des grilles de calculs (systèmes distribués à grande distance).

2.3.2.3 Fondements du traitement numérique de l'image

Les projets dans lesquels nous sommes impliqués nous conduisent à l'étude de problèmes théoriques concernant l'analyse d'images et l'algorithmique de l'image, qui sont valorisés dans le cadre d'un groupe très actif au niveau national sur la géométrie discrète.

2.3.2.3.1 Géométrie discrète

Les premiers travaux réalisés dans le cadre de la géométrie discrète ont consisté en l'étude de la notion de courbure discrète. Nous avons alors développé des algorithmes efficaces de calcul de la courbure en chaque point d'une courbe discrète. Ces derniers ont été appliqués dans le cadre du projet « Stèles » notamment dans le but de décrire les profils. Nos activités s'orientent maintenant vers le calcul d'autres caractéristiques géométriques, telles que l'estimation de la longueur d'une courbe discrète, l'estimation de la surface d'un objet discret ou le calcul de la normale et de la courbure en chaque point d'un objet 3D discret. De tels paramètres pourront être utilisés dans le cadre de divers projets en cours, comme par exemple le projet « Mammo » pour caractériser la forme de la zone tumorale, dans le projet « Neige » pour estimer l'état d'un échantillon neigeux. Par ailleurs, cette « extraction des connaissances » à partir d'images 2D ou 3D servira dans le cadre de l'apprentissage par le contenu dans des bases de données d'images.

---

[60]S. MIGUET. Un Environnement Parallèle pour l'Imagerie 3D. *Mémoire d'Habilitation à Diriger des Recherches. Laboratoire de l'informatique du Parallélisme de l' Ecole Normale Supérieure de Lyon.* Décembre 1995.

2.3.2.3.2 Transformations géométriques

La mise en correspondance d'images implique la recherche de transformations géométriques maximisant la similarité entre une image de référence et la transformée d'une image observée. Nous nous sommes restreints dans un premier temps à l'étude de transformations affines (translations, rotations, changement d'échelles), et éventuellement de projection perspective. Dans certains cas, ces transformations peuvent être effectuées en plusieurs passes, amenant des calculs plus simples et un accès plus régulier aux données. En revanche, cette simplification peut amener des imprécisions numériques dans les résultats qu'il est intéressant de savoir maîtriser, dans la mesure où l'on se base sur le résultat de la transformation pour effectuer le recalage. Nous sommes en train d'élargir la classe des transformations jusque-là étudiées, et notamment considérer les transformations homéomorphiques (déformables élastiques).

2.3.2.4 Travaux relatifs à l'apprentissage

2.3.2.4.1 Stèles

Les travaux effectués sur les stèles permettent d'ores et déjà de construire une vectorisation des profils. Nous souhaitons faire collaborer les pôles Image et Apprentissage à l'étape suivante du projet, qui consiste à effectuer une classification de ces profils à l'aide des techniques étudiées au sein du pôle Apprentissage. Les outils les plus fréquemment utilisés au sein du laboratoire travaillent sur des vecteurs d'attributs au nombre fixé de composantes. Or le profil que nous construisons actuellement pour une stèle consiste en une chaîne a priori non bornée d'éléments géométriques. Deux axes de recherche semblent se dessiner : adapter les outils de classification développés au laboratoire pour leur permettre de traiter des chaînes de taille arbitraire ou au contraire, extraire de la vectorisation un nombre fixe d'attributs, choisis de manière pertinente pour permettre une discrimination efficace.

2.3.2.4.2 Mise en correspondance

Dans le cadre de la recherche d'une mesure de similarité entre images, nous avons entamé une collaboration avec des membres du pôle Apprentissage concernant d'une part l'évaluation des performances des tests statistiques et d'autre part, l'étude de nouvelles mesures de similarité.

2.3.3 Prospective – Pôle Bases de données décisionnelles

Les travaux de recherche du pôle Bases de Données Décisionnelles (BDD) visent à apporter des solutions dans l'élaboration de systèmes d'aide à la décision, notamment dans la conception d'entrepôts de données évolutifs bâtis à partir de sources de données multi-formes et multi-supports et dans la construction d'espaces d'analyse efficaces. Nous portons un intérêt particulier à la possibilité d'utiliser les bases de données non seulement comme une source de données pour le *data mining*, mais aussi comme un outil d'analyse en soi, que ce soit dans un contexte de bases de données traditionnelles ou d'entrepôts de données. Dans cette optique, nous avons prévu de développer les projets suivants.

2.3.3.1 Structuration de données multi-formes, multi-sources et multi-supports

Les bases de données à vocation décisionnelle (comme les entrepôts de données) peuvent nécessiter des sources de données externes. Par exemple, une entreprise qui souhaite faire de la veille concurrentielle ne pourra se contenter d'analyser uniquement des données provenant de ses bases de production. Une source de données primordiale dans un tel contexte est le web. Cependant, les données diffusées sur ce média sont très hétérogènes et le problème de leur intégration et de leur homogénéisation se pose avant de pouvoir envisager leur appliquer des méthodes d'ECD. Notre objectif est de parvenir à utiliser le web comme source de données à part entière de façon transparente. Ceci soulève des problèmes de recherche à différents niveaux.

- Structuration de données multiformes en provenance du web – texte libre, données multimédia (sons, images, vidéo...), données semi-structurées (HTML, XML...) – dans une base de données (travail de DEA ECD de S. Miniaoui).
- Intégration de ces données dans l'architecture particulière d'un entrepôt de données (datamarts, tables de faits, tables de dimension).
- Stratégies d'évolution de l'entrepôt face à des données nouvelles.

2.3.3.2 Apports du *data mining* dans le data warehousing

Dans un entrepôt de données, les données sont agrégées et représentées sous forme de magasins de données selon différentes dimensions représentant les axes d'analyse. L'analyste choisit lui-même les données à analyser en se fiant à son expérience et à sa connaissance du domaine. Nous proposons de recourir à la fouille de données afin d'aider l'analyste dans son choix des axes à

analyser et dans la sélection des données. L'utilisation de différentes techniques de *data mining* pourra aussi mettre en lumière des associations entre les différentes dimensions, par exemple, ou entre une valeur d'une mesure et une ou plusieurs dimensions. Les règles ainsi extraites peuvent expliquer le déroulement d'une activité ou prédire ses retombées. Nous proposons par ailleurs d'intégrer la connaissance ainsi obtenue dans l'entrepôt de données lui-même.

2.3.3.3 Auto-administration d'entrepôts de données à l'aide de techniques d'extraction de données

Avec le développement des bases de données en général et des entrepôts de données en particulier, il est devenu très important de réduire la fonction d'administration de base de données. Les systèmes auto-administratifs ont pour objectif de s'administrer et de s'adapter eux-mêmes, automatiquement, sans perte de performance[61]. Notre idée est d'utiliser des techniques de *data mining* pour extraire à partir des données stockées des connaissances utiles pour l'administration. C'est une approche très prometteuse, notamment pour les entrepôts de données, où les requêtes sont très hétérogènes et ne peuvent pas être interprétées facilement. Notre objectif est de rechercher une manière d'extraire dans les des données des connaissances utilisables pour appliquer des techniques d'auto-administration existantes (principalement des techniques d'optimisation des performances comme l'indexation, la gestion de cache, le préchargement et le groupement). Ce projet est mené en collaboration avec le laboratoire de Bases de données de l'université d'Oklahoma, USA (OUDB) dirigé par M^me^ Le Gruenwald.

2.3.3.4 Évaluation des performances de méthodes de *data mining*

De nombreux algorithmes sont proposés dans le domaine de la fouille de données. L'évaluation de leurs performances est cruciale pour leurs concepteurs, que ce soit pour comparer leurs travaux à d'autres ou pour appréhender «en situation» le comportement de leur algorithme, et pour leurs utilisateurs. Cette évaluation est généralement effectuée en appliquant les algorithmes étudiés à un ensemble de données synthétiques ou issues de cas réels. Ces différents « data sets » sont actuellement en nombre restreint et pratiquement tous fixes (ils ne représentent chacun

[61] CHAUDHURI S., NARASAYYA V. *An Efficient, Cost-Driven Index Selection Tool for Microsoft SQL Server*. 23rd Conference on Very Large Databases (VLDB '97), Athens, Greece, 1997

qu'un cas d'utilisation possible). Il nous paraît donc intéressant de proposer un protocole de génération de bases de données synthétiques paramétrable (à partir d'une base de règles comme des règles d'association, par exemple) permettant d'obtenir différentes configurations de grandes bases de données susceptibles d'être employées pour l'évaluation des performances d'algorithmes de fouille de données. Il s'agit à la fois de concevoir un schéma générique de base de données représentatif des besoins en *data mining* et d'assurer que la sémantique des donnés générées pour ce schéma est respectée (problème NP-complet).

2.3.3.5 Opérateurs Base de Données pour le Data *mining*

Dans le contexte de l'apprentissage supervisé (construction de graphes d'induction), nous proposons une nouvelle approche de *data mining* qui consiste à considérer les populations successives d'apprentissage comme des vues relationnelles. Le but de ce travail est d'intégrer des opérateurs de *data mining* au sein d'un SGBD de manière à : (1) pouvoir traiter des masses importantes de données et (2) bénéficier des techniques d'optimisation du SGBD utilisé afin de réduire le temps d'exécution.

ANNEXE : Sujets de thèses en cours

Céline AGIER

Date de naissance : 28/03/1975

Directeur de thèse : R. Rakotomalala

Financement : allocation du ministère

Date de début de thèse : 01/10/1999

Date de soutenance : fin 2002

Titre de la thèse : Simplification des bases de règles en ECD

Mots clés : Règles d'association, Méthodes de simplification, ECD

Résumé de thèse:

La problématique de l'ECD (Extraction de Connaissances à partir de Données) est maintenant bien connue : comment, à partir de bases de données de très grande taille, extraire des informations pertinentes, souvent sous forme d'ensemble de règles. La généralisation des très nombreuses techniques existantes de production de règles dans différents domaines, notamment industriels, pose néanmoins un problème qui réduit l'efficacité de l'extraction de règles : le nombre de règles produit peut être tellement grand que la lecture des bases de règles est impossible, empêchant toute interprétation cohérente de la connaissance produite dans son intégralité.

L'objectif de la thèse est de mettre en place des procédés de simplification de bases de règles en s'appuyant sur les techniques d'analyse de données. Une règle dans une base AGIERpouvant être considérée comme un tuple, en regroupant des règles «similaires «, on peut faire émerger des groupes qui correspondrait à des concepts plus généraux, en tous les cas plus élaborés, qu'il faudra ensuite traduire en connaissance interprétable (soit sous forme d'une règle «représentative», soit sous une autre forme).

L'étude se porte plus particulièrement sur les règles d'association. Plusieurs méthodes possibles permettant la simplification des bases de règles d'association sont : imposer une contrainte sur les items des règles produites, l'élagage lors de la production des règles pour avoir une «couverture» minimale des règles (Toivonen & al.), le clustering sur les règles d'association produites (Toivonen & al.).

Il existe aussi des techniques plus «classiques « de simplification de bases de règles, celles fondées sur une approche symbolique pure (algorithme type VLSI par exemple), ou encore celles fondées sur une optimisation numérique d'une mesure mettant en balance la complexité et la précision (algorithmes génétiques ou de recuit simulé pour la réduction des

bases de règles). De plus, représenter les règles d'associations sous forme d'objets symboliques (en utilisant notamment la notion de bornes et d'intervalle de confiance sur les mesures d'intérêt liées aux règles d'association que sont le support, la confiance, le lift, etc…) permettrait d'appliquer à un ensemble de règles d'association les méthodes de regroupement de l'Analyse de Données Symbolique (Diday).

Il pourrait être intéressant, une fois que l'on disposera d'une méthode réellement applicable pour les règles d'association, de voir si on peut l'adapter pour l'appliquer à d'autres types de règles produites par d'autres algorithmes d'apprentissage.

Laurent BAUMES

Date de naissance : 14/06/1977

Directeurs de thèse : Claude Mirodatos, Nicolas Nicoloyannis et Ferdi Schuth

Financement : CNRS et Max Planck Institute

Date de début de thèse : Janvier 2001

Date de soutenance : janvier 2004

Titre de la thèse : New methodological approach to generate a database in a sequential way by combining optimisation and learning methods : An application for efficient heterogeneous catalysts' combinatorial production.

Résumé de thèse:

Catalyse et approche combinatoire

Grâce au développement de robots de préparation et de techniques d'évaluation rapide des catalyseurs, il apparaît possible aujourd'hui d'accélérer la recherche et ainsi de trouver de bons catalyseurs dans un domaine de temps réduit. La méthode de recherche consiste à faire un screening de nombreux échantillons (première bibliothèque), à sélectionner les meilleurs candidats, puis à les modifier pour créer une seconde bibliothèque qui sera évaluée et ainsi de suite. Dans cette approche de recherche de catalyseurs, l'acte catalytique est considéré comme une boîte noire. Cette nouvelle méthodologie de recherche, qui est communément appelée Chimie Combinatoire, produit des données à haut débit. Nous nous proposons de gérer ces données à deux niveaux :

1. concevoir des bibliothèques pertinentes lors des cycles combinatoires
2. trouver des corrélations caractéristiques – performances sur de larges bases de données

<u>L'informatique</u>

La chimie combinatoire permet d'engendrer de larges banques de données incluant à la foi certains caractères physico-chimiques et les performances catalytiques. Nous proposons de trouver des corrélations entre les caractéristiques des matériaux et leurs performances. La problématique fait appel à diverses techniques informatiques. Cette étude comprend deux domaines : Optimisation et Modélisation

L'alimentation de la base de données ne peut se faire d'un seul trait en quantité suffisante : la production des données est en effet contrainte par les capacités des robots employés. Pour cette

raison, la base de données sera alimentée de façon séquentielle, c'est à dire que la base augmentera de façon régulière du nombre possible de tests que peuvent effectuer les machines. Il paraît alors naturel de faire croître la taille des échantillons tout en recherchant à optimiser les performances des catalyseurs. Ainsi sur le cycle d'alimentation de la base vient se greffer des méthodes informatiques d'optimisation. L'exercice d'optimisation retiendra notre attention dans un premier temps via des méthodes stochastiques, déterministes ou hybrides (Algorithmes génétiques (AG), programmation génétique (PG), méthodes du simplexe, du gradient et bien d'autres sont envisageables) selon le niveau d'adaptation de celles-ci au problème posé. Le caractère multi-critères de l'analyse est donné par la non-unicité du critère de performance. En effet, plusieurs niveaux d'appréciation d'un catalyseur existent. Certains seront de type discret et d'autres continus.

Cependant une réflexion plus globale semble le réel enjeu de cette étude ; à savoir la détermination de correspondances qui peuvent être formalisées sous forme de règles entre les caractéristiques d'un coté et les critères de performance de l'autre. La partition des résultats de performance en groupe de catalyseurs de niveaux de qualité ordonnés et la correspondance avec des groupes définis de caractéristiques qualitatives et quantitatives requiert des techniques de *Data mining*. Nous nous situons ainsi à la frontière entre la classification et l'extraction de règles.

Originalité

En premier lieu, le rapprochement des techniques informatiques et de la chimie catalytique dans un travail de fond, donne à ce sujet une dimension originale non négligeable. La segmentation du catalyseur en plusieurs morceaux sert à optimiser pas à pas car le nombre de variables endogènes est grand relativement aux nombre de variables exogènes ou exemples. La mise en place d'un algorithme d'optimisation et de techniques d'apprentissage simultanés n'est pas commun et rend l'exercice tout à fait novateur. Les tests d'intégration possible suite aux tests de validité des résultats issus des méthodes d'apprentissage semble une voie intéressante de recherche théorique dans le cas d'une base de données qui augmente séquentiellement.

Programme et déroulement envisagé

Premièrement : production d'un algorithme d'optimisation qui permettrait de fournir les valeurs des paramètres de catalyseurs qui soient potentiellement meilleurs que les précédents.

Dans un second temps : faire une correspondance entre la qualité du catalyseur et ses caractéristiques. Ainsi, disposant d'un début de base de données, il sera testé chez MPI (Max

Planck Insitute) et IRC (Institut de Recherches en Catalyse) les techniques d'apprentissage (classification, règles, arbres d'induction, réseaux de neurones …). Au préalable, il sera nécessaire de mener à bien une réflexion sur les méthodes d'agrégation des critères de performance.

Une extension de ce travail sera d'effectuer une analyse sur un fichier contenant des informations sur plusieurs réactions, le but étant de construire une correspondance entre les caractéristiques d'une réaction complète et les caractéristiques des catalyseurs potentiels. L'intérêt est de partir sur une nouvelle réaction avec des informations dès le départ.

Jérémy CLECH

Date de naissance : 16 avril 1976

Directeur de thèse : Djamel A. ZIGHED

Financement : salarié, société Diagnos Dev

Date de début : 1er septembre 2000

Date de soutenance prévu : décembre 2003

Titre de la thèse : Recherche dans les corpus non structurés

Mots clefs : Systèmes d'Informations, Moteurs de Recherche, Classification, Text Mining, Knowledge Management.

Résumé de thèse :

Le volume des bases de données double chaque année. La plupart de ces bases se trouvent au sein de réseaux comme les intranet, extranet ou encore, le réseau des réseaux, internet. A leur échelle, ces bases sont la première source d'informations de leurs groupes d'utilisateurs. Pour faire face à la densité de documents en lignes, ces réseaux sont équipés de moteurs de recherche. Ainsi, l'intérêt de la consultation de ces médias est directement lié à l'efficacité de ces moteurs. Les principaux d'entre eux fonctionnent essentiellement sur la base de mots clés. Les limites de cette technique d'interrogation sont très facile à voir : des centaines de pages d'inégal intérêt sont proposées à chaque requête. En outre, pour ne citer que l'exemple du Web, une grande part des ressources du réseaux est contenue dans des bases de données dynamiques qui construisent des pages éphémères, et ne sont pas explorées / indexées par les moteurs de recherche classique. Enfin, une étude américaine de 1999 montre la non adéquation des moteurs et des habitudes des utilisateurs en raison de la complexité d'écriture de requêtes précises (utilisation d'opérateurs booléens, guillemets, parenthèses, …). Dans cette thèse, il s'agit de réfléchir sur le moyen de mise en œuvre des méthodes d'apprentissage automatique pour réaliser des moteurs de recherche plus intelligents. Pour ce faire, cette recherche pourra débuter sur l'analyse des techniques de représentations de texte (e.g. n-grammes, résumé, syntaxique, sémantique, …) afin de définir leurs qualités face aux problèmes de classifications et de recherche. Ensuite, il s'agit de définir une architecture de moteur de recherche apprenant de façon incrémentale, à partir d'exemples (corpus de textes traitant du sujet souhaité) ou de mots clés, et offrant la possibilité de chercher des informations dans des bases de données dynamiques. Ces travaux devront conduire à la réalisation d'un prototype qui devra être testé dans des cas

concrets comme la recherche d'informations dans de grandes bases de données médicales. Enfin, cette recherche devra s'orienter dans le cadre de la problématique du partage et d'appropriation du cycle de Knowledge Management.

Publications :

J. Clech, «Développement d'un système informatique pour la construction de graphes d'induction afin de prédire des paramètres de structures complexes», rapport de DEA, Laboratoire ERIC, 6 juillet 2000.

[F32] A. Ciampi, D. A. Zighed, J. Clech, «Trees and Induction Graphs for Multivariate Response», PKDD 2000, septembre 2000.

David COEURJOLLY

Date de naissance : 03/09/1977

Directeur de thèse : Serge Miguet et Laure Tougne

Financement : MENRT

Date de début de thèse : 1 Octobre 2000

Date de soutenance prévue : année 2002-2003

Titre : Géométrie discrète et outils d'indexation d'images

Mots clés : Traitement d'Images, Extraction de connaissances à partir des données, Imagerie médicale, Géométrie discrète.

Résumé de la thèse

Un des principaux axes développés par le pôle Images du Laboratoire ERIC concerne la recherche par le contenu dans des bases de données multimédia 2D ou 3D. En effet, il est parfois impossible d'indexer a priori les images et par suite, de faire des requêtes «classiques». Un exemple concerne le domaine médical dans lequel un tel processus peut fournir une aide au positionnement de patients ou bien, dans lequel un médecin peut vouloir rechercher dans une base de données d'images toutes celles comportant une tumeur «visuellement similaire».

Une première approche développée jusque-là est basée sur des méthodes de recalage multimodal consistant à mettre en correspondance un couple d'images de façon à pouvoir fusionner les informations de chacune d'elles. Une évaluation précise du placement du patient est ainsi obtenue grâce à une recherche par le contenu dans une série d'images pré-calculées.

Une deuxième approche complémentaire que le Laboratoire souhaiterait développer concerne l'extraction de caractéristiques géométriques discrètes 3D communes aux deux images. Il s'agit classiquement de points, de courbes, de segments ou de facettes. Des techniques plus évoluées associent d'autres attributs, souvent de nature différentielle, aux coordonnées des primitives. Un exemple concerne les lignes de crête (ou crest line) définies comme une suite de points dont la courbure est un extrémum local. Les intersections entre ces lignes de crête pouvant également servir de repères.

Dans le cadre de cette thèse, il s'agit de mettre en œuvre des méthodes de géométrie discrète permettant de calculer la courbure en tout point d'une surface discrète. En préparation à ce travail et dans le cadre du stage de DEA, une étude bibliographique des différents algorithmes de calculs de courbure, discrets ou non, 2D ou 3D a permis de mettre en évidence la nécessité d'un algorithme efficace purement discret. Un premier travail concerne une version purement discrète du calcul de courbure en tout courbe 2D.

Publications :

[G3] COEURJOLLY D., SARRUT D., AND TOUGNE L. *Décimation en imagerie médicale 3d*. In Courbes Surfaces et Algorithmes, Journées du Groupe de Travail «Modélisation Géométrique», Grenoble, France, September 99, **(1999)**

[F36] COEURJOLLY D., MIGUET S. ET TOUGNE L. *Discrete Curvature based on Osculating Circle Estimation*. 4th international workshop on visual form (IWVF4'2001) **(2001)**

[F47] COEURJOLLY D., GERARD Y., REVEILLES J.P. ET TOUGNE L. *An elementary algorithm for digital arc segmentation*. International Workshop on Combinatorial Image Analysis (IWCIA'2001) **(2001)**

Sébastien CLIPPE

Date de naissance : 30/09/1970

Directeur de thèse : Serge Miguet

Financement : Hospices Civils de Lyon puis Centre Léon Bérard

Date de début de thèse : 1 Octobre 2000

Date de soutenance prévue : année 2002-2003

Titre : Aide au positionnement du patient en radiothérapie par des techniques d'imagerie.

Mots clés : radiothérapie conformationnelle, positionnement de patient, recalage d'images, mesures de similarité

Résumé de la thèse :

Contexte : En radiothérapie conformationnelle, le but est de délivrer une dose maximale à la tumeur en épargnant les tissus sains environnants. Cependant le principal facteur limitant est la reproductibilité quotidienne du positionnement du patient. Une solution passe par la réalisation d'images de contrôle et leur comparaison avec une image de référence afin de corriger éventuellement cette mauvaise installation. Ceci est généralement effectué manuellement par un médecin, ce qui est imprécis et consommateur de temps. Des travaux préliminaires [cars2000], dont une partie dans le cadre de mon DEA, ont débuté afin de développer une méthode automatisée de quantification des erreurs de positionnement et nous nous proposons de les poursuivre.

Une collaboration a débuté depuis quatre ans entre le laboratoire ERIC de l'université Lumière Lyon 2 et le «Centre Régional de Lutte contre le Cancer» Léon Bérard autour du thème de l'aide au positionnement du patient en radiothérapie. Ces travaux de thèse s'intègrent également dans un projet régional nommé ADEMO (Acquisition et Décision conduites par le Modèle) concernant, entre autres, le positionnement du patient.

Objectifs : Dans un premier temps, nous proposons la poursuite du développement d'une méthode d'aide au positionnement utilisant des techniques de recalage d'images 2D/3D basées sur l'intensité des pixels. Pour cela nous prévoyons de porter nos efforts sur l'amélioration des processus d'optimisation, actuellement sources d'un certain nombre d'échecs. Nous travaillerons également sur les différentes mesures de similarité afin de déterminer la plus efficace. Sur le plan expérimental, nous allons acquérir un fantôme

adapté à nos mesures dans des conditions réelles. Le Centre Léon Bérard vient par ailleurs d'acquérir un nouvel appareil d'acquisition d'images portales (Iview, Elekta) permettant de disposer d'images de contrôle de meilleure qualité. Dans un second temps nous passerons à une validation «pré-clinique» de la méthode (fiabilité et précision) par des tests réels sur un grand nombre de positions du fantôme. La troisième étape sera, selon les résultats, une étude pilote sur quelques patients avant de proposer un prototype totalement automatisé d'aide au positionnement.

Parallèlement, nous travaillerons sur la prise en compte du déplacement des organes relativement aux structures osseuses. D'autre part, une autre voie d'aide au positionnement du patient réside en l'utilisation de marqueurs radio-opaques et nous nous proposons d'explorer cette méthode. Nous prévoyons tout d'abord de travailler sur la détection des marqueurs dans les images portales (sur fantôme, puis sur patient) avant d' étudier leur mobilité potentielle puis d'entreprendre une étude pilote sur quelques patients. Le travail d'extraction de caractéristiques dans des images sera également mis à profit dans un projet de CAD (Computer Assisted Diagnosis) concernant la mammographie.

Publications :

[F31] SARRUT D., CLIPPE S. *Patient positioning in radiotherapy by REGISTRATION of 2D portal to 3D CT images by a contend-based research with similarity measures*. In Computer Assisted Radiology and Surgery, San Francisco, USA, pages 707-712. Elsevier Science, Amsterdam, NL **(2000)**.

[F40] CLIPPE S., SARRUT D. *Patient positioning in radiotherapy*. In 92nd Annual Meeting - American Association for Cancer Research, à paraître **(2001)**.

[G17] SARRUT D., CLIPPE C. *Positionnement de patient en radiothérapie par recalage 2d/3d sans segmentation*. In Radiothérapie Assistée par l'Image (RAI 2001), Centre Antoine-Lacassagne, Nice, Mai 01 **(2001)**.

Serge DI PALMA

Date de naissance : 22.06.1970

Directeur de thèse : D. A. Zighed

Financement : salarié, société Diagnos Dev

Date de début de thèse : septembre 1997

Date de soutenance prévue : Fin 2001

Titre : Caractérisation, Recherche et Evaluation d'espaces de représentation sur des données hétérogènes pour des applications dans le domaine de la découverte automatique de connaissances. Application aux données textuelles.

Mots clés : Apprentissage supervisé à partir de textes, sélection et construction d 'attributs.

Résumé de la thèse :

Nous nous sommes intéressés dans notre thèse aux techniques de représentation des textes pour les problèmes d'apprentissage supervisés, privilégiant celles qui s'imposent de construire les descripteurs à partir du seul contenu typographique des documents à classer. Nous avons plus particulièrement étudié la technique dite des « n-gram » consistant à utiliser comme descripteurs potentiels l'ensemble des séquences de n signes à l'intérieur des textes constituant l'échantillon d'apprentissage.

La dimensionnalité (nombre de variables) très grande de ces problèmes nous a naturellement conduit à étudier les différentes techniques de sélection d'attributs concrètement utilisées en catégorisation de textes. Nous avons proposé une réinterprétation de plusieurs critères de sélection, associée à une méthodologie plus générale de définition d'un critère de sélection d'attributs prenant en compte les spécificités du problème d'apprentissage et de l'algorithme de catégorisation de textes que l'utilisateur souhaite utiliser.

Nous travaillons actuellement à la caractérisation des conditions d'équivalence (du point de vue de l'ordre induit sur les attributs) des différents critères de sélection utilisés en catégorisation des textes. Cette caractérisation permettrait de fournir une explication théorique au constat empirique d'équivalence des différents critères que de nombreuses études ont diagnostiqué suite à des tests sur des bases de textes exemples.

Parallèlement à ces techniques de sélection, nous avons également étudié l'application au domaine du texte des techniques de construction d'attributs fondées sur la recherche de

combinaisons (le plus souvent linéaires) des attributs initiaux. Cette étude a abouti à la définition d'un cadre cohérent permettant de situer les différentes pratiques dans ce domaine (essentiellement occupé par les techniques dites de « Latent Semantic Indexing »). Elle a également permis la mise au point d'une technique originale de construction d'attributs pour l'apprentissage supervisé à partir de textes dans une optique d'indexation. Cette technique repose sur l'utilisation de méthodes traditionnelles d'analyse conjointe de plusieurs tableaux (analyse canonique ou analyse de co-inertie) appliquées simultanément à une matrice de fréquences de termes et à une matrice booléenne d'indexation.

Afin de valoriser les différentes études que nous venons de décrire, nous avons intégré à la plate-forme Sipina©, développée au sein du laboratoire ERIC, un module de classification automatique de textes permettant de :

- Définir de manière interactive un échantillon d'apprentissage constitué de textes bruts étiquetés par rapport à une liste de catégories prédéfinie.
- Construire, pour un problème donné, un fichier de description d'un ensemble de textes d'apprentissage en sélectionnant automatiquement les attributs (mots, n-grammes, racines) les plus discriminants. Différents critères de sélection des attributs peuvent être testés. L'extraction des racines des mots (stemming) ne concerne pour l'instant que les textes anglais.
- Appliquer sur ce fichier toutes les méthodes d'apprentissage implémentées dans la plate-forme Sipina© (graphes d'induction, réseaux de neurones, analyse discriminante…) et valider les connaissances produites à l'aide de différentes procédures de réechantillonnage (bootstrap, validation croisée …)
- Utiliser les connaissances validées pour classer et indexer automatiquement de nouveaux textes, sans codage préalable.

Publications

S. Di Palma, J. Da Rugna et D.A. Zighed. Apport d'approches basées sur la décomposition de relations binaires en apprentissage supervisé. Note de recherche. Laboratoire ERIC, Lyon2, 1997.

S. Di Palma, J. Da Rugna et D.A. Zighed. Apprentissage supervisé polythétique : une évaluation. Actes des 5eme rencontres de la Société Francophone de Classification. 1997

[F5] G. Gavin, S. Di Palma and D.A. Zighed. Generalisation properties of convex linear combinations of Classifiers belonging to finite VC-dimension family. Fourth International Conference on Computer Science and Informatics, octobre 1998.

M. Sebban, D.A. Zighed, S. Di Palma: Selection and Statistical Validation of Features and Prototypes, Proceedings of the 3rd European Symposium «Principles of *Data mining* and Knowledge Discovery» PKDD'99, Prague (Lecture Notes in Artificial Intelligence 1704), Springer Verlag 1999, 184-192

[F14] R. Rakotomalala, S. Lallich, S. Di Palma: Studying the Behavior of Generalized Entropy in Induction Trees Using a M-of-N Concept. Proceedings of the 3rd European Symposium «Principles of *Data mining* and Knowledge Discovery» PKDD'99, Prague (Lecture Notes in Artificial Intelligence 1704), Springer Verlag 1999, 510-517

Djamel Zighed and Serge Di Palma , «A tutorial on Machin-Learning Text Categorization". 4th European Conference on Knoweldge Discovery in DataBases, PKDD'2000, September 2000.

Radwan JALAM

Date de naissance : 20/05/1968

Directeur de thèse : J.-H. Chauchat

Financement : ATER

Date de début de thèse : Décembre 1998

Date de soutenance : début 2002

Titre de la thèse : Catégorisation de textes multilingues

Mots clés : Apprentissage, Text-Mining, Cross-languages Information Access.

Résumé de la thèse

L'objectif de mes travaux porte sur la conception et la réalisation d'un agent informatique intelligent qui apprend à sélectionner les documents pertinents parmi les pages WEB multilingues selon les thèmes qui intéressent l'utilisateur. Le jugement porté par l'utilisateur sur la pertinence des documents est utilisé par l'agent en auto apprentissage pour affiner le profil de ce qui est recherché. Le flot de documents est constitué de pages WEB multilingues. A l'autre extrémité, une interface permet à un utilisateur de définir un profil pour chaque langue sur la base de différents paramètres (requêtes, mots clés, documents pertinents, date...). Nous utilisons un ensemble d'apprentissage de textes pour générer automatiquement une liste d'attributs discriminants qui peuvent être des mots clefs, des séquences de lettres (n-grammes), des séquences de mots. Par exemple, on filtre les trigrammes pertinents en ne retenant que ceux bien corrélés à l'étiquette, et parmi ceux-ci une sélection par apprentissage permet de se limiter aux plus efficaces. Ensuite, on prend en compte les phrases ou les mots clefs déclarés par l'utilisateur que l'on transforme en attributs binaires voire plus complexes. Si un attribut est multivalué, tel que auteur, date, ..., on passe en forme disjonctive complète. Le profil final de l'utilisateur pour une langue i rassemble ces différents attributs.

Dans un second temps on revient aux textes de l'ensemble d'apprentissage en utilisant l'ensemble des attributs. Par l'une ou l'autre des méthodes usuelles, on construit une fonction discriminante ou un ensemble de règles. Les règles facilitent l'intégration de connaissances du domaine, telle que la date, l'institution ...

Après une vaste recherche bibliographique concernant les algorithmes permettant l'identification automatique de langues (O. Teytaud and R. Jalam, Identification de la

langue basée sur les N-grammes, Rapport de recherche ERIC, Juillet 2000) nous avons proposé un nouvel algorithme et conçu un devineur de langue capable d'identifier plus de 69 langues à partir de textes de taille de 50 octets avec un taux de réussite supérieur à 96% (O. Teytaud and R. Jalam, Kernel-based text-categorization, ERIC, Juillet 2000).

Publications :

[G20] Jalam R., Teytaud O. Identification de la langue et catégorisation de textes basées sur les N-grammes, ECA, Vol. 1, n°1-2/2001, pages 227-238 (2001).

Pierre-Emmanuel JOUVE

Date de naissance : 16/12/1976

Directeur de thèse : Nicolas Nicoloyannis

Financement : Bourse Région Rhône-Alpes

Date de début de thèse : 1 Octobre 2000

Date de soutenance prévue : année 2002-2003

Titre de la thèse : Apprentissage et Structuration de données

Mots-clés : Structuration de données, Apprentissage, modélisation géométrique, modélisation prétopologique, modélisation statistique

Résumé de la thèse

Les techniques d'extraction de connaissances à partir de grandes bases de données multi-sources sont souvent utilisées dans les processus d'aide à la décision. En matière d'extraction de connaissances et plus particulièrement dans le domaine d'apprentissage automatique nous sommes très souvent confrontés au problème de la représentation et de la structuration de l'information à traiter. La qualité de l'apprentissage étant fortement liée à l'espace de représentation des données, l'objectif de cette thèse est l'étude de l'impact des problèmes de structuration des données sur l'apprentissage. Il s'agit de mettre au point des outils méthodologiques capables de mesurer l'apprenabilité d'une structure retenue pour représenter les données.

Ces outils nous permettront de définir la structure optimale parmi l'ensemble des structures possibles et ceci indépendamment de toute méthode d'apprentissage. Il s'agit donc de concevoir des outils méthodologiques capables de caractériser la structure des données à utiliser pour effectuer le meilleur apprentissage et ceci indépendamment de la méthode utilisée.

L'approche méthodologique pour mettre en oeuvre les objectifs de ce travail fera appel à la modélisation géométrique, prétopologique et statistique.

Approche géométrique : Dans le cadre méthodologique général qui dépasse la simple construction d'une fonction de classement, nous sommes intéressés par la construction de

tests de séparabilité des classes. Parmi les modèles possibles on peut citer les graphes de Delaunay, les graphes de Gabriel, les graphes de voisins relatifs etc.
Approche prétopologique : Cette approche nous permettra de structurer les données d'une façon optimale, dans le cas où la seule information disponible sur les données n'est pas une information quantitative mais qui s'exprime sous une forme qualitative donnée par une relation binaire.

Approche statistique : Par cette approche on envisage de mettre en évidence les agrégations nécessaires qu'on doit effectuer sur les données initiales pour optimiser la représentation finale des données qui vont être soumis en apprentissage.

Une extension de la problématique dans le cas des espaces non métriques est prévue. L'approche théorique sera complétée par la réalisation d'un outil logiciel qui va enrichir la plate-forme logiciel déjà existante au sein du laboratoire.

Fabrice MUHLENBACH

Date de naissance : 16/12/1972

Directeur de thèse : Djamel A. Zighed

Financement : MENRT

Date de début de thèse : Octobre 1999

Date de soutenance prévue : Décembre 2002

Titre de la thèse : Intégration, fusion et fouille de données non structurées

Mots clés : Fusion de données, Fouille de données, ECD, Sciences cognitives

Résumé de la thèse :

La fusion de données est l'intégration globale de données issues de sources disparates. À partir d'informations de formes diverses et séparées, les processus de fusion de données produisent un objet composite gardant l'intégrité des données. La fusion de données est un domaine récent, pluridisciplinaire et en plein essor.

Les recherches théoriques sur ce domaine font appel aux techniques de *data mining*, l'imagerie, la reconnaissance de forme et l'apprentissage à partir de données numériques et symboliques. Le domaine de la fusion de données est riche en applications : navigation et surveillance (à travers la gestion d'informations issues de différents capteurs), veille technologique ou analyse d'informations (commerciales, météorologiques, etc.) sur Internet. Toutefois ce domaine aux applications multiples souffre encore d'une insuffisance d'ordre méthodologique globale.

Au cours de cette thèse, nous cherchons à développer des méthodes et des outils informatiques permettant de fusionner des données de types multiples (bases de données structurées, images ou textes libres) et d'en extraire automatiquement des connaissances. Nous avons ainsi conçu un logiciel intitulé « Data Squeezer » qui fusionne des catégories de données en cherchant à en minimiser l'incertitude générale afin d'en retirer des règles, à la manière d'un presse-citron où la pulpe du fruit – l'ensemble des données – est pressée pour en extraire le jus – la connaissance exprimée sous forme de règles d'inférence –.

Ce travail puise ses modèles auprès des sciences cognitives. En particulier, nous cherchons à prolonger les réflexions déjà commencées dans le laboratoire ERIC sur la perception multimodale, le système perceptif procédant en effet par la fusion des informations issues des différentes modalités sensorielles.

Publication :

[E3] MUHLENBACH F., ZIGHED D.A. et D'HONDT S. « Génération de règles par compression », *ECA*, *Vol. 1*, n°1-2/2001, p93-104 **(2001)**.

Vasile-Marian SCUTURICI

Date de naissance : 30 octobre 1971

Directeur de thèse : Serge Miguet

Financement : bourse CNOUS - financement Compagnie des Signaux

Date de début de thèse : mars 1998

Date de soutenance : novembre 2001

Titre de la thèse : Serveur Web haute performance et application à la

Mots clé : multimédia, vidéo à la demande (VoD), protocole HTTP, serveur vidéo, benchmarking pour VoD

Résumé de thèse:

Avec les avancées dans le domaine des réseaux d'ordinateurs, de nouveaux services interactifs sont développés sur Internet ou Intranet, parmi lesquels la vidéo à la demande (VoD). Actuellement, le nombre de solutions VoD croît de manière continue, ces systèmes deviennent de plus en plus complexes et coûteux, sans pour autant répondre totalement aux besoins des utilisateurs.

Un certain nombre d'utilisateurs ne peuvent accéder aux serveurs VoD que par l'intermédiaire du protocole HTTP (le protocole standard sur le Web), en raison des firewalls (pare-feux) mis en place dans les réseaux pour filtrer tous les autres paquets de données. L'installation d'un système VoD nécessiterait la modification des architectures réseau actuelles, en ajoutant des serveurs et des protocoles réseau supplémentaires. Hormis leur coût élevé, les serveurs VoD nécessitent souvent des plate-formes matérielles et logicielles propriétaires, dont le système de stockage est lié au serveur. Cependant, les serveurs Web sont présents dans presque tous les réseaux, fonctionnent sur toutes les plate-formes avec des systèmes de fichiers standards et utilisant HTTP. Par conséquent, l'utilisation d'un serveur HTTP comme serveur VoD présenterait de nombreux avantages. Mais l'utilisation d'un système VoD s'attend à disposer de fonctionnalités similaires à celles qu'il trouve sur un système de vidéo traditionnelle (VCR), c'est à dire les fonctions usuelles d'un magnétoscope (avance et retour rapide, arrêt sur une image, ralentir) ou d'un vidéo disque (incluant le saut à une plage aléatoire). Malheureusement, pour l'instant, les serveurs VoD utilisant HTTP souffrent d'un manque d'interactivité en terme de fonctionnalités VCR (notamment «saut» et «pause»), en raison principalement de stratégies d'utilisation de HTTP inadéquates.

Dans cette thèse nous avons analysé et mis en place *une stratégie d'utilisation de HTTP* qui permet l'utilisation efficiente de ce protocole pour la VoD. Cette idée permet le développement d'une solution simple et efficace d'un système VoD, articulé autour de serveurs Web existants, utilisés comme serveurs VoD.

Afin de mesurer les performances d'un système VoD (et du notre en particulier), nous avons développé une méthodologie de tests (*benchmarking*) pour un serveur VoD. Nous avons défini la *qualité de perception*, une mesure qui reflète le degré de satisfaction d'un utilisateur qui regarde un flux vidéo en utilisant un système VoD. À partir de cette mesure, nous avons déterminé des facteurs équivalents au niveau serveur. En mesurant ces facteurs, nous proposons une stratégie de benchmarking pour les serveurs VoD qui permet de trouver les performances réelles des différents serveurs VoD en termes d'utilisateurs satisfaits.

De plus, pour optimiser les performances d'un serveur Web utilisant notre stratégie en VoD tout en garantissant à l'utilisateur une certaine qualité de perception, nous avons mis en place une *politique de contrôle d'admission* et différentes *politiques de travail avec le système de stockage*. Ceci nous permet d'atteindre des performances proches des limites du système matériel VoD. La conception d'un serveur Web qui implémente ces politiques est aussi présenté dans le cadre de cette thèse.

Publications :

[F49] Vasile-Marian Scuturici, Mihaela Scuturici, Serge Miguet, Jean-Marie Pinon, «Measuring Web Servers Performance in VoD», International Conference on Telecommunications (ICT2001), Romania, Bucharest, June 4-7, 2001

[F29] Vasile-Marian SCUTURICI, Mihaela SCUTURICI, «Video on Demand Using HTTP», PROtocols for Multimedia Systems - PROMS2000, Cracow, Poland, October 22-25, 2000, pages 228-235, (http://proms2000.kt.agh.edu.pl/index.html)

[F20] Serge Miguet, Vasile-Marian Scuturici , «Using the Hyper Text Transfer Protocol for Multimedia Streaming», International Symposium on Intelligent Multimedia and Distance Education, Baden-Baden, Germany, 2-7 august 1999, pages 215-223, (http://www.ece.ndsu.nodak.edu/ece/research/conferences/isimade/)

Dorel Bozga, Dan Chiorean, Vasile-Marian Scuturici, Dan Suciu, Dan Vasilescu, «Present and Perspectives Of the CASE Tools Used In Object-Oriented Analysis & Design -The RO_CASE Experience» 3rd International Symposium in Economics Informatics» pag. 21-28., Bucharest, 1997

Vasile-Marian Scuturici, L'utilisation de serveurs Web pour la VoD, COmpression et REprésentation des Signaux Audiovisuels, CORESA'99, Sophia-Antipolis, juin 1999, pages 203-209

Serge Miguet, Vasile-Marian Scuturici, «Metrics for Measuring the Web Servers Performance», Studia Universitatis Babes-Bolyai, Series Computer Science, Cluj-Napoca, Romania, 1998

Vasile-Marian Scuturici, Dan Suciu, Mihaela Scuturici, Iulian Ober, «The Active Objects Description Using Statecharts» Studia Universitatis Babes-Bolyai, Cluj-Napoca, 1997

Michael Papathomas, Vasile-Marian Scuturici, «An Active Object Model for Multimedia Presentations», Studia Universitatis «Babes-Bolyai», Informatica, Volume I, Number 1 (1996), Pages 21 - 40

Olivier TEYTAUD

Date de naissance : 05/03/1975

Directeur de thèse : Hélène Paugam-Moisy

Financement : Allocation Couplée

Date de début de thèse : 15/09/99

Date de soutenance : année 2001/2002

Titre de la thèse : Représentation interne dans les systèmes d'apprentissage

Mots clés: Apprentissage, Text-Mining, Image-Querying, Représentation interne, Contrôle, Bornes, Support Vector Machines, Bayes Point Machines, Classes de Donsker, Apprentissage Bayesien.

Résumé de la thèse:

Cette thèse porte sur la théorie de l'apprentissage et notamment sur les Support Vector Machines, système d'apprentissage qui se développe abondamment ces dernières années et destiné à fournir un bon contrôle sur l'erreur en généralisation, et sur des variantes de ces machines (Bayes Point Machines notamment).

Elle devrait comporter:

- Une étude théorique du comportement asymptotique des Support Vector Machines (erreur en généralisation, nombre de support vectors, comparaison des bornes de sparsity et de marges), réalisée en collaboration avec J. Droniou.
- Une étude des bornes sur l'erreur en généralisation des algorithmes bayésiens, utilisant la fat-shattering dimension et les nombres de couverture.
- Des applications en text mining et classification d'image, toutes deux basées sur un noyau permettant de classifier des distributions par Support Vector Machine (ou autres méthodes à noyaux) de manière à avoir un feature space dans lequel la distance est naturelle pour des distributions. Le cas du text-mining a été étudié en collaboration avec R. Jalam, et un travail est en cours sur de la classification d'images, en collaboration avec D. Sarrut. Ces études insistent sur l'importance de la représentation utilisée par rapport à l'algorithme et au paradigme d'apprentissage.
- Une application en prédiction de systèmes chaotiques (en collaboration avec J.-M. Friedt, D., D. Gillet, M. Planat), qui devrait se prolonger en diverses études de turbulence (avec A. Moreau). Le lien avec la taille de la représentation interne est rendu visible par la dimension de l'attracteur et ses implications.

- Des essais d'application à la détection de pollution (travail réalisé avec ECRIN)
- Un contrat Elf cherchant des critères permettant de choisir entre différents systèmes d'apprentissage (ciblant notamment les cas de bons et mauvais fonctionnement des support vector machines).

Un contrat Elf supplémentaire devrait bientôt être signé, portant notamment sur la sélection de bases d'apprentissage réduites et l'apprentissage en temps réel, dans le cas des Support Vector Machines notamment.

Apprentissage dans le cadre d'une VC-dimension infinie

Publications :

J. Droniou, A. Elisseeff, H. Paugam-Moisy and O. Teytaud. *Contrôle de l'architecture et des représentations internes dans les réseaux de neurones multicouches*. In Actes de la Conférence sur l'Apprentissage, CAP'99, Palaiseau, FRANCE, pages 185--194 **(1999)**.

O. Teytaud and J. Droniou. *Comportement asymptotique des Support Vector Machines*. Rapport de Recherche ERIC 99-07, Université Lyon 2, juin **(1999)**.

O. Teytaud. *Bayesian learning / Structural risk minimization*. Rapport de Recherche ERIC 00-05, Université Lyon 2, juin **(2000).**

O. Teytaud. *Bounds on covering numbers and fat-shattering dimension for Bayesian learning*. Rapport de Recherche ERIC 00-06, Université Lyon 2, juillet **(2000)**.

R. Jalam and O. Teytaud. *Kernel-based text categorization*. Rapport de Recherche ERIC 00-07, Université Lyon 2, juillet **(2000)**.

J.-M. Friedt, O. Teytaud, D. Gillet and M. Planat. *Simultaneous amplitude and frequency noise analysis in Chua's circuit - neural network based prediction and analysis*. In Proc. of VIII Van Der Ziel Symposium on Quantum 1/f Noise and Other Low Frequency Fluctuations in Electronic Devices, St Louis, **(2000)**.

O. Teytaud, J.-M. Friedt and M. Planat. *Neural network based prediction of Chua's chaotic circuit frequency - possible extension to quartz oscillators*. In Proceeding of the European Time And Frequency Forum, Neuchâtel **(2001).**

J.-M. Friedt, O. Teytaud and M. Planat. *Learning from Noise in Chua's Oscillator. Accepted* . In the 16th International Conference on Noise in Physical Systems and 1/f Fluctuations, **(2001)**.

Jalam R. , Teytaud O. *Identification de la langue et catégorisation de textes basées sur les N-grammes*, ECA, Vol. 1, n°1-2/2001, pages 227-238 **(2001)**.

O. Teytaud. *Decidability of the halting problem for Matiyasevich Deterministic Machines*. Accepted for the special issue of TCS on WAD'14 **(2001)**.

O. Teytaud. *Convergence speed of deformable models*. Accepted for IJCNN **(2001)**.

O. Teytaud, R. Jalam. Kernel based text categorization. Accepted for IJCNN **(2001)**.

O. Teytaud. Robust Learning: Regression Noise. Accepted for IJCNN **(2001)**.

G. Gavin, O. Teytaud. Equivalence in the worst case between the training error estimate and the Leave-One-Out estimate. Accepted in IJCNN **(2001)**.

Tiffany TWEED

Date de naissance : 30/11/1975

Directeur de thèse : Serge Miguet

Financement : salariée, société Diagnos Dev

Date de début de thèse : 15/09/2000

Date de soutenance : année 2002/2003

Titre de la thèse : Aide au diagnostic du cancer du sein par extraction de connaissances dans les bases de données d'images mammographiques

Mots clés : Système d'aide au diagnostique(CAD), mammographies, détection de masses et de micro-calcifications

Résumé de thèse:

Contexte de l'étude

Le laboratoire ERIC de l'Université Lumière Lyon 2 débute une collaboration avec la société canadienne Diagnos, afin de développer une plate-forme d'aide à la décision capable de traiter des données de natures très diverses (textes, bases de données, images, etc). La société Diagnos souhaite en particulier utiliser cette plate-forme pour concevoir un système d'aide au diagnostic du cancer du sein, à partir de grandes bases de données d'images mammographiques.

Description des travaux

L'une des premières étapes des travaux qui seront effectués dans cette thèse sera la recherche des familles de paramètres pertinents pour discriminer les cas cancéreux ou douteux des cas sains. Une étude de la littérature montre qu'il est nécessaire de se focaliser sur la distribution, la taille et la forme des microcalcifications et des opacités observables sur la plupart des clichés. Nous prévoyons donc d'étudier d'une part des techniques géométriques locales de segmentation et de description de formes, d'autre part des outils statistiques basés sur des caractéristiques plus globales comme les distributions de niveaux de gris ou l'extraction de paramètres de texture.

Retombées attendues

L'un des objectifs attendus de ce projet de recherches est la conception d'outils automatisés pour l'aide au diagnostic du cancer du sein dont la caractéristique essentielle devra être de

ne jamais déclarer bénin un cas qui aurait été diagnostiqué cancéreux ou douteux par un expert (faux négatif).

Dans un cadre plus général, nous voulons proposer un environnement logiciel permettant le traitement automatisé de grandes bases de données d'images d'origines diverses, et la sélection de paramètres pertinents pour chaque corpus.

Publications :

TWEED T., MIGUET S. *Sélection automatique de régions d'intérêt pour la détection de zones cancéreuses dans les mammographies* in : Coopération Analyse d'Image et Modélisation, 14 Juin 2001, Lyon, Université Claude Bernard, Lyon **(2001)**.

TWEED T., SAADANE A. *L'apport d'un bloc de segmentation d'erreur dans l'évaluation de la qualité d'images codées.* GRETSI 2001, 10-13 Sept. 2001, Toulouse **(2001)**.

Gaëlle LEGRAND

Date de naissance : 24/05/1977

Directeur de thèse : Nicolas Nicoloyannis

Financement : Allocation de Recherche

Date de début de thèse : 1 Octobre 2001

Date de soutenance prévue : année 2003-2004

Titre de la thèse : Extraction de motifs séquentiels à partir des données structurées et non structurées

Mots-clés : patterns, motifs séquentiels, localisation et extraction de régularités, séries temporelles.

Résumé de la thèse

En matière de data mining nous sommes très souvent confrontée à la recherche de structures stables dans les données.

L'objectif de ce travail est la représentation des différentes problématiques associées à la découverte, la localisation et l'extraction de régularités ou encore de patterns stables dans les données qui peuvent être structurées ou non.

Plus particulièrement, nous sommes intéressée par les trois axes suivants :

- Les données sont structurées par une relation binaire (ordre, preordre, …)
- Les données sont structurées par un ordre naturel ou par une composante temporelle (séries chronologiques, séquences temporelles)
- Les données ne sont pas structurées a priori par une relation donnée (texte et web mining).

L'approche méthodologique que nous adopterons initialement pour la découverte et la localisation des motifs sera basée autour de l'algorithmique classique sur les séquences.

Après un tour d'horizon des différentes approche classiques et l'étude de l'art en matière de recherche de patterns stables dans les données, nous allons nous intéresser à la conception d'un système « modulaire coopératif » capable de traiter les trois types de données.

La méthodologie choisie sera inspirée des approches utilisées en reconnaissance des formes ainsi que l'approche prétopologique adaptée au contexte du sujet. La modélisation type multiagent et la théorie des jeux seront utilisées pour la partie cooperation.

ERRAY Walid

Date de naissance : 07/04/1977

Directeur de thèse : Djamel Zighed

Financement : Bourse EGIDE

Date de début de thèse : 01 Novembre 2001

Date de soutenance prévue : année 2003/2004

Titre de la thèse : Segmentation généralisée et application à l'identification des Churns en téléphonie mobile.

Mots clés : Segmentation généralisée, Sélection d'attributs, Construction d'attributs, Churns.

Résumé de la thèse

Ce travail a pour objectif la poursuite des recherches engagées avec Walid Erray dans le cadre de son mémoire de DEA. Il s'agira d'approfondir l'étude théorique et algorithmique de la généralisation du concept de graphes d'induction aux cas où toutes les variables utilisées en apprentissage sont quelconques et à nombre de modalités pouvant être très grand (plusieurs centaines).

Dans le cadre d'un contrat avec France Telecom, nous appliquerons ces techniques pour apprendre à identifier les clients qui risquent de rompre leur abonnement.