UNIVERSITÉ PARIS 13- INSTITUT GALILÉE

LABORATOIRE D'INFORMATIQUE DE PARIS NORD
UMR 7030 DU CNRS

Mémoire présenté en vue de l'obtention du diplôme d'

Habilitation à diriger des travaux de recherche

Spécialité : Informatique

Mustapha LEBBAH
Maître de conférences 27^{e} section du CNU

Contributions en apprentissage non supervisé à partir de données complexes

Document de synthèse

Soutenue le 27 janvier 2012

devant le jury composé de :

M.	Gilles	Venturini	Professeur	Président
M.	Djamel	Bouchaffra	Professeur	Rapporteur
M.	Marc	Gelgon	Professeur	Rapporteur
M.	Gérard	Govaert	Professeur	Rapporteur
M.	Djamel Abdelkader	Zighed	Professeur	Rapporteur
M.	Fouad	Badran	Professeur	Examinateur
M.	Younès	Bennani	Professeur	Examinateur
M.	Gérard	Duchamp	Professeur	Examinateur
Mme	Céline	Rouveirol	Professeur	Examinatrice

Remerciements

Tout d'abord j'exprime ma gratitude et mes respects à M. Djamel Bouchaffra (Professeur à Grambling State University), M. Marc Gelgon (Professeur à Polytech'Nantes), M. Gérard Govaert (Professeur à l'UTC de Compiègne) et M. Djamel Abdelkader Zighed (Professeur à l'Université Lumière Lyon 2), de m'avoir fait l'honneur d'accepter de rapporter sur cette thèse. Je tiens à remercier Mme Céline Rouveirol (Professeur à l'Université de Paris 13), M. Younès Bennani (Professeur à l'Université de Paris 13), M. Fouad Badran (Professeur au CNAM -Paris), Gérard Duchamp (Professeur à l'Université de Paris 13) et M. Gilles Venturini (Professeur à l'Université François-Rabelais de Tours) qui ont bien voulu marquer leur intérêt pour mon travail en acceptant d'être membre du jury. J'ai une haute estime pour chacun des membres de mon jury et je sais qu'il me faut encore travailler pour arriver à leur niveau.

Ce travail n'aurait pu exister sans le soutien des personnes qui m'ont accueilli dans leur équipe et laboratoire : Mme Sylvie Thiria (Professeur à l'université de Versailles), M. Fouad Badran (Professeur au CNAM -Paris), M. Jean-Daniel Zucker (Dr IRD), M. Younès Benanni (Professeur à l'Université de Paris 13) et Christophe Fouqueré (Professeur à l'Université de Paris 13). Ils ont beaucoup de qualités humaines et j'ai beaucoup appris d'eux. A travers eux j'ai pu connaitre et rencontrer des personnalités de valeurs humaines qui ont aussi participé de loin ou de près à ce travail. Je ne saurais leur exprimer suffisamment ma gratitude. J'ai eu le plaisir au LIPN comme au LIM&BIO de travailler dans des équipes dynamiques dont je remercie tous les membres.

Je remercie particulièrement M. Younès Bennani pour son soutien et les échanges fructueux et Mme Hanane Azzag non seulement pour la collaboration scientifique, mais aussi pour son soutien permanent. Plusieurs travaux ont été réalisés avec des doctorants avec qui j'ai trouvé une grande satisfaction à travailler. Cela a été pour moi un véritable plaisir. Je souhaite exprimer ma sympathie à tous les membres de l'équipe A3. Je remercie tous ceux qui ont participé de loin ou de près à ce travail. Je souhaite aussi saluer mes collègues de l'IUT de Villetaneuse.

Je termine ces remerciements par ceux destinés à ma petite et grande famille, particulièrement ma femme *louiza*, mes enfants *lina* et *ilyes* et mes parents qui n'ont pas besoin d'un long discours, ils savent très bien que je ne serais rien sans eux.

Résumé

Ce mémoire de synthèse est consacré à l'analyse des données complexes pour lesquelles la représentation des variables qui est toujours numérique rencontre des limites. L'ensemble des travaux présentés dans ce mémoire s'inscrit dans le cadre de l'apprentissage non supervisé dont la problématique consiste à construire des représentations simplifiées de données sans connaissance a priori des classes. Il existe actuellement un nombre conséquent de méthodes de partitionnement, mais elles ne s'adaptent pas toujours aux particularités de certains types de données (binaires, mixtes, séquences). On peut distinguer deux grandes familles de modèles de classification non supervisée : les modèles probabilistes et les modèles déterministes ou tout simplement les modèles de quantification. Dans ce mémoire, une importance particulière est accordée aux modèles des cartes topologiques auto-organisatrices. Deux modèles sont proposés pour le traitement des données mixtes (continues et qualitatives). Dans le premier modèle des modifications de la distance sont apportées pour prendre en compte le type de variables. Dans le deuxième modèle, des cartes topologiques dédiées aux données binaires et mixtes sont proposées, utilisant la distribution gaussienne et de Bernoulli. Un autre axe étudié dans ce mémoire est celui de l'apprentissage de données structurées en séquences (non i.i.d). Un lien étroit est montré entre les chaînes de Markov cachées et les cartes à base de modèles de mélanges. Enfin, un bilan des travaux est présenté tout en fournissant des perspectives générales.

Mots clés : apprentissage non-supérvisé, cartes auto-organisatrices, modèles de mélanges, chaînes de Markov, données binaires, données catégorielles, données mixtes, données séquentielles.

Table des matières

Préface

Ce mémoire fait la synthèse de mes activités de recherche et d'enseignement les plus importants effectués en tant que maître de conférences à l'université de Paris 13.

Mes travaux de recherche portent sur l'analyse des données complexes pour lesquelles la représentation des variables qui est toujours numérique rencontre des limites. L'ensemble de mon activité de recherche s'inscrit dans le cadre de l'apprentissage non supervisé dont la problématique consiste à construire des représentations simplifiées de données, pour mettre en évidence les relations existantes entre les caractéristiques relevées sur des données et les ressemblances ou dissemblances de ces dernières, sans avoir aucune connaissance sur les classes. La quantité et la complexité croissantes de données posent des problèmes pour les experts du domaine ou les utilisateurs qui analysent ces données. Ils ne peuvent pas s'appuyer sur des techniques entièrement automatiques pour l'analyse des données. Ainsi, la visualisation est une manière d'interagir avec les experts et de pouvoir communiquer avec eux afin d'affiner les paramètres. Une visualisation intelligente des données et son interaction avec les méthodes issues de l'apprentissage traditionnelle jouent un rôle central dans ce processus. Parmi les objectifs de mon travail, l'un consiste à développer des approches fournissant une classification automatique des données et permettant à la fois une réduction des dimensions afin d'offrir un espace de visualisation convivial.

Nous disposons actuellement d'un nombre conséquent de méthodes de clustering/partitionnement, mais elles ne correspondent pas pour autant toujours aux particularités de certains types de données (binaires, mixtes, séquences, graphe). Dans mes travaux, j'accorde une importance particulière aux modèles des cartes topologiques auto-organisatrices. Ce type de modèles permet de bien s'adapter aux traitements des données complexes, aussi de proposer à la fois un clustering et une réduction des dimensions en proposant des visualisations 1D, 2D. Je montre dans ce mémoire leur capacité aussi, de traiter des données non i.i.d. Toutes les contributions que je propose sont valables pour les différents algorithmes de partitionnement de type topologique (Neural Gas [MS91], Growing Neural Gas [Fri95] ou le H^2SOM [OR06]). Ainsi, les modèles topologiques, que je propose, ont en commun avec les autres modèles : l'utilisation de la notion de grille ou de graphe comme espace de projection.

Dans de nombreux domaines, par exemple l'analyse des enquêtes d'opinion ou les données médicales, les variables observées ne sont pas des données continues classiques. Les variables ainsi définies sont appelées variables qualitatives nominales ou catégorielles. Si cette variable n'a que deux modalités, on parle souvent simplement de variable binaire. Souvent aussi, ces variables sont mixtes. Ainsi, les praticiens utilisent parfois des méthodes statistiques destinées aux données quantitatives, sans se préoccuper de la spécificité des données et de la perte d'information. On peut distinguer deux grandes familles de modèles de classification

non supervisée : les modèles probabilistes et les modèles déterministes ou tout simplement les modèles de quantification.

Mon travail apporte des contributions aux deux familles. Je me suis intéressé à l'analyse de données mixtes : continues et qualitatives, donc le but est la définition et l'étude de méthodes de partitionnement pour des observations décrites par des variables qualitatives, binaires et continues. Pour la première famille, j'ai proposé des modifications de la distance afin qu'elle prenne en compte le type de variables. J'ai toujours veillé à ce que les modèles que je propose puissent fournir des résultats interprétables. Ceci se traduit par la conservation de la structure initiale des données en proposant des prototypes du même type que les données initiales. La difficulté principale en ce qui concerne la deuxième famille ou l'application des modèles de mélange est qu'ils sont consacrés pour un type de données : les distributions normales pour les données quantitatives. Cela est regrettable puisque la plupart des ensembles de données impliquent différents types. Dans cet axe j'ai proposé un modèle de cartes topologiques dédié aux données binaires en utilisant la distribution de Bernoulli et puis un modèle probabiliste pour les données mixtes.

Le second axe de recherche que j'étudie est celui de l'apprentissage de données structurées en séquences. Je me suis intéressé à étendre le modèle probabiliste topologique présenté pour les données i.i.d aux données séquentielles (non i.i.d.). De la même manière, les approches que je propose au chapitre 5, mes travaux sur les données séquentielles, relèvent du même objectif que celui de prendre en considération la nature et l'aspect séquentiel des données au cours de l'apprentissage des cartes auto-organisatrices. Je montre un lien étroit entre les chaînes de Markov cachées et les cartes à base de modèles de mélanges. En fait, les HMMs ne sont qu'un cas particulier des modèles de mélanges topologiques. Le modèle proposé dans cet axe de recherche permet de visualiser les séquences dans un espace réduit présenté sous forme d'une grille 2D ou 1D.

Organisation du mémoire

J'ai tenté dans ce mémoire de rester concis, tout en donnant suffisamment de détails pour rendre la lecture intelligible. les détails techniques se trouvent aussi dans les papiers joints à ce mémoire. Ce mémoire est structuré en deux parties :

La partie I est découpée en trois chapitres qui fournissent une synthèse de mon activité de recherche et d'enseignement. Le chapitre 1 dresse un curriculum vitae décrivant mon parcours professionnel et académique. Il dresse un bilan de mes activités en responsabilité d'enseignement et administrative ; ma participation à la valorisation de la recherche par des coopérations industrielles. Une section est consacrée aux différents encadrements de doctorants ou de masters 2. Le chapitre 2 dresse mon parcours scientifique en se focalisant sur les travaux menés depuis mon recrutement à l'université de Paris 13 en septembre 2005. Je termine ce chapitre par une synthèse des collaborations et des collaborateurs avec qui j'ai eu le plaisir de travailler. Le chapitre 3 dresse une liste complète des brevets, publications nationales et internationales.

La partie II dresse une synthèse scientifique des travaux de recherche. Le chapitre 4 rappelle le contexte et les problématiques sur lesquelles repose l'ensemble de mes travaux de recherche. Je fournis dans ce chapitre une synthèse des différentes contributions. Quelques travaux prospectif de par les données traitées et les approches abordées sont aussi présentées.

Le chapitre 5 se focalise sur les principaux travaux menés en apprentissage non supervisé dédié au données binaires catégorielles et mixtes (continues et binaires). Le chapitre 6 se concentre sur les résultats principaux concernant l'apprentissage non supervisé avec des données séquentielles ou non i.i.d. Enfin, je termine par un bilan de mes travaux tout en fournissant quelques perspectives générales et propres à chaque axe.

Première partie

Rapport d'activités

Chapitre 1

Curriculum Vitæ

Mustapha LEBBAH

Maître de Conférences (CNU 27)

Né le	:	16/05/1975, à Oran
Etat Civil	:	Marié, 2 enfants
Laboratoire	:	Laboratoire d'Informatique de Paris-Nord **(LIPN, UMR 7030)**
Equipe	:	Apprentissage Artificiel & Applications**(A3)**
Adresse	:	Université Paris 13, LIPN-UMR 7030 99, av. J.-B. Clément F-93430 Villetaneuse
Enseignement	:	Université Paris 13, IUT de Villetaneuse, département Informatique
Tél	:	01 49 40 28 31
Fax	:	01 48 26 07 12
Courriel	:	mustapha.lebbah@univ-paris13.fr, lebbah@gmail.com
Page web	:	http://www-lipn.univ-paris13.fr/~lebbah/

1.1 Résumé du CV

Je suis maître de conférences à l'université de Paris 13 depuis 2005 et membre permanent de l'équipe A3 (Apprentissage Artificiel et Applications) du laboratoire LIPN (UMR 7030 CNRS). Je participe activement à l'organisation d'événements scientifiques au niveau national et international. Je suis membre et animateur du groupe de travail "Fouille de Données Complexes" de l'association EGC (Extraction et Gestion des Connaissances). Je suis aussi membre du bureau Data Mining et Apprentissage de la SFds (Société Française de Statistique) et membre de INNS (International Neural Network Society). Je suis titulaire de la Prime d'excellence scientifique (PES) (2009/2013). Mon thème de recherche principal concerne les systèmes d'apprentissage statistique pour la fouille de grands volumes de données complexes. Une de mes expertises concerne la classification probabiliste non supervisée à base de modèles de mélanges.

J'ai commencé ma carrière de chercheur en master 2 (ex DEA) dans le cadre d'un stage de recherche chez L'OREAL R&D. Par la suite, j'ai occupé un poste d'ingénieur de recherche chez Renault R&D pendant trois ans (2000-2003, contrat CIFRE). J'ai obtenu en 2003 mon doctorat en informatique à l'université Versailles Saint-Quentin en Yvelines. J'ai toujours dépassé le cadre applicatif en développant des modèles généraux. Ainsi par la suite, j'ai pu faire un postdoc au laboratoire CETP (Centre d'étude des Environnements Terrestres et Planétaires) et d'être recruté au laboratoire de bioinformatique le LIM&BIO en 2005. Mon parcours m'a permis d'évoluer dans différents domaines : le secteur de l'automobile, l'analyse des enquêtes, la bioinformatique, la recherche d'informations, et l'environnement. Mes recherches sont le résultat de l'influence du contexte, des projets en cours et des différentes rencontres, avec une constante qui est celle de développer des modèles non supervisés en tenant compte de la structure spatiale et de la nature des données.

Synthèse des Enseignements
J'ai commencé à enseigner en 2000 en tant que vacataire au CNAM (Conservatoire National des Arts et métiers). Ensuite, j'ai été ATER à l'ENSIIE (l'Ecole Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise) pendant deux années (2003-2005). Depuis septembre 2005, date à laquelle j'étais recruté, j'enseigne à l'université de Paris Nord où j'effectue entre 192 heures et 220 heures équivalent TD par an avec seulement 128 heures au cours des deux premières années (2005/2007 : Modulation des heures d'enseignement décidée par la CS de l'université de Paris 13 pour les jeunes MCF recrutés). Pour l'année 2010/2011, je n'ai pas eu d'enseignement car j'étais en délégation CNRS au LIPN.

Publications
- Revues : internationales : 6 ; nationales : 3
- Brevet : 2 (RENAULT (2001) ; LIPN&THALES (2009))
- Chapitre de livre : National : 2 ; International : 2
- Conférences internationales avec comité de sélection : 20
- Conférences nationales avec comité de sélection : 15

1.2 Cursus universitaire

-2003 Doctorat en informatique de l'université de Versailles Saint-Quentin en Yvelines. **Mention :** Très honorable

Titre : *Carte topologique pour données qualitatives : application à la reconnaissance automatique de la densité du trafic routier. Brevet déposé.*

Soutenue devant le **jury** composé de **:** M. Jean-Michel **FOURNEAU**(Président), M. Patrick **GALLINARI**(Rapporteur), M. Ludovic **LEBART**(Rapporteur), Mme. Sylvie **THIRIA** (Directeur de thèse), M. Christian **CHABANON**(Examinateur), M. Fouad **BADRAN** (Examinateur), M. Gérard **GOVAERT** (Examinateur).

Financement : Contrat CIFRE (Université UVSQ[1]-RENAULT R&D). Laboratoire d'Informatique PRiSM et LOCEAN-Univ-Paris 6[2].

-1999 DEA d'intelligence artificielle. Université de Paris 13

04/1999 - 09/1999 Stage de DEA chez L'OREAL : Centre de recherche avancée Etude des méthodes de classification appliquées à une base de données de molécules de shampooings. Application des cartes topologiques (réseau de neurones), aux données binaires représentant les molécules.

-1998 Diplôme d'Ingénieur en Informatique (Génie Logiciel), USTO-Algérie.

11/1997- 07/1998 Stage de fin d'études d'ingénieur
Université des Sciences et de la Technologie d'Oran (USTO-Algérie)
Laboratoire de communication homme-machine. -Etude des méthodes de compression du signal avec et sans perte d'informations (image/parole) -Etude des bancs de filtres et leurs applications. Filtrage multicadence, filtres QMF.

[1] Université de Versailles Saint-Quentin-en-Yvelines
[2] LOCEAN : Laboratoire d'Océanographie et du Climat : Expérimentation et Approches Numériques à l'université de Paris 6

1.3 Parcours professionnel

09/2010-09/2011 Délégation CNRS au LIPN-UMR 7030, Laboratoire d'Informatique de Paris Nord. Université de Paris 13.

09/2005 à aujourd'hui Maître de Conférences à l'IUT de Villetaneuse , Université de Paris 13. Membre du laboratoire LIPN-UMR 7030. **Equipe :** Apprentissage Artificiel & Applications (A3)

09/2004-09/2005 Attaché temporaire d'enseignement et de recherche à l'Ecole Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise (ENSIIE ex IIE-CNAM).

10/2003-09/2004 Attaché temporaire d'enseignement et de recherche (1/2 ATER) à l'Ecole Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise (ENSIIE).

10/2003-09/2004 -Ingénieur de recherche à mi-temps au LOCEAN[3] Unité Mixte de Recherche 7159 CNRS / IRD /P6 -Contrat de recherche avec L'OREAL
-Equipe : Modélisation et Méthodes Statistiques Avancées du LOCEAN/UPMC
Extraction d'information à partir d'une base de données médicales appelées SuviMax .

02/2003- 09/2003 Ingénieur de recherche (Post-Doc) au Centre d'étude des Environnements Terrestre et Planétaires (CETP) : UMR CNRS 8639
Application d'une technique d'assimilation variationnelle basée sur les méthodes adjointes et sur les techniques de la programmation modulaire actuellement développées au LODYC. Cette approche est appliquée au modèle physique de surface ISBA qui décrit les interactions entre le sol, la végétation et l'atmosphère. Développement d'une interface homme-machine dédié à la programmation modulaire.

02/2000- 01/2003 Ingénieur de recherche (CIFRE) : Technocentre de RENAULT
Direction de la recherche ; groupe. Détection de l'Environnement et Aide à la Conduite -Analyse et développement d'un modèle breveté de réseau de neurones dédié à l'estimation de la densité de trafic. -Prototypage rapide du modèle sur cible calculateur (Simulink/RTW, Dspace, TargetLink).-Proposition de solution logicielle pour le développement temps-réel embarqué utilisant le bus CAN (client/serveur). Validation sur VelSatis dans le cadre du projet ACC (Autonomus Cruise Control). Préparation de la campagne d'essais clients. -Encadrement de deux stagiaires du DESS TRIED : Traitement de l'Information et Exploitation des Données.

[3]LOCEAN : Laboratoire d'Océanographie et du Climat :Expérimentation et Approches Numériques

1.4 Activités et responsabilités d'enseignement

J'ai commencé à enseigner en 2000 en tant que vacataire au ***CNAM*** *(Conservatoire National des Arts et métiers). Ensuite, j'ai été* ***ATER à l'ENSIIE*** *(l'Ecole Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise) pendant deux années (2003-2005). Depuis septembre 2005, j'enseigne à* ***l'université de Paris Nord*** *où j'effectue entre* ***192 heures et 220 heures équivalent TD par an avec seulement 128*** *heures au cours des deux premières années (2005/2007 : Modulation des heures d'enseignement décidée par la CS de l'université de Paris 13 pour les jeunes MCF recrutés).* ***Pour l'année 2010/2011, je n'ai pas eu d'enseignement car en délégation CNRS au LIPN.***

Une synthèse des quelques enseignements est donnée ci-dessous :

Base de la programmation
Chargé de TD/TP (67.5h éq TD) - 2 ans (2008-2010).
Enseignement destiné à des étudiants de première année à l'IUT. Il s'agit d'initier les étudiants à l'algorithmique et aux bases de la programmation, en particulier en C.

Base de données
Chargé de cours et de TD/TP (65h éq TD) - 4 ans (2007-2008). Enseignement destiné à des étudiants de première année à l'IUT. Le cours comprend : le modèle relationnel, conception d'une base de données, algèbre et calcul relationnel, SQL. Aspects pratiques : création d'un schéma relationnel, vues, contraintes.

Apprentissage numérique
Enseignement destiné à des étudiants en Master 2 recherche. Chargé de cours et de TD/TP (20h éq TD) - 1 an (2007-2008). Introduction à l'apprentissage supervisé et non supervisé, réseaux de neurones (les réseaux multicouches). Les modèles topologiques.

Introduction à l'Intelligence Artificielle
Enseignement destiné à des étudiants de 3^e année d'ingénieur (ENSIIE). Chargé de cours et de TD/TP (22h éq TD) - 1 an (2004-2005). Programme du cours : Historiques et applications de l'apprentissage. Introduction à l'apprentissage supervisé et non supervisé. Arbre de décision, réseaux de neurones (les réseaux multicouches). Extraction de règle à partir d'un réseau multicouches. Les cartes topologiques de Kohonen. Les SVMs. Le boosting et le bagging.

Responsabilités. J'ai par ailleurs été responsable pédagogique en informatique à l'IUT pendant 2 ans (2008-2010). Ceci se traduit par la répartition des services d'enseignement, recrutement de vacataires. Je vais être directeur des études du semestre 2 pour l'année 2011/2012. Ceci se traduit par la mise en place de l'emploi du temps, le suivi des étudiants et préparation des commissions.

Tableau récapitulatif

(**CM** : Cours Magistraux, **TD** : Travaux Dirigés ; **TP** : Travaux Pratiques ; **GT** : Groupe de Travail). **Ex** : Existant. **Vol** :Volume.
Pour l'année universitaire 2010/2011, j'étais en délégation CNRS à 100%.

Année	Filière	Matière	Support	Type	Vol
2009-2010 (MCF)	1^{re} année (IUT-INFO)	Base de données (PostgreSQL)	Créé	CM/ TD/ TP	65h
	2^{e}année (IUT-INFO)	Réseaux	EX	TD	36h
	1^{e} année (IUT-INFO)	Base de la programmation	EX	TD/TP	67.5h
	2^{e} année (IUT-INFO)	Programmation Web	EX	TD/TP	21h
	2^{e} année (IUT-INFO)	Projet	Créé	TD	11h
2008-2009 (MCF)	1^{re} année (IUT-INFO)	Base de données (PostgreSQL)	Créé	CM/TD/ TP	65h
	2^{e}année (IUT-INFO)	Réseaux	Créé/EX	CM/TD	25h
	1^{re} année (IUT-INFO)	Base de la programmation	EX	TD/TP	67.5h
	1^{re} année (IUT-INFO)	Info et signaux	EX	TD/TP	18h
	2^{e} année (IUT-INFO)	Projet	Créé	TD	11h
2007-2008 (MCF)	Master 2 MICR	Apprentissage numérique	Créé	TD	20h
	1^{re}année (IUT-INFO)	Projet Personnel et Professionnel	EX	TD	25h
	1^{re} année (IUT-INFO)	Base de données	Créé/EX	CM/TD	65h
	2^{e}année (IUT-INFO)	Réseaux	Créé/EX	CM/TD	60h
	1^{re} année (IUT-INFO)	Base de la programmation	EX	TD/TP	67.5h

Année	Filière	Matière	Support	Type	Vol
2006-2007 (MCF)	1^{re} année (IUT-INFO)	Projet Personnel et Professionnel	EX	TD	22.50h
	1^{re} année (IUT-INFO)	Base de données	EX	TD/TP	36h
	2^{e} année (IUT-INFO)	Réseaux	Créé/EX	CM/TD	69.50h
	2^{e} année (IUT-INFO)	Conception avancée de base de données	EX	TD/TP	9h
2005-2006 (MCF)	1^{re} année (IUT-INFO)	Projet Personnel et Professionnel	EX	TD	22.50h
	1^{re} année (IUT-INFO)	Base de données	EX	TD/TP	36h
	2^{e} année (IUT-INFO)	Réseaux	EX	TD/TP	69.50h
2004-2005 (ATER)	$3^{\text{ème}}$ année (ENSIIE[4])	Intelligence Artificielle	Créé	CM	22h
	3^{e} année (ENSIIE)	Intelligence Artificielle	Créé	TP	9h
	3^{e} année (ENSIIE)	Intelligence Artificielle	Créé	GT	16h
	3^{e} année (ENSIIE)	Réseaux de neurones	Créé	CM	18h
			Créé	TP	9h
	1^{re} année (ENSIIE)	Algorithmique et programmation	EX	TD/TP	32h
	1^{re} année (ENSIIE)	Base de données	EX	TD/TP	48h
	2^{e} année (ENSIIE)	Recherche Opérationnelle	Créé	TP	40h
2003-2004 (ATER)	Master Bio-Informatique	Introduction aux réseaux et à la programmation Internet	Créé	CM	25h
	1^{re} année (ENSIIE)	Algorithmique et programmation	EX	TD/TP	32h
	2^{e} année (ENSIIE)	Recherche Opérationnelle	Créé	TP	40h
1999 - 2000	Cycle A au CNAM	Structure de données	EX	TD	60h
1999 - 2000	Salariés de Renault	Programmation C/C++	Créé	CM	–

1.5 Coopérations industrielles et valorisation

2010-fin 2013 Participation au **Projet ANR E-FRAUD**
Durée : 2010-2013
Participant : KXEN, LIP6, THALES, LIPN
Le projet E-fraud Box vise à développer une boîte à outils intégrée, dédiée à la détection et à l'investigation de la fraude à la carte bancaire sur Internet.
http://efraud.lip6.fr/pmwiki.php?n=Main.PresentationDuProjet.
Productions : [CI-3], [CN-5]

2006–fin 2009 Participation au **Projet Infom@gic** [Dans le cadre du pôle de compétitivité Cap Digital (IMVN : Image, Multimédia et Vie Numérique)].
Pilote industriel : THALES **Durée :** 2006-2009.
Ce projet consiste à mettre en place, sur une période de trois ans, un laboratoire industriel de sélection, de test, d'intégration et de validation sur des applications opérationnelles des meilleures technologies franciliennes dans le domaine de l'ingénierie des connaissances. Ce laboratoire s'appuie sur une plate-forme commune d'interopérabilité. Cette plateforme doit couvrir les grands domaines fonctionnels et techniques de l'analyse d'information que sont la recherche et l'indexation, l'extraction de connaissances, et la fusion d'informations multimédias, sur tous les types de sources : texte, données, images, son.
Productions : [BR-1], [CI-7, CI-9], [CHN-1, CHN-2] ; [CN-7, CN-9] [COL-2] ***et plusieurs rapports internes***

2007-2008 Porteur du projet Bonus Qualité Recherche (BQR Université de Paris 13) : Partitionnement de grandes bases de données médicales à très large dimensionnalité : Utilisation des modèles bio-inspirés.
Durée : 1 an
Productions : [CI-12, CI-8], [CN-10]

2005 -2007 LIM&BIO-Hopital Hôtel-Dieu : Participation aux recherches sur l'obésité au service nutrition de l'hôpital Hôtel-Dieu. Développement de nouveaux modèles de visualisation et de prédiction de perte de poids.
Productions : [CN-11], [CN-13, CI-16]

2005-2007 CSTB[5]-LOCEAN[6] (UMR 7159) : Participation aux contrats de recherche CNRS-LOCEAN-UPMC & CSTB Analyse des déterminants de la qualité de l'air intérieur dans les logements. Suivi des travaux de recherche et proposition de nouvelles stratégies d'analyse de données. (Participation à l'encadrement des stagiaires sur le sujet).
Productions : [COL-1]

Avant 2005 , j'ai été conduit à administrer les tâches menées par l'équipe pour le contrat de recherche avec RENAULT (Contrat CIFRE) et L'OREAL (Contrat de recherche),

[5]Centre Scientifique et Technique du Bâtiment
[6]LOCEAN : Laboratoire d'Océanographie et du Climat :Expérimentation et Approches Numériques

d'une part en proposant des solutions pour le projet, et d'autre part en rédigeant des livrables. J'ai également contribué à l'élaboration de la suite du contrat de recherche avec L'OREAL.

1.6 Encadrement de doctorants et de masters

1.6.1 Thèses

Dès ma nomination en tant que maître de conférences le **01/09/2005**, j'ai participé à l'encadrement de thèses. Je suis titulaire de la Prime d'excellence scientifique (*PES*) (2009/2013).

Thèses en cours

Doctorant *:* M. Amine Chaibi
- ***Titre*** : Fouille de données et Apprentissage statistique pour la prévision
- ***Etat : Soutenance prévue*** *décembre 2013*
- ***Encadrement : 50% Lebbah, 50%Azzag***
- ***Financement :*** *Contrat CIFRE*

Doctorant *:* M. Nhat-Quang Doan
- ***Titre*** : Fouille de grands graphes et visualisation hiérarchique
- ***Etat : Soutenance prévue*** *décembre 2013*
- ***Encadrement : 50% H. Azzag, 50% Lebbah***
- ***Financement :*** *Bourse vietnamienne*

Doctorante *:* Mlle Rakia Jaziri
- ***Titre*** : Apprentissage non supervisé de données structurées en séquences
- ***Etat : Soutenance prévue*** *décembre 2012*
- ***Encadrement : 50% Lebbah, 50% Bennani***
- ***Financement :*** *Contrat CIFRE*
- ***Productions*** : [CI-1, CI-2], [CN-2]

Thèses soutenues

Doctorante *:* Mlle Nicoleta Rogovschi
- ***Titre*** : Apprentissage topographique par des modèles de mélanges pour des données mixtes
- ***Etat :*** Soutenue le 4 décembre 2009
- ***Directeur de thèse :*** Younès Bennani *(Professeur Univ. Paris 13)*
- ***Encadrement : 50% Lebbah, 50% Bennani***
- ***Productions :*** [RI-5, RI-4], [CI-13, CI-10], [CN-8, CN-12]
- ***Situation actuelle :*** *Maître de conférence à l'université Paris 5*

Doctorant *:* M. Mohamed-Ramzi Temanni
- ***Titre*** : Combinaison de sources de données pour l'amélioration de la prédiction en apprentissage : une application à la prédiction de la perte de poids chez l'obèse à partir

de données transcriptomiques et cliniques
- ***Etat :*** Soutenue le 26 juin 2009
- ***Directeur de thèse :*** Jean-Daniel Zucker *(DR IRD)*
- ***Encadrement :*** *20% Lebbah, 80% Zucker*
- ***Productions*** : [CN-11, CN-13], [CI-16]
- ***Situation actuelle*** *: Consultant en bioinformatique*

1.6.2 Masters

1) Mlle Liu YuQiong (Masters EID : Exploration Informatique des Données et Décisionnel, Univ-Paris 13)
Période : 04/2011-09/2011
Titre : Classification et structuration des séquences audiovisuelles
Encadrement : Jaziri, Lebbah
2) M. Zayed Yakoubi (Master 2 informatique, Univ-Dauphine)
Période : 04/2010-09/2010
Titre : Classification par modèles de mélange des données séquentielles
Encadrement : Lebbah
3) M. Aymen Arfaoui (Master 2 informatique, Univ-Dauphine)
Période : 04/2010-09/2010
Titre : Classification topologique et hiérarchique des grands graphes
Encadrement : Lebbah, Azzag
Productions : [CI-4]
4) M. Brahim Laouissed (Master 2 informatique, Univ-Paris 8).
Période : 04/2009-09/2009
Titre : Sous-échantillonnage topographique pour bases déséquilibrées
Encadrement : Lebbah
5) M. Chafik Mebarek (Master 2 MSIR : Modèles, systèmes, Imagerie, robotique, Université Blaise Pascal)
Période : 04/2009-09/2009
Titre : Classification hiérarchique et règles de décision
Encadrement : Lebbah, Azzag
6) Mlle Maia Iordatii (Master 2 ITCN :Ingénierie des Textes et Contenus Numériques, Univ-Paris 13)
Période : 04/2008-09/2008
Titre : Apprentissage topographique avec mémoire
Encadrement : Lebbah, Bennani
Productions : [CHN-2, CI-9]
7) M. Ali Lajnef (Master 2 ITCN :Ingénierie des Textes et Contenus Numériques, Univ-Paris 13)
Période : 04/2008-09/2008
Titre : Apprentissage topographique et points d'intérêts
Encadrement : Lebbah, Azzag
8) M. Nabil Ferradj (Master 2 TRIED : Traitement de l'Information et Exploitation des Données, Univ-UVSQ).

Période : 04/2007-09/2007.
Titre : Agrégation automatique et méthode de partitionnement : application aux données médicales
Encadrement : Lebbah

9) M. Mohamed Al Othman (Master 2 ITCN : Ingénierie des Textes et Contenus Numériques, Univ-Paris 13)
Période : 04/2007-09/2007
Titre : Apprentissage non supervisé et segmentation automatique
Encadrement : Lebbah, Azzag
Productions : [CI-8, CI-12, CN-10]

10) Mme Hanane Benaribi (Master 2 TRIED : Traitement de l'Information et Exploitation des Données, Univ-UVSQ)
Période : 04/2005-09/2005
Titre : Analyse des déterminants de la qualité de l'air intérieur dans les logements
Encadrement : Lebbah, Thiria
Productions : [COL-1]

11) Mr Heiko AGNOLI (Master 2 TRIED : Traitement de l'Information et Exploitation des Données, Univ-UVSQ)
Période : 04/2002-09/2002
Titre : Etude des transitions du trafic routier : Utilisation des cartes topologiques
Encadrement : Lebbah

12) Mme LI BERNIER (Master 2 TRIED : Traitement de l'Information et Exploitation des Données, Univ-UVSQ)
Période : 04/2001-09/2001
Titre : Reconnaissance du trafic à partir des descripteurs symboliques
Encadrement : Lebbah

1.7 Participation à la vie scientifique et administrative

1.7.1 Membre de sociétés savantes

- **Membre de** : International Neural Network Society (INNS), http://www.inns.org/
- **Membre et animateur du groupe de travail FDC** : Fouille de Données Complexes de l'association EGC : Extraction et Gestion des Connaissances. http://eric.univ-lyon2.fr/~gt-fdc/
- **Membre du bureau de la SFds**. Fonction : Correspondant entre la société française de classification (SFC) et la Société Française de Statistique (SFds). http://www.sfds.asso.fr/
- **Membre du bureau du groupe** Data Mining et Apprentissage de la SFds. Fonction : Webmaster. http://www.sfds.asso.fr/83-Presentation

1.7.2 Responsabilités administratives et scientifiques

1.7.2.1 Comité d'organisation de colloques/workshops

National

- **Co-organisateur du colloque sur la fouille de données complexes et grands graphes. (GT :EGC-FDC-FGG). Les 20 et 21 juin 2011** au CNAM. http://eric.univ-lyon2.fr/~gt-fdc/ ; Organisateurs : Hanene Azzag (LIPN, Univ. Paris 13), Guillaume Cleuziou (LIFO, Univ. d'Orléans), Cyril de Runz (CReSTIC, Univ. de Reims), Mustapha Lebbah (LIPN, Univ. Paris 13), Fabien Picarougne (LINA, Univ. Nantes), Bruno Pinaud (LABRI, Bordeaux), Cédric Wemmert (LSIIT-AFD, Univ. Strasbourg).

- **Co-organisateur de la** 8^{e} **édition** de l'atelier, associé à la conférence **EGC'2011**, Fouille de données complexes -complexité liée aux données multiples-. Organisateur : Guillaume Cleuziou (LIFO, Université d'Orléans), Cyril de Runz (CReSTIC, Université de Reims Champagne-Ardenne), Mustapha Lebbah (LIPN, Université de Paris 13), Cédric Wemmert (LSIIT, Université de Strasbourg). https://sites.google.com/site/fdcegc11.

- **Co-organisateur d'une session spéciale** Apprentissage et modèles de mélanges dans le cadres des 43^{es} journées de statistique. Organisateur : Younès Bennani (LIPN, Université Paris 13), Christophe Biernacki (Laboratoire de Mathématiques de Lille), Mustapha Lebbah (LIPN, Université Paris 13), Mohamed Nadif (LIPADE, Université de Paris 5). **2011** http://jds2011.tn.refer.org/

- **Co-organisateur du colloque sur la Fouille de Données Complexes (GT :EGC-FDC) : Complexité liée aux données multiples. le 28 Juin 2010** à l'université de Paris 13. http://eric.univ-lyon2.fr/~gt-fdc/ ; Organisateurs : Guillaume Cleuziou, (LIFO, Université d'Orléans), Mustapha Lebbah, (LIPN, Université Paris 13), Cédric Wemmert, (LSIIT-AFD, Université de Strasbourg).

- **Membre du comité d'organisation de l'école d'hiver EGC'10 (é-EGC)** . http://www-lipn.univ-paris13.fr/~bennani/e-egc/.

- **Co-organisateur de la 7^{e} édition** de l'atelier, associé à la conférence **EGC'2010**, Fouille de données complexes -complexité liée aux données multiples-. Organisateur : Boutheina Ben Yaghlane (LARODEC, IHEC Carthage), Guillaume Cleuziou (LIFO, Université d'Orléans), Mustapha Lebbah (LIPN, Université de Paris 13), Arnaud Martin (ENSIETA, Brest). http://sites.google.com/site/afdcegc10.

- **Membre du comité d'organisation de la conférence** CAp (Conférence d'Apprentissage) **2009** et la plateforme AFIA (Association Française pour l'Intelligence Artificielle) 2009. http://cap09.lipn.fr/.

- **Membre du comité d'organisation et d'animation du Groupe de Travail EGC-FDC** sur la Fouille de Données Complexes (GT : EGC-FDC). Complexité liée aux données multiples. Dernière réunion **le 18 Juin 2009**. Organisateurs : Arnaud Martin (ENSEITA), Guillaume Cleuziou (Univ. Orleans), Mustapha Lebbah (Univ. Paris 13). http://eric.univ-lyon2.fr/~gt-fdc/ ; http://www.ensieta.fr/e3i2/jfdc/Programme_reunion_18juin2009.htm.

- **Responsable du séminaire AAPN : Apprentissage Artificiel à Paris Nord de 2006 à 2009**. Je m'occupe de l'organisation du séminaire apprentissage en commun LIPN (Invitation des personnes, programme...). http://www-lipn.univ-paris13.fr/actualites/tag/aapn

- **Participation à l'organisation de 3^{e} Journées** thématiques "Apprentissage Artificiel & Fouille de Données" **AAFD'2008**. mardi 8 et mercredi 9 avril 2008. http://www-lipn.univ-paris13.fr/A3/AAFD08/cfp.html.

International

- **Participation à l'organisation de la session spéciale à ICMLA** (the Eighth International Conference on Machine Learning and Applications) **2009** : Machine Learning Methods for Modeling Treatment Outcomes in Cancer and Radiation Therapy . Organisateurs : Dr Issam El Naqa, Washington University in St. Louis & Dr Steve Jiang, University of California, San Diego. http://www.icmla-conference.org/icmla09/CFP_SpecialSession1.html

- **Co-organisateur de la session spéciale** : Special Session : Incremental Topological Learning Models and Dimensional Reduction. The 5^{th} International Conference on Neural Network and Artificial Intelligence (ICNNAI'2010) ICNNAI-2010. **1-4 June, 2010**. Brest State Technical University, Belarus. Organisateurs : Nistor GROZAVU (Univ. Paris 13), Mustapha LEBBAH (Univ. Paris 13), Younès BENNANI (Univ. Paris 13).

1.7.2.2 Comité de programme & rapporteur

- **2011 :** Rédacteur invité du numéro spécial de la revue RNTI : Revue des nouvelles technologies de l'information. Fouille de Données Complexes.
- **2009 :** Rédacteur invité du numéro spécial de la revue RNTI : Revue des nouvelles technologies de l'information. Apprentissage & Visualisation.
- **2008-2010 :** Membre du comité de lecture de la revue : RNTI : Revue des nouvelles technologies de l'information (Apprentissage artificiel et fouille de données).

- **Membre du Comité de programme & rapporteur** :
  - **Nationales :** EGC 11, EGC'10, EGC'08, EGC'09, MCSEAI'08, NTICRI'09, CAp'10, CAp'09, SFC'09, SFC'10,SFC'11. **Atelier** : MLV'09, AGS'09, Atelier de Fouille de données et algorithmes biomimétiques-EGC 2007

  - **Internationales :** ICMLA 2009, Compstat'10, IJCNN'10, IJCNN'11,NaBIC 2011, ICMLA 2011

- **Chairman** dans une session à CAp'2007, ICMLA'2008, SFC'2011.

- **Relecteur** pour des revues internationales : régulièrement relecteur de Neurocomputing journal, neural computation.

1.7.2.3 Expertises

- **Expert** : J'ai rapporté un projet CIFRE soumis à l'ANRT.

- **Jury de thèse** : Membre du jury des thèses de :

  - **M. Nistor Grozavu**, Université de Paris 13, LIPN. Sujet de thèse : "Classification Topologique pondérée : approches modulaires, hybrides et collaboratives". Soutenue le : 8 décembre 2009 (Très Honorable). Jury : A. Aussem, Y. Bennani, M. Lebbah, J-F. Marcotorchino, JL. Zarader, JD. Zucker. Rapporteurs : P. Kuntz, M. Verleysen.

  - **Mlle Nicoleta Rogovschi**, Université de Paris 13, LIPN. Sujet de thèse : "Classification à base de modèles de mélanges topologiques des données catégorielles et continues". Soutenue le : 4 décembre 2009 (Très Honorable). Jury : F. Alexandre, K. Benabdeslem, Y. Bennani, M. Lebbah, B. Denby, C. Recanati. Rapporteurs : D. Bouchaffra, M. Nadif.

  - **M. Firas Abou Latif**, LITIS, INSA de Rouen. Sujet de thèse : "Identification du profil des utilisateurs d'un hypermédia encyclopédique à l'aide de classifieurs basés sur des dissimilarités". Soutenue le : 8 juillet 2011. Jury : N. Delestre, M. Lebbah, J-P Pécuchet. Rapporteurs : Y. Bourda, J-M Labat.

1.7.3 Responsabilités administratives et pédagogiques

1.7.3.1 Comité de sélection de MCF

2009-2010 Membre de comité de sélection MCF à l'université de Paris 13 et à l'université de Paris 8
2010 Membre externe du comité de sélection MCF à Univ. Paris 5
2009 Membre externe du comité de sélection MCF à Univ. Lyon 1

1.7.3.2 Responsabilité à l'IUT

- **Directeur des études** du semestre 2. Ceci se traduit par le suivi des étudiants, la mise en place de l'emploi du temps, organisation des réunions pédagogiques...etc, de **2011 à 2012**.

- **Responsable Informatique à l'IUT** : Ceci se traduit par la répartition des services d'enseignement, recrutement de vacataires, de **2008 à 2010**.

- **Coordonnateur** du serveur de note au département d'informatique à l'IUT. Ceci se traduit par la gestion informatique du serveur de notes (création de formations, création de comptes), **depuis 2007**.

- **Responsable** de la matière Base de données (1^{re} année) à l'IUT (en semestre 1 et 2), de **2007-2010**.

- **Responsable** de la matière Réseaux (2^{e} année en semestre 4 décalé), de **2008 à 2009**.

Chapitre 2

Descriptif du parcours de recherche

Je suis maître de conférence nommé à l'université de Paris 13 en 2005. Actuellement, je suis membre permanent au LIPN dans l'équipe A3 (Apprentissage Artificiel & Applications). De 09/2005 à 09/2008, j'étais membre permanent du LIM&BIO (Laboratoire d'Informatique Médicale&Bio-Informatique, EA3969, Université de Paris 13). La nature et l'évolution de mes recherches en apprentissage non supervisé sont le résultat d'une influence de l'environnement de recherche. J'étais confronté à différentes applications manipulant des données complexes : chez l'OREAL R&D, dans le cadre du stage de master 2, j'ai eu l'occasion de manipuler des molécules représentées par des vecteurs de dimension 1000, codées en binaires, chez Renault R&D, c'était des données multidimensionnelles de nature catégorielle et continue, captées par le lidar ou le radar installé sur le véhicule. Par la suite j'ai effectué un post-doc dans le domaine de l'environnement où j'ai manipulé des données environnementales. J'ai toujours dépassé le cadre complexe des applications en développant des modèles théoriques robustes. Mon thème de recherche principal concerne les systèmes d'apprentissage statistique pour la fouille de grands volumes de données complexes.

2.1 Avant mon recrutement à l'université de Paris 13

Thèse (Contrat CIFRE)

Après un DEA d'Informatique en intelligence artificielle à l'université de Paris 13, j'ai réalisé ma thèse dans le cadre d'un contrat CIFRE entre l'université de Versailles Saint-Quentin en-Yvelines et RENAULT-R&D. J'ai travaillé durant 3 années sous la direction de Mme Sylvie Thiria (Prof-Univ UVSQ), M. Fouad Badran (Prof CNAM), et M. Christian Chabanon (Ingénieur de recherche chez RENAULT-R&D). Le travail de recherche a été guidé par deux contraintes : la première industrielle demandait d'aboutir à une prestation complète sur véhicule et la seconde, académique, demandait de réaliser un travail de recherche théorique. La partie théorique de ma thèse est à situer à la frontière de l'informatique et de l'intelligence artificielle : il s'agit de modélisation neuronale. Il concerne le nouveau domaine du data mining qui résout des problèmes comme l'extraction, à partir de gros volume de données, d'informations ou de connaissances originales, auparavant inconnues, potentiellement utiles.

Mon travail de recherche concernait le traitement des données qualitatives et catégorielles. Il s'agit de variables, dites discrètes, ne pouvant prendre par nature qu'un nombre restreint

de valeurs. Les calculs dans un espace discret diffèrent des calculs sur l'espace des réels et demandent des approches spécifiques. Mon travail de thèse a consisté à proposer des approches neuronales (cartes topologiques) dédiées aux données binaires et d'une manière globale aux données catégorielles ou qualitatives, [CI-20, CN-16, RN-3, CI-18, CI-19]. J'ai validé mes modèles chez RENAULT R&D sur un problème réel lié à la classification et la reconnaissance du trafic routier. Le travail a été réalisé dans le cadre du développement du système d'aide à la conduite, les données étaient issues de capteur du type radar et lidar [BR-2].

Post-Doc au Centre d'étude des Environnements Terrestre et Planétaires

A la fin de ma thèse, j'ai effectué un Post-doc au CETP (Centre d'étude des Environnements Terrestre et Planétaires) qui fait partie des 7 laboratoires regroupés dans le cadre de l'Institut Pierre-Simon Laplace (IPSL). Le CETP souhaitait écrire le code adjoint du modèle numérique de la météo nationale nommé ISBA qui décrit les interactions entre le sol, la végétation et l'atmosphère. J'ai réalisé dans ce cadre, de Post-Doc, une recherche sur l'application de l'assimilation variationnelle basée sur les méthodes adjointes et l'approche de programmation modulaire actuellement développée par l'équipe "Modélisation et Méthodes Statistiques Avancées" du LOCEAN dans le cadre du développement du logiciel YAO, http://www.locean-ipsl.upmc.fr/~yao.

J'ai été amené à reformuler le modèle direct ISBA afin de pouvoir le traiter par le logiciel YAO. D'autre part, étant donné que cette application était l'une des premières application de YAO, j'ai été amené à travailler avec l'équipe de Modélisation et Méthode Statistique du LOCEAN et j'ai apporté des contributions au développement de YAO, notamment en développant une interface homme-machine qui permet à l'utilisateur de spécifier d'une manière modulaire son modèle.

Je considère cette période comme une mobilité géographique et thématique. Toujours est-il que ce changement d'environnement (d'un monde industriel-R&D au monde laboratoire de recherche) m'a amené a être entouré de chercheurs en physique spécialistes des phénomènes météo et à travailler sur des problématiques très intéressantes liées à l'environnement.

2.2 Travaux de recherche à l'université de Paris 13

2.2.1 Au sein du LIM&BIO (2005/2007)

Durant ma première année au sein du LIM&BIO (Laboratoire d'Informatique Médicale et Bioinformatique) de l'Université de Paris 13, j'ai participé aux activités de recherche du laboratoire. Plus précisément, mes recherches s'inscrivaient dans le cadre général des modèles d'apprentissage pour l'analyse de données médicales. Il s'agissait d'étendre mes travaux de recherche de la thèse pour traiter des données médicales.

En collaboration avec un enseignant chercheur (Jean-Daniel Zucker (DR IRD)), un doctorant du LIM&BIO, (Ramzi Temanni) et un médecin (Christine Poitou) de l'Hôpital Hôtel-Dieu, nous avons développé un modèle de prédiction à partir de données médicales liées à l'obésité. L'objectif est de prendre en compte les différentes données biologiques et cliniques (données mixtes) pour prédire la perte de poids à 3 mois et à 6 mois avant de subir une opération chirurgicale. Afin d'aboutir à un système d'aide à la décision pour la perte de poids, une combinaison des cartes topologiques et le modèle SVM a été développée et testée au sein

de l'équipe [CN-13, CN-11, CI-16].

Enfin, travailler avec Jean-Daniel Zucker, fut une grande chance pour moi. Jean-Daniel m'a fait découvrir d'autres thématiques de recherche, en particulier la bioinformatique. J'ai vu comment faire avancer des travaux communs à des personnes de compétences très diverses (biologistes, informaticiens, médecins, ...).

2.2.2 Au sein de l'équipe A3 du LIPN (depuis 2007)

J'effectue ma recherche au sein de l'équipe A3. Les thèmes de l'équipe s'articule autour de 5 axes de recherche : (1) Aspects théoriques et algorithmiques de l'apprentissage supervisé, (2) Apprentissage pour et par l'action, (3) Apprentissage collaboratif et incrémental non supervisé, (4) Apprentissage de modèles de mélanges, (5) Analyse exploratoire de données complexes. Les applications liées à ma recherche sont motivés par mes collaborations externes et internes à l'université, notamment via les projets ANR, BQR et CIFRE. La nature de mes recherches en apprentissage statistique s'articule autour de la fouille de données complexes en utilisant des approches topologiques robustes. Mes recherches sont aussi le résultat de l'influence du contexte de l'équipe, en particulier la composante numérique, des projet en cours et des différentes rencontres. Ces recherches m'ont amené à participer et à encadrer des stages de master 2 et des thèses sur l'apprentissage numérique. J'ai toujours essayé d'avoir des constantes dans ma recherche qui sont celles de développer des modèles non supervisés en tenant compte de la structure spatiale et de la nature des données. Mes travaux sont en adéquation avec l'augmentation et la difficulté des bases de données actuelles (graphes, données mixtes, séquences, RFID...). Elle sont aussi en phase avec les besoins des chercheurs en apprentissage, en fouille de données et des industriels.

Classification topologique à base de modèles de mélanges

Je vise dans ce thème le développement de méthodes de classification automatique à base de modèles de mélanges où l'espace des données et des variables est pris en compte. Je porte un intérêt particulier aux modèles d'apprentissage qui consistent à découvrir un concept sous une forme géométrique et topologique. Des systèmes d'apprentissage à base de modèles de mélanges sont proposés en collaboration avec Mlle Nicoleta Rogovschi[1], et M. Younès Bennani, pour : -le traitement de données binaires [RI-5, CI-13, CN-12] -le traitement de données mixtes binaires et continues [CI-10, CN-8]. En effet, dans le but de donner une interprétation probabiliste des cartes auto-organisatrices, j'ai développé des modèles probabilistes des cartes topologiques adaptées à la nature des données. Chaque cellule de la carte est associée à une distribution de Bernoulli et/ou une distribution gaussienne selon la nature des données (binaires, mixtes : binaires et continues). L'apprentissage dans ces modèles a pour objectif d'estimer la fonction densité sous forme d'un mélange de densités élémentaires. Chaque densité élémentaire est elle aussi un mélange de lois définies sur un voisinage. Ces nouveaux algorithmes d'apprentissage probabiliste et non supervisé utilisent l'algorithme EM pour maximiser la vraisemblance des données afin d'estimer les paramètres du modèle de mélange. Ces algorithmes ont une portée pratique aussi bien en classification qu'en visualisation.

A la suite de la thèse (de Mlle Nicoleta Rogovschi), j'ai orienté la suite de ce thème vers

[1] devenue depuis septembre 2010 MCF à l'université de Paris 5

l'apprentissage à partir de données séquentielles (non i.i.d[2]). En effet, je pense que c'est un sujet de recherche intéressant qui pose à la fois des questions théoriques et pratiques lorsque la topologie des données est prise en compte. D'un point de vue algorithmique, je souhaite que mes recherches s'articulent autour des modèles des mélanges topologiques. Pour cette problématique, je suis par ailleurs guidé par un projet CIFRE avec l'INA (Institut national de l'audiovisuel). Une thèse a démarré fin 2009 pour étendre ces travaux sur les données structurées en séquences (Doctorante : Mlle Rakia Jaziri) [CI-11, CN-2, CI-1, CN-1, CI-2]. L'idée principale de ce travail est de définir des modèles de Markov cachés topologiques et auto-organisés, qui s'adaptent à la nature, la structure et la dynamique des séquences. Ceci paraît très pertinent puisque l'information véhiculée par les séquences (structure, ordre, nature) sera prise en compte à l'intérieur des modèles. Dans la partie synthèse scientifique, je reviens plus en détail sur quelques avancés réalisées.

Combinaison de modèles et visualisation de grandes masses de données

L'équipe A3 travaille, dans le cadre du pôle de compétitivité Cap Digital - IMVN, sur des problèmes liés à la discrimination de données structurées. Par ailleurs, certains de mes travaux sont guidés par les projets ANR, dont la combinaison de modèles. La combinaison de modèles consiste à fusionner plusieurs modèles ou résultats d'apprentissage en se basant sur l'hypothèse que deux (ou plus de) résultats ou modèles valent mieux qu'un. L'objectif du développement des systèmes hybrides est de tirer parti des points forts de chaque paradigme utilisé afin d'obtenir de meilleures performances pour la résolution des problèmes. Les modèles combinés visés couvrent un vaste spectre de modèles : probabilistes, les réseaux connexionnistes, les méthodes statistiques traditionnelles. Une approche a été, récemment, brevetée avec notre partenaire THALES dans le cas non supervisé pour la fusion/combinaison/consensus du partitionnement utilisant l'analyse relationnelle [BR-1, CI-9]. L'intérêt de l'approche proposée est de pouvoir projeter les données à très forte dimensionnalité dans un espace de faible dimensionnalité, l'espace de partitionnement, sans aucune information a priori sur le nombre de classes. L'aspect novateur de la démarche proposée consiste à transformer le problème de fusion en un problème d'optimisation. L'approche proposée a permis d'associer plusieurs types de visualisations à plusieurs niveaux : une visualisation globale grâce à l'analyse relationnelle, puis une visualisation plus fine à l'aide des cartes auto-organisatrices.

D'autres approches ont été proposés dans le cadre du projet ANR E-fraud Box (section 1.5), qui concerne le traitement et l'investigation de la fraude à la carte bancaire sur internet en faisant collaborer différentes techniques de classification et d'analyse de réseaux sociaux. Les données traitées sont essentiellement déséquilibrées et transactionnelles. Récemment, j'ai proposé, une méthode de sous-échantillonnage adaptatif pour traiter ce type de bases déséquilibrées. Le processus procède par le sous-échantillonnage des données majoritaires, guidé par les données minoritaires tout au long de la phase d'un apprentissage semi-supervisé. Nous utilisons comme modèle d'apprentissage les cartes auto-organisatrices [CN-5, CI-3].

Au gré des rencontres, particulièrement avec Mme H. Azzag qui est maître de conférences dans mon équipe, mes centres d'intérêt ont évolué. J'ai découvert avec elle d'autres approches bioinspirées et peu classiques en apprentissage automatique. Nous avons pu combiner et même unifier avec succès deux modèles d'apprentissages. Le premier modèle est celui des cartes topologiques et le deuxième est l'algorithme de classification hiérarchique bioinspirée

[2] independently and identically distributed

AntTree. Ce dernier est inspiré du comportement des règles d'accrochage des fourmis réelles. Le modèle vise à développer un modèle de classification simultanée : topologique et hiérarchique, d'une manière à produire une partition des données organisées sur une grille 2D et en même temps une organisation hiérarchique sous forme d'arbre au niveau de chaque cellule de la grille [CI-8, CI-12, CN-10]. Je continue à travailler dans cette perspective à l'extension de ce formalisme à la classification et la visualisation de données structurées sous forme de graphe [CN-6]. En assimilant la donnée traditionnelle à un nœud d'un graphe, cette approche peut rejoindre les autres méthodes consistant à décomposer le graphe de départ (i.e. graphe initial) en une succession de sous-graphes. Je m'intéresse plus particulièrement au formalisme d'apprentissage non supervisé. Cet axe de recherche est très passionnant car il nous permet de s'inspirer des modèles biologiques et de proposer d'autres algorithmes robustes.

En plus des thématiques de recherche abordées, d'autres problématiques commencent dans notre équipe, et ceci guideés par un autre projet CIFRE. Dans ce cadre avec M. Amine Chaibi, j'ai commencé à m'intéresser aux problématiques liées à la classification croisée (co-clustering). C'est un sujet ouvert et passionnant et il existe différents travaux assez avancés. Ceci nous oblige et nous pousse à étudier des problématiques non résolues ou partiellement résolues.

2.3 Synthèse des différentes collaborations

Les premiers collaborateurs sont évidemment Mme. Sylvie Thiria et M. Fouad Badran qui m'ont initié à la recherche, et que j'ai eu le plaisir d'avoir comme encadrants, du DEA (Master 2) et de ma thèse. J'ai eu l'occasion de collaborer avec eux depuis ma thèse [CI-20, CN-16, RN-3, CI-18, CI-19, BR-2, CN-16, CI-15, CI-17, CHI-1, CHI-2, RI-6]. C'est toujours avec plaisir que je reçois les courriels de leurs doctorants ou stagiaires en cours. Les discussions que j'ai avec Sylvie et Fouad lorsque je passe à leur laboratoire sont toujours très intéressantes scientifiquement. Leurs travaux de recherche sont toujours guidés par des applications passionnantes liée particulièrement à l'environnement.

Jean-Daniel Zucker est une personne auprès de qui j'ai également appris beaucoup de choses. Je continue à avoir de ses nouvelles et à avoir des discussions lorsque nous nous rencontrons. A travers lui, j'ai rencontré des médecins et des bioinformaticiens de grande qualité. J'ai en outre beaucoup de plaisir à travailler avec les doctorants de son équipe, en particulier Ramzi Temmani [CN-13, CN-11, CI-16]. Faute de temps et vu la mobilité interne que j'ai effectuée à Paris 13 en passant du laboratoire LIM&BIO au laboratoire LIPN, j'ai abandonné cet axe, mais j'espère un jour revenir à ce type de problématiques qui intéressent les médecins. Par contre, je suis persuadé que ceci ne peut se faire sans l'aide des médecins ou bioinformaticiens. L'expert du domaine est primordial.

Bien évidemment, j'entretiens des collaborations étroites avec les membres de notre équipe A3. Une grande partie de mes travaux de recherche sont en collaboration avec M. Younès Bennani, que ce soit sur les modèles de mélanges pour données binaires [CI-9, CI-10, CI-13, CN-5] ou récemment sur les données séquentielles [CI-1, CI-2, CI-11, CN-2, CN-8, CN-12, RI-5]. J'apprécie spécialement sa rigueur et sa franchise. Et aussi son implication dès qu'il s'agit de discuter de problématique de recherche. Par ailleurs j'ai partagé des collaborations avec M. Nistor Grosavu et Mlle Nicoleta Rogovschi au moment où ils étaient doctorants et actuellement en tant que maîtres de conférences [CI-7, CN-7, CI-11, CN-9, CHN-1, CHN-2, CHI-2, RI-2].

Recemment, j'ai initié avec Mme Hanane Azzag (MCF à Univ-Paris 13) des travaux de recherche intéressants [CN-4, CN-10, CN-6, CI-5, CI-12, RI-1]. Hanane m'a permis de m'ouvrir et de me familiariser à d'autres approches d'apprentissage. Le bio-inspiré est un domaine de recherche intéressant et prometteur. J'apprécie également chez Hanane sa sincérité et sa disponibilité. Je me dois aussi de citer M. Khalid Benabdeslem (MCF à Lyon 1), avec qui j'ai gardé une collaboration scientifique efficace [CN-15, CN-14, RI-3]. Mes dernières collaborations concernent les doctorants que je co-encadre, je me doit de citer Mlle Rakia Jaziri [CN-2, CI-1, CI-2], M. Nhat-Quang Doan et M. Amine Chaibi.

J'ai par ailleurs la chance d'être impliqué dans différents projets avec des industriels ou des institutions. Je me dois ici de les remercier tous pour les discussions intéressantes et la confiance qu'ils ont eue pour les travaux de notre équipe.

Avec mes activités d'animation au sein des groupes de travail, j'ai eu l'occasion de collaborer avec des personnalités différentes. Je m'abstiens de citer certains au risque d'oublier d'autres. Je me dois de les remercier ici pour le partage et pour leur disponibilités dans l'organisation des événements de partage scientifique (colloque, ateliers, réunions...). Cette collaboration fut un véritable plaisir.

Chapitre 3

Publications

Ce chapitre dresse une liste complète de mes publications, communications et brevets. Je souhaite signaler que la plupart de mes travaux acceptés à des conférences internationales sont également présentés dans leur majorité à la conférence francophone Extraction et Gestion des Connaissances (EGC). Une partie de mes travaux sont aussi présentés à la Conférence francophone d'apprentissage (CAp). Ceci est ma volonté et mon souhait de partager mes travaux avec la communauté francophone.

3.1 Brevet (2)

[BR-1] Hamid Benhadda, Younès Bennani, Mustapha Lebbah, and Nistor Grozavu. System for searching visual information. WO/2010/066774 - PCT/EP2009/066702. In *European Patent Office (EPO)*. Brevet (THALES, Université de Paris 13), 2009.

[BR-2] Mustapha Lebbah. Procédé de reconnaissance du trafic routier. In *INPI*, number FR2822576. brevet, 2000.

3.2 Chapitres de livres d'audience internationale (2)

[CHN-1] Nistor Grozavu, Younès Bennani, and Mustapha Lebbah. Cluster-dependent feature selection through a weighted learning paradigm. In Fabrice Guillet, Gilbert Ritschard, Djamel Zighed, and Henri Briand, editors, *Advances in Knowledge Discovery and Management*, volume 292 of *Studies in Computational Intelligence*, pages 133–147. Springer Berlin / Heidelberg, 2010.

[CHN-2] Mustapha Lebbah, Younès Bennani, Hamid Benhadda, and Nistor Grozavu. Relational analysis for clustering consensus. In *Invited Book Chapter Machine Learning*, pages 45–60. IN-TECH, 2009.

3.3 Chapitres de livres d'audience nationale (2)

[CHI-1] Fouad Badran, Mustapha Lebbah, and Sylvie Thiria. Cartes auto-organisatrices et classification automatique. livre apprentissage statistique. In *Livre Apprentissage statistique*. Eyrolles, 2008.

[CHI-2] Mustapha Lebbah, Sylvie Thiria, and Fouad Badran. Les perceptrons multicouches. In *Livre Apprentissage statistique*. Hermès, 2006.

3.4 Revues internationales avec comité de sélection (6)

[RI-1] Hanane Azzag and Mustapha Lebbah. Self-organizing tree using artificial ants. *Special Issue on Applications of Nature Inspired Computing. Journal of Information Technology Research*, 4(2) :1–16, 2011.

[RI-2] Nicoleta Rogovschi, Mustapha Lebbah, and Younès Bennani. A self-organizing map for mixed continuous and categorical data. *in IJC, International Journal of Computing*, 10(1), 2011.

[RI-3] Mustapha Lebbah and Khalid Benabdeslem. Visualization and clustering of categorical data with probabilistic self-organizing map. *Neural Computing and Applications*, 19(3) :393–404, 2010.

[RI-4] Nicoleta Rogovschi, Mustapha Lebbah, and Younès Bennani. Learning self-organizing mixture markov models. *Journal of Nonlinear Systems and Applications. ISSN 1918-3704.*, 1(1-2) :63–71, 2010.

[RI-5] Mustapha Lebbah, Younès Bennani, and Nicoleta Rogovschi. A probabilistic self-organizing map for binary data topographic clustering. *International Journal of Computational Intelligence and Applications*, 7(4) :363–383, 2008.

[RI-6] Martin Saraceno, Christine Provost, and Mustapha Lebbah. Biophysical regions identification using an artificial neuronal network : A case study in the south western atlantic. *Advances in Space Research*, 37(4) :793 – 805, 2006. Natural Hazards and Oceanographic Processes from Satellite Data.

3.5 Revues nationales avec comité de sélection (3)

[RN-1] Nicoleta Rogovschi, Mustapha Lebbah, and Younès Bennani. Modèles de mélanges topologiques pour la classification de données catégorielles et mixtes. *Numéro spécial : Apprentissage Artificiel et Fouille de Données*, RNTI-E-21, 2011.

[RN-2] Mustapha Lebbah, Mohamed Ramzi Temanni, Christine Poitou-Bernert, Karine Clément, and Jean-Daniel Zucker. Partitionnement des données pour les problèmes de classement difficiles : Combinaison des cartes topologiques mixtes et svm. *Numéro spécial : Apprentissage Artificiel et Fouille de Données*, RNTI-A-2, 2007.

[RN-3] Mustapha Lebbah, Sylvie Thiria, and Fouad Badran. Visualisation et classification avec les cartes topologiques catégorielles. *Numéro spécial sur la fouille de données complexes*, RNTI-E-4, 2005.

3.6 Conférences internationales avec comité de sélection (20)

[CI-1] Rakia Jaziri, Mustapha Lebbah, and Nicoleta Rogovschi Younès Bennani. Probabilistic self-organizing maps for multivariate sequences. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN 2011,San Jose, California July 31 - August 5, 2011*, pages 851–858. IEEE, 2011.

[CI-2] Rakia Jaziri, Mustapha Lebbah, Younès Bennani, and Jean H Chenot. Sos-hmm : Self-organizing structure of hidden markov model. In *Artificial Neural Networks - ICANN 2011, International Conference, June 14-17th, 2011, Espoo, Finland, Proceedings*, Lecture Notes in Computer Science. Springer, 2011.

[CI-3] Fatma Hamdi, Mustapha Lebbah, and Younès Bennani. Topographic under-sampling for unbalanced distributions. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN 2010, IEEE World Congress on Computational Intelligence., Barcelona , Spain, 18-23 July 2010*, pages 1–6. IEEE, 2010.

[CI-4] Hanene Azzag, Mustapha Lebbah, and Aymen Arfaoui. Map-treemaps : A new approach for hierarchical and topological clustering. In Sorin Draghici, Taghi M. Khoshgoftaar, Vasile Palade, Witold Pedrycz, M. Arif Wani, and Xingquan Zhu, editors, *The Ninth International Conference on Machine Learning and Applications, ICMLA 2010, Washington, DC, USA, 12-14 December 2010*, pages 873–878. IEEE Computer Society, 2010.

[CI-5] Mustapha Lebbah and Hanane Azzag. Topological hierarchical tree using artificial ants. In Kok Wai Wong, B. Sumudu U. Mendis, and Abdesselam Bouzerdoum, editors, *Neural Information Processing. Theory and Algorithms - 17th International Conference, ICONIP 2010, Sydney, Australia, November 22-25, 2010, Proceedings, Part I*, volume 6443 of *Lecture Notes in Computer Science*, pages 652–659. Springer, 2010.

[CI-6] Hanane Azzag and Mustapha Lebbah. A new method for topological and hierarchical clustering. In *34th Annual Conference of the German Classification Society (GfKl) July 21 -23*, 2010.

[CI-7] Nistor Grozavu, Younès Bennani, and Mustapha Lebbah. From variable weighting to cluster characterization in topographic unsupervised learning. In *International Joint Conference on Neural Networks, IJCNN 2009, Atlanta, Georgia, USA, 14-19 June 2009*, pages 1005–1010. IEEE, 2009.

[CI-8] Hanane Azzag and Mustapha Lebbah. A new approach for auto-organizing a groups of artificial ants. In Istváan Karsai George Kampis and Eors Szathmáry, editors, *ECAL'09*, Lecture Notes in Computer Science, pages 434–640. Springer Berlin / Heidelberg, 2009.

[CI-9] Mustapha Lebbah, Younès Bennani, and Hamid Benhadda. Relational analysis for consensus clustering from multiple partitions. In M. Arif Wani, Xue wen Chen, David Casasent, Lukasz A. Kurgan, Tony Hu, and Khalid Hafeez, editors, *Seventh International Conference on Machine Learning and Applications, ICMLA 2008, San Diego, California, USA, 11-13 December 2008*, pages 218–223. IEEE Computer Society, 2008.

[CI-10] Nicoleta Rogovschi, Mustapha Lebbah, and Younès Bennani. Probabilistic mixed topological map for categorical and continuous data. In M. Arif Wani, Xue wen Chen, David Casasent, Lukasz A. Kurgan, Tony Hu, and Khalid Hafeez, editors, *Seventh International Conference on Machine Learning and Applications, ICMLA 2008, San Diego, California, USA, 11-13 December 2008*, pages 224–231. IEEE Computer Society, 2008.

[CI-11] Mustapha Lebbah, Younés Bennani, and Nicoleta Rogovschi. Learning self-organizing maps as a mixture markov models. In *Proceedings of The 3rd International Conference on Complex Systems and Applications*, ICCSA'09, pages 54–59, Le Havre, Normandy, France, June 29 - July 02, 2009, 2009.

[CI-12] Hanane Azzag and Mustapha Lebbah. Clustering of self-organizing map. In *ESANN 2008, 16th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 23-25, 2008, Proceedings*, pages 209–214, 2008.

[CI-13] Mustapha Lebbah, Nicoleta Rogovschi, and Younès Bennani. Besom : Bernoulli on self-organizing map. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN 2007, Celebrating 20 years of neural networks, Orlando, Florida, USA, August 12-17, 2007*, pages 631–636. IEEE, 2007.

[CI-14] Khalid Benabdeslem and Mustapha Lebbah. Feature selection for self-organizing map. In *Information Technology Interfaces, 2007. ITI 2007. 29th International Conference on*, pages 45–50, 2007.

[CI-15] Arnaud Quesney, Eric Jeansou, Christian Ruiz, Nathalie Steunou, Bruno Cugny, Nicolas Picot, Jean-Claude Souyris, Sylvie Thiria, and Mustapha Lebbah. Unsupervised classification of altimetric waveform over all surface type. In *Ocean Surface Topography Science Team Meeting*, OSTST 2007, (Poster), 2007.

[CI-16] Mohamed Ramzi Temanni, Mustapha Lebbah, Christine Poitou-Bernert, Karine Clément, and Jean-Daniel Zucker. Combining mixed topological maps and svm to improve accuracy of hard classification problems : application to the biomedical data. In *Integrative Post-Genomics*, IPG'06, (Poster), 2006.

[CI-17] Mustapha Lebbah, Aymeric Chazottes, Fouad Badran, and Sylvie Thiria. Mixed topological map. In *ESANN 2005, 13th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 27-29, 2005, Proceedings*, pages 357–362, 2005.

[CI-18] Mustapha Lebbah, Fouad Badran, and Sylvie Thiria. Visualization and classification with categorical topological map. In *ESANN 2004, 12th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 28-30, 2004, Proceedings*, pages 459–464, 2004.

[CI-19] Mustapha Lebbah, Christian Chabanon, Fouad Badran, and Sylvie Thiria. Categorical topological map. In José R. Dorronsoro, editor, *Artificial Neural Networks - ICANN 2002, International Conference, Madrid, Spain, August 28-30, 2002, Proceedings*, volume 2415 of *Lecture Notes in Computer Science*, pages 890–895. Springer, 2002.

[CI-20] Mustapha Lebbah, Fouad Badran, and Sylvie Thiria. Topological map for binary data. In *ESANN 2000, 8th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 26-28, 2000, Proceedings*, pages 267–272, 2000.

3.7 Conférences nationales avec comité de sélection (15)

[CN-1] Rakia Jaziri, Mustapha Lebbah, Younès Bennani, and Jean-Hugues Chenot. Apprentissage non supervisé des structures des hmms. In *in Proc. SFDS, 43éme Journées de Statistiques, Gammarth, Tunisie, 23-27 Mai 2011*, 2011.

[CN-2] Rakia Jaziri, Mustapha Lebbah, Younès Bennani, and Jean-Hugues Chenot. Structuration automatique des flux télévisuels par apprentissage non supervisé des répétitions. In Ali Khenchaf and Pascal Poncelet, editors, *Extraction et gestion des connaissances (EGC'2011), Actes, 25 au 29 janvier 2011, Brest, France*, volume RNTI-E-20 of *Revue des Nouvelles Technologies de l'Information*, pages 311–312. Hermann-Éditions, 2011.

[CN-3] Nicoleta Rogovschi, Mustapha Lebbah, and Nistor Grozavu. Pondération et classification simultanée de données binaires et continues. In *Extraction et gestion des connaissances (EGC'2011), Actes, 25 au 29 janvier 2011, Brest, France*, pages 65–70, 2011.

[CN-4] Hanane Azzag and Mustapha Lebbah. Une nouvelle approche visuelle pour la classification hiérarchique et topologique. In *Extraction et gestion des connaissances (EGC'2011), Actes, 25 au 29 janvier 2011, Brest, France*, pages 677–688, 2011.

[CN-5] Mustapha Lebbah and Younès Bennani. Sous-échantillonnage topographique par apprentissage semi-supervisé. In *Extraction et gestion des connaissances (EGC'2010), Actes, 26 au 29 janvier 2010, Hammamet, Tunisie*, pages 121–126, 2010.

[CN-6] Hanane Azzag and Mustapha Lebbah. Auto-organisation topologique et hiérarchique des données. In *Extraction et gestion des connaissances (EGC'2010), Actes, 26 au 29 janvier 2010, Hammamet, Tunisie*, pages 555–560, 2010.

[CN-7] Nistor Grozavu, Younès Bennani, and Mustapha Lebbah. Caractérisation automatique des classes découvertes en classification non supervisée. In *Extraction et gestion des connaissances (EGC'2009), Actes, Strasbourg, 27 au 30 janvier 2009*, volume RNTI-E-15 of *Revue des Nouvelles Technologies de l'Information*, pages 43–54, 2009.

[CN-8] Nicoleta Rogovschi, Mustapha Lebbah, and Younés Bennani. Un algorithme pour la classification topographique simultanée de données qualitatives et quantitativesp. In *Cap'2009 : Conférence francophone sur l'apprentissage automatique. Plate-forme AFIA. 25-29 Mai, Hammamat-Tunisie*, 2009.

[CN-9] Nistor Grozavu, Younès Bennani, and Mustapha Lebbah. Pondération locale des variables en apprentissage numérique non-supervisé. In *Extraction et gestion des connaissances (EGC'2008), Actes des 8èmes journées Extraction et Gestion des Connaissances, Sophia-Antipolis, France, 29 janvier au 1er février 2008, 2 Volumes*, volume RNTI-E-11 of *Revue des Nouvelles Technologies de l'Information*, pages 321–330, 2008.

[CN-10] Mustapha Lebbah and Hanane Azzag. Segmentation hiérarchique des cartes topologiques. In *Extraction et gestion des connaissances (EGC'2008), Actes des 8èmes journées Extraction et Gestion des Connaissances, Sophia-Antipolis, France, 29 janvier au 1er février 2008, 2 Volumes*, volume RNTI-E-11 of *Revue des Nouvelles Technologies de l'Information*, pages 631–642, 2008.

[CN-11] Mustapha Lebbah, Mohamed Ramzi Temanni, Christine Poitou-Bernert, Karine Clément, and Jean-Daniel Zucker. Partitionnement des données pour les problèmes de classement difficiles : Combinaison des cartes topologiques mixtes et svm. *Numéro spécial : Apprentissage Artificiel et Fouille de Données*, RNTI-A-2, 2007.

[CN-12] Mustapha Lebbah, Nicoleta Rogovschi, and Younés Bennani. Besom : Bernoulli on self organizing map. In *Cap'2007 : conférence francophone sur l'apprentissage automatique, Plate-forme AFIA*, 2007.

[CN-13] Mohamed Ramzi Temanni, Mustapha Lebbah, Christine Poitou-Bernert, Karine Clément, and Jean-Daniel Zucker. Combinaison des cartes topologiques mixtes et des machines à vecteurs de support : une application pour la prédiction de perte de poids chez les obèses. In *Extraction et gestion des connaissances (EGC'2007), Actes des cinquièmes journées Extraction et Gestion des Connaissances, Namur, Belgique, 23-26 janvier 2007, 2 Volumes*, volume RNTI-E-9 of *Revue des Nouvelles Technologies de l'Information*, pages 33–44. Cépaduès-Éditions, 2007.

[CN-14] Khalid Benabdeslem, Mustapha Lebbah, Alexandre Aussem, and Marilys Corbex. Approche connexionniste pour l'extraction de profils cas-témoins du cancer du nasopharynx à partir des données issues d'une étude épidémiologique. In *Extraction et gestion des connaissances (EGC'2007), Actes des cinquièmes journées Extraction et Gestion des Connaissances, Namur, Belgique, 23-26 janvier 2007, 2 Volumes*, volume RNTI-E-9 of *Revue des Nouvelles Technologies de l'Information*, pages 445–454. Cépaduès-Éditions, 2007.

[CN-15] Khalid Benabdeslem, Mustapha Lebbah, Alexandre Aussem, Marilys Corbex, and CHELGHOUM N. Learning based system for knowledge discovery from nasopharyngeal cancer data. In *Colloque sur l'Optimisation et les Systèmes d'Information COSI 2007. 11-13 juin 2007 - Oran Algérie*, 2007.

[CN-16] Mustapha Lebbah, Fouad Badran, and Sylvie Thiria. Carte topologique et données binaires. In *32 èmes Journées de la société française des statistiques, SFds. Mai 16 au 25/ 2000*, 2000.

3.8 Colloques (3)

[COL-1] Julien Brajard, Cédric Duboudin, Hanane Bénaribi, Mustapha Lebbah, and Sylvie Thiria. Typologie des logements et lien avec la multipollution. In *Colloque "Comment concilier énergie, qualité de l'air intérieur et santé". Pendant le salon Pollutec. 4 Dec 2008. http ://www.cstb.fr/actualites/webzine/editions/editionde-decembre-2008/colloque-comment-concilier-energie-qualite-de-lair-interieuret-sante.html*, 2008.

[COL-2] Younès Bennani, Mustapha Lebbah, Nistor Grozavu, Marcotorchino J.F., Hamid Benhadda, and Stephane Lorin. Analyse relationnelle comme algorithme de fusion de partitionnement. In *Workshop Infomagic, Analyse multimodale de l'information, Pôle de compétitivité Cap digital,10 juin 2008 Telecom-ParisTec*, 2008.

[COL-3] Mustapha Lebbah, Sylvie Thiria, and Fouad Badran. Visualisation avec les cartes topologiques catégorielle. In *ATELIERS Fouille de données complexes dans un processus d'extraction de connaissances, EGC'2004*, 2004.

Deuxième partie

Synthèse scientifique

Chapitre 4

Contexte et contributions

4.1 Cadre des travaux de recherche

Mes travaux de recherche se placent dans le cadre de l'apprentissage automatique. Je m'intéresse essentiellement au problème de l'apprentissage non supervisé dont la problématique consiste à construire des représentations simplifiées de données, pour mettre en évidence les relations existantes entre les caractéristiques relevées sur des données et les ressemblances ou dissemblances de ces dernières, sans avoir aucune connaissance sur les classes. On peut distinguer deux grandes familles : les méthodes probabilistes et les méthodes déterministes ou tout simplement les méthodes de quantification.

L'apprentissage non supervisé rentre dans le cadre de la statistique dite descriptive. Du point de vue statistique, l'apprentissage non supervisé cherche à modéliser la fonction densité $p(\mathbf{x})$ sous-jacente aux données à partir de l'ensemble d'apprentissage $\mathcal{A}$. D'une manière plus précise, la classification automatique cherche à déterminer des régions convexes de l'espace $\mathcal{D}$ qui contiennent les modes de $p(\mathbf{x})$, donc les régions correspondant aux fortes concentrations de données. Les approches hiérarchiques [War63](Classification Hiérarchique Ascendante), l'algorithme des k-means [Mac67], les modèles de mélange [MP00], et les approches spectrales [NJW01] basées sur la théorie des graphes sont des méthodes souvent utilisées.

Deux types d'approches seront présentées : l'approche probabiliste qui suppose explicitement que les données sont générées par des mélanges (ou mixtures) de fonctions densités plus simples, et l'approche par quantification vectorielle qui cherche à déterminer directement une partition de $\mathcal{D}$ (espace des données) et qui repose souvent sur des hypothèses probabilistes implicites.

Les méthodes que je vais présenter dans les chapitres suivants cherchent à déterminer directement une partition de $\mathcal{D}$ en K sous-ensembles, groupes ou clusters. Cette partition et les groupes associés seront notés par $\mathcal{P} = \{P_1, \ldots, P_c, \ldots, P_K\}$. A chaque sous-ensemble P_c, on associe un vecteur référent $\mathbf{w}_c \in \mathcal{D}$ qui sera le représentant ou le "résumé" de l'ensemble des observations de P_c. On notera par la suite $\mathcal{W} = \{\mathbf{w}_c; c = 1, ..., K\}$ l'ensemble des vecteurs référents. La nature de ce référent dépendra de la nature des données traitées. La partition $\mathcal{P}$ de $\mathcal{D}$ peut être définie d'une manière équivalente par une fonction d'affectation χ qui est une application de $\mathcal{D}$ dans un ensemble fini d'indices $\varphi = \{1, 2, ..., K\}$. Si on utilise cette définition le sous-ensemble P_c est alors représenté par $\{\mathbf{x} \in \mathcal{D}/\chi(\mathbf{x}) = c\}$. L'utilisation de cette fonction d'affectation permet d'associer à une observation $\mathbf{x}$ de $\mathcal{D}$ son vecteur référent $\mathbf{w}_{\chi(\mathbf{z})}$.

Mes travaux de recherche présentés dans ce rapport concernent l'apprentissage non supervisé, la combinaison de modèles pour le partitionnement et la visualisation de grandes masses de données complexes. Je m'intéresse plus particulièrement dans mes travaux de recherche aux modèles dits modèles auto-organisés [KKL97, Koh01b]. Plusieurs modèles différents sont présentés dans la littérature avec des architectures variables (hiérarchiques ou hyperboliques) [Fri95, RMD02, OR06], mais tous partagent les mêmes caractéristiques qui sont celles de présenter les résultats du partitionnement, souvent sur une grille de faible dimension (1D, 2D ou 3D). Dans le cadre de l'apprentissage non supervisé tel que défini par Kohonen, les cartes topologiques utilisent un algorithme d'auto-organisation (Self-Organizing Map, SOM) qui permet d'une part de quantifier de grandes quantités de données en regroupant les observations similaires en groupes ou "clusters" et d'autre part de projeter les groupes obtenus de façon non linéaire sur une grille, permettant ainsi de visualiser la structure de données en deux dimensions, tout en respectant la topologie initiale des données.

Il est très difficile de donner une définition à la complexité des données. Je ne pense pas qu'il en existe une, mais, dans mes travaux de recherche, je me suis intéressé au début aux traitements de données catégorielles, binaires, continues et mixtes. Puis par la suite, je me suis intéressé aux données structurées en séquence.

Modèle des cartes auto-organisatrices

Dans ces modèles, l'ensemble des indices φ possède une structure de graphe qu'on appellera par la suite la "carte" et sera notée $\mathcal{C}$. La carte $\mathcal{C}$ se présente sous la forme d'un graphe non orienté dont les sommets représentent des cellules. Il s'agit le plus souvent d'un treillis de faible dimension (grille 1D, 2D ou 3D). Cette structure de graphe permet de définir une topologie sur $\mathcal{C}$, ainsi, pour toute paire de cellules (c,r) de cette "carte", on définit la distance discrète $\delta(c,r)$ sur $\mathcal{C}$ comme étant la longueur du plus court chemin entre c et r sur la carte $\mathcal{C}$.

Pour chaque cellule c, cette distance discrète permet de définir le voisinage. Cette notion de voisinage peut être introduite à l'aide de fonctions noyaux $\mathcal{K}$ qui sont des fonctions numériques, positives et symétriques et telles que $\lim_{|y|\to\infty} \mathcal{K}(y) = 0$. La fonction noyau permet de quantifier l'influence relative entre deux cellules c et r de la carte, cette influence est numériquement représentée par la valeur $\mathcal{K}(\delta(c,r))$. Souvent, dans les algorithmes proposés, on a besoin de faire évoluer la taille du voisinage associé à chaque cellule en le faisant décroître au cours des itérations. Pour cela, il est nécessaire d'utiliser la famille de fonctions paramétrées $\mathcal{K}^T(\delta) = \mathcal{K}(\delta/T)$. Le paramètre T sera dit température pour une fonction noyau $\mathcal{K}^T$.

-Modèle de quantification "déterministe"

Les modèles de quantification cherchent à projeter des observations multidimensionnelles $\mathbf{x}$ ($\mathcal{D} \subset \Re^n$) sur un espace discret de faible dimension (en général 1, 2 ou 3) qui est la carte $\mathcal{C}$. Nous allons noter par la suite $\mathcal{A} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ l'ensemble des observations $\mathbf{x}$ constitué de n composantes $\mathbf{x} = (x^1, ..., x^n)$. La "projection" doit respecter la propriété de "conservation" de la topologie, c'est-à-dire que deux cellules c et r qui sont voisines par rapport à la topologie discrète de la carte, doivent être associés à deux vecteurs référents $\mathbf{w}_c$ et $\mathbf{w}_r$ proches par en terme de distance euclidienne sur $\mathcal{D}$. Dans l'espace euclidien, l'algorithme classique minimise une fonction de coût notée $\mathcal{G}_{som}^T$ définie par l'expression :

$$\mathcal{G}_{som}^{T}(\chi, \mathcal{W}) = \sum_{\mathbf{x}_i \in \mathcal{A}} \sum_{c \in \mathcal{C}} \mathcal{K}^{T}(\delta(c, \chi(\mathbf{x}_i)) \|\mathbf{x}_i - \mathbf{w}_c\|^2 \tag{4.1}$$

L'expression (4.1) représente une famille de fonctions coûts paramétrée par le paramètre T dans laquelle la distance euclidienne d'une observation $\mathbf{x}$ à son référent $\mathbf{w}_{\chi(\mathbf{z})}$ est remplacée par une distance pondérée, notée d_T,

$$d_T(\mathbf{x}_i, \mathbf{w}_{\chi(\mathbf{x}_i)}) = \sum_{c \in \mathcal{C}} \mathcal{K}^{T}(\delta(c, \chi(\mathbf{x}_i)) \|\mathbf{x}_i - \mathbf{w}_c\|^2 \tag{4.2}$$

On observe que la distance entre $\mathbf{x}$ et $\mathbf{w}_{\chi(\mathbf{x})}$ relativement à la distance d_T est une somme pondérée de la distance euclidienne de $\mathbf{x}$ à tous les vecteurs référents $\mathbf{w}_c$ du voisinage d'influence de la cellule $\chi(\mathbf{x})$.

L'algorithme des cartes auto-organisatrices minimise la fonction de coût qui correspond aux k-means ou aux k-moyennes régularisées par un terme qui assure la conservation de la topologie. Dans tous les modèles qui seront présentés par la suite dans le cas de quantification, j'utilise l'approche des nuées dynamiques pour minimiser la fonction de coût. Lorsqu'on fait décroître la température T, l'algorithme des cartes auto-organisatrices tend vers l'algorithme des k-means. Implicitement, l'algorithme de Kohonen peut donc s'interpréter d'une manière probabiliste puisque, lorsque la température T devient suffisamment petite, il cherche à modéliser la densité sous-jacente des observations avec des contraintes et des hypothèses implicites fortes.

-Modèle probabiliste

Le formalisme que je viens de présenter peut être étendu afin de transformer le modèle de cartes auto-organisatrices en modèle probabiliste [ABT97, ABT98, BSI98, MP00]. Il est tout à fait connu, en général, que les méthodes de classification non supervisées cherchent à approximer la densité $p(\mathbf{x})$ des observations. Ainsi, nous savons que l'algorithme de quantification cherche à approximer $p(\mathbf{x})$ sous les hypothèses que $p(\mathbf{x})$ est un mélange de K lois de probabilités particulières [DH73], c'est-à-dire, il est fait l'hypothèse implicite que :

$$p(\mathbf{x}) = \sum_{c=1}^{K} p(c) p(\mathbf{x}/c) \qquad \text{où} \quad \sum_{c=1}^{K} p(c) = 1 \tag{4.3}$$

Lorsqu'on est dans l'espace euclidien et qu'on manipule des données continues, alors on peut considérer $p(\mathbf{x}/c)$ comme suit :

$$p(\mathbf{x}/c) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp\left(-\frac{\|\mathbf{x} - \mathbf{w_c}\|^2}{2\sigma^2}\right) \tag{4.4}$$

Une modélisation plus complexe va permettre de proposer un modèle de cartes topologiques qui conserve les propriétés d'ordre, mais s'interprète maintenant comme un mélange de mélange locales de lois de probabilités permettant d'approximer les densités sous-jacentes des données binaires, catégorielles et mixtes.

Au niveau théorique, je me suis intéressé aux modèles topologiques. Ces modèles constituent toujours un thème de recherche dans le domaine de l'apprentissage non supervisé. En particulier, les cartes topologiques auto-organisatrices présentent un certain nombre d'intérêts :

- Elles réalisent une partition des données en des sous-ensembles "homogènes". Compte tenu du nombre élevé de sous-ensembles générés, elles représentent l'organisation d'une base de données en sous-ensembles pouvant être représentés par des prototypes. De ce point de vue, l'organisation proposée par la carte facilite la fouille de données.

- Elles sont utilisées comme un outil de visualisation de données. En effet, la "projection" des données sur l'espace discret de la carte (généralement de dimension 2) fournit des représentations visuelles conviviales facilement interprétables.

- L'organisation des données en une partition de plusieurs sous-ensembles "homogènes" constitue une base pour faire de la classification. En effet, il suffit pour cela de procéder à des fusions entre sous-ensembles afin d'obtenir une partition en un nombre limité de sous-ensembles.

4.2 Contributions

4.2.1 Apprentissage non supervisé et données catégorielles et continues

Dans une grande partie de mes travaux de recherche, je m'intéresse aux observations binaires, c'est-à-dire aux observations $\mathbf{x} = (x^1, ..., x^k, ..., x^n)$ dont les composants x^j sont des variables de $\beta = \{0, 1\}$, ainsi $\mathbf{x} \in \mathcal{D} \subset \beta^n$, [Cox70, Mar89]. D'une manière générale, on s'intéressera aux observations $\mathbf{x}$ dont les composantes sont des variables qualitatives à plusieurs modalités et continues.

Une grande partie de mes travaux depuis mon arrivé à l'université Paris 13 s'est focalisée sur l'apprentissage non supervisé à base de modèles de cartes auto-organisatrices. L'objectif était d'être en mesure d'apprendre à partir de données catégorielles, binaires ou encore mixtes. Je me suis intéressé aux modèles de quantification, mais essentiellement au cas de modèles de mélanges permettant de tenir compte de la nature des variables. Les modèles de mélanges statistiques sont utilisés pour modéliser des situations dans lesquelles certaines variables sont mesurées, mais où une variable catégorielle est manquante.

Ainsi, parmi les résultats que j'ai obtenus, je peux citer les modèles suivants. Tout d'abord, un modèle probabiliste de cartes auto-organisatrice pour les données catégorielles sans aucun codage [RI-3]. Ensuite, j'ai travaillé avec Mlle Nicoleta Rogovschi[1] que j'ai encadré en thèse sur la question d'étendre les modèles de mélanges aux données binaires [RI-5, CI-13, CN-12] et aux données mixtes [CI-10, CN-8]. Nous avons montré qu'il était possible d'apprendre sur des données binaires ou mixtes avec les cartes auto-organisatrices déterministes et probabilistes, tout en respectant la nature et la structure initiale des données, qu'elles soient simplement binaires, catégorielles ou mixtes. Dans les modèles probabilistes proposés, chaque cellule de la carte auto-organisatrice est associée à une distribution de Bernoulli et/ou une

[1] devenue depuis septembre 2010 MCF à l'université de Paris 5

distribution gaussienne selon la nature des données (binaires, mixtes : binaires et continues). L'apprentissage dans ces modèles a pour objectif d'estimer la fonction densité sous forme d'un mélange de densités élémentaires. Chaque densité élémentaire est, elle aussi, un mélange de lois définies sur un voisinage. Ces nouveaux algorithmes d'apprentissage probabiliste et non supervisé utilisent l'algorithme EM pour maximiser la vraisemblance des données, afin d'estimer les paramètres du modèle de mélange. Ces algorithmes ont une portée pratique, aussi bien en classification qu'en visualisation des données. Concernant les modèles déterministes, j'ai proposé une extension de mon algorithme déterministe (MTM : Mixed Topological Map) [CI-17] aux données mixtes en introduisant un système de pondération des variables adaptative [CN-3, RI-2]. L'apprentissage est combiné à un mécanisme de pondération des différentes variables sous forme de poids d'influence sur la pertinence des variables. L'apprentissage des pondérations et des prototypes est réalisé d'une manière simultanée en favorisant une classification optimisée des données.

4.2.2 Apprentissage non supervisé et données structurées en séquences

Comme il est souligné dans la littérature, l'hypothèse i.i.d est généralement faite dans le cadre de l'apprentissage non supervisé. Je me suis intéressé à étendre le modèle probabiliste des cartes auto-organisatrices aux données séquentielles (non i.i.d.). Les données séquentielles consistent à représenter des ensembles d'unités d'une longueur fixe ou variable et éventuellement d'autres caractéristiques intéressantes, telles que les comportements dynamiques et les contraintes de temps. Ces propriétés rendent les données séquentielles distinctes par rapport aux autres types de données (continues, binaires). Il est souvent impossible de les représenter comme des points dans un espace multidimensionnel et d'utiliser les algorithmes de classification existants.

Malheureusement, les paradigmes de l'auto-organisation ne peuvent pas être facilement transférés à des données non i.i.d. Différentes approches ont été développées pour intégrer l'information temporelle dans la carte d'auto-organisation. La façon la plus directe consiste à inclure les versions temporelles en entrée ou à ajouter un prétraitement pour la capture spatiale dynamique [Kan90, Wie03]. Une variété de modèles existe pour les cartes auto-récurrentes : la carte de Kohonen temporelle (TKM), la SOM récurrente (RSOM), la SOM récursive (RecSOM), et SOM pour les données structurées (SOMSD) [SH03, CT93, HSTM03, Voe02]. Les chaînes de Markov cachées (les HMMs) sont connues pour être la façon la plus naturelle de traiter les données séquentielles ou temporelles. Afin de surmonter les limites des HMMs, des travaux récents dans [BT06, Bou08, BT04] proposent un nouveau paradigme, appelé *topological HMM*, qui manipule les nœuds du graphe associé au HMM et ses transitions dans un espace euclidien. Cette approche modélise la structure locale d'un HMM et extrait leur forme en définissant une unité d'information comme une forme composée d'un groupe de symboles d'une séquence. Dans [FS08a, FS08b], les auteurs proposent un modèle hybride SOM-HMM dans lequel chaque cellule de la carte représente un HMM. Cependant, le processus d'organisation n'est pas intégré explicitement dans l'approche HMM. Le modèle GTM a été étendu au modèle de série chronologique [BHS97] et aux données structurées [BMS10]. Cependant, la manière dont il réalise l'organisation topographique est tout à fait différente de celles utilisées dans mes approches. Dans GTM le mélange des composantes est paramétré par une combinaison linéaire de fonctions non linéaires des positions des cellules de la carte dans l'espace latent.

De la même manière, mes travaux sur les données séquentielles, ou données non i.i.d, relèvent du même objectif, celui de prendre en considération la nature et l'aspect séquentiel des données au cours de l'apprentissage. Ce thème de recherche est vraisemblablement l'un des axes que je vais investir de l'énergie. Dans ce cadre, je propose une extension de l'algorithme de la carte topologique probabiliste pour des données séquentielles multivariées en supposant que les données séquentielles sont générées de la même manière que les HMMs (Hidden Markov Models). Il est intéressant de voir que les chaînes de Markov cachées peuvent être un cas particulier de l'un de nos modèles lorsque la contrainte topologique est relaxée. Ce modèle a l'intérêt d'intégrer la notion d'auto-organisation dans les modèles de chaînes de Markov cachées (HMMs) ; d'ailleurs, j'aurai pu le nommer HMM auto-organisé. Nous avons introduit aussi la réduction de dimension. Le modèle proposé dans cet axe de recherche permet de visualiser les séquences dans un espace réduit présenté sous forme d'une grille 2D ou 1D. Je suis par ailleurs guidé pour cette problématique par un projet CIFRE avec l'INA (Institut national de l'audiovisuel). Une thèse a démarré en octobre 2009 pour travailler sur les données structurées en séquences (Doctorante : Mlle Rakia Jaziri) [CI-11, CN-2, CI-1, CN-1, CI-2].

4.2.3 Autres travaux

Apprentissage non supervisé et données structurées en graphes

Une partie de mes travaux commence à s'inscrire dans des applications aux données structurées en graphe. C'est un axe nouveau pour moi qui est plein de verrous scientifiques. Cet axe de recherche est intéressant puisqu'il a des visées pratiques et souleve des questions algorithmiques, la manipulation de graphes, la notion de référents (hub, point d'intérêt, leader) et l'espace de représentation des graphes. Evidemment, le problème est difficile et ouvre des perspectives de recherche intéressantes puisque les données ne sont pas représentées d'une manière classique "plate" individus/variables. Même si les résultats, pour l'instant, ne sont pas extraordinaires, mais nous proposons une nouvelle vision de la classification et de la visualisation de graphes.

Le point commun avec les travaux issus de la littérature, est d'étudier comment les méthodes vectorielles peuvent être appliquées au clustering de graphe. L'idée commune est d'utiliser une transformation de graphe dans un nouvel espace où la similarité peut être calculée [Chu97, Moh91]. Des généralisations d'approches utilisant les modèles des cartes auto-organisées comme la dissimilarité SOM (D-SOM) [KS02] et le Kernel SOM [MF00, And02, BJRV08] ont été définies pour s'adapter à ce type de données en utilisant les fonctions du noyau.

Je me suis intéressé aux modifications des cartes auto-organisatrices, leur permettant de traiter des données tout en introduisant la notion de structure des données au niveau de chaque cellule. Ainsi, j'ai pu, en collaboration avec H. Azzag (maître de conférences dans mon équipe), proposer avec succès un modèle unifié utilisant les cartes auto-organisatrices et le modèle inspiré du comportement des règles d'accrochage des fourmis artificielles [AGV06a]. Le modèle vise à développer un algorithme de classification simultanée : topologique et hiérarchique, d'une manière à produire une partition des données organisées sur une grille 2D et en même temps une organisation hiérarchique sous forme d'arbre au niveau de chaque cellule de la grille. Dans un premier temps, nous avons appliqué notre approche sur des données classique (individus/variables) [CI-8, CI-12, CN-10, CN-6]. En assimilant la donnée traditionnelle à un

nœud d'un graphe, cette approche peut rejoindre les autres méthodes de clustering consistant à décomposer le graphe de départ (i.e. graphe initial) en une succession de sous-graphes.

Les questions qui restent ouvertes renvoient à la fois à des questions algorithmiques, comme celle de trouver un espace de représentation des graphes, théorique, comme celle de la maximisation de la modularité et pratique comme la manipulation et la visualisation de très graphes réels.

Apprentissage non supervisé et données déséquilibrées

Enfin je tenais à mentionner parmi mes travaux ceux auquels j'ai collaboré avec Fatma Hamdi (doctorante) sur l'apprentissage avec des données déséquilibrées. Plusieurs aspects pourraient influencer les systèmes d'apprentissage existants. Un de ces aspects est lié au déséquilibre des classes dans lequel le nombre d'observations appartenant à une classe, dépasse fortement celui des observations dans les autres classes. Dans ce type de cas qui est assez fréquent, le système d'apprentissage a des difficultés au cours de la phase d'entraînement liées au déséquilibre inter-classe. Nous avons proposé une méthode de sous-échantillonnage adaptative pour traiter des bases de données déséquilibrées. Le processus procède par le sous-échantillonnage des données majoritaires, guidé par les données minoritaires [CN-5, CI-3].

Tous mes autres travaux actuels sont guidés par deux contraintes : la nécessité de construire des représentations simplifiées de données pour mettre en évidence les relations existantes, et le respect de la nature des données manipulées.

Chapitre 5

Apprentissage non supervisé et données catégorielles et continues

5.1 Introduction

Ce chapitre fait une synthèse des différents travaux que j'ai réalisés en classification non supervisée des données catégorielles, binaires et mixtes avec des variables continues. Au début de mes travaux dans cet axe, je me suis focalisé sur les données catégorielles/qualitatives et binaires puis j'ai évolué vers les données mixtes. On peut distinguer deux grandes familles de modèles de classification non supervisée : les modèles à base de modèles de mélanges et déterministes ou tout simplement des modèles de quantification. Dans ce chapitre, je développe principalement mes contributions pour la famille des modèles à base de modèles de mélanges.

Dans certains domaines tels que la finance, le marketing, la santé, la quantité de données stockées a eu une augmentation explosive. Il existe de nombreux cas de bases de données où la description des observations n'est ni continue ni exclusivement catégorielle. Les deux types de valeurs peuvent apparaître simultanément. Une pratique courante dans le clustering de données mixtes est de transformer les valeurs catégorielles en valeurs continues, puis d'utiliser un algorithme de clustering dédié aux données continues. Une autre approche consiste à discrétiser toutes les variables continues, puis à utiliser un algorithme dédié aux données catégorielles. Néanmoins, ces méthodes ne prennent pas en compte toute l'information en entrée. Par conséquent, les résultats du clustering ne sont pas fidèles à la structure initiale des données [Hsu06, HW06, HH08].

Je présente ci-dessous un état de l'art rapide sur la classification non supervisée à base des modèles de mélanges ayant un lien avec les modèles que je présente dans ce chapitre. Les modèles de mélanges statistiques sont utilisés pour modéliser des situations dans lesquelles certaines variables sont mesurées, mais où une variable catégorielle est manquante. La définition ci-dessus caractérise le problème le plus simple. En faisant des hypothèses de distribution d'un ensemble d'observations, on obtient un modèle de mélange, dont les paramètres peuvent être estimés, par exemple, par la méthode du maximum de vraisemblance. Les estimations des paramètres peuvent être utilisées pour trouver des estimations des probabilités a posteriori d'appartenance à un groupe ou cluster. L'approche des modèles de mélange est très flexible à la fois pour les distributions simples et complexes. Je présenterai dans ce chapitre mes contributions à la manipulation de données catégorielles, binaires et mixtes.

Travaux similaires

Des algorithmes de clustering ont été proposés pour les des données catégorielles. k-modes [Hua97a], ROCK [GRS99], CoolCat [BCL02], CACTUS [GGR99], MNDBIN [Gov90a] sont des algorithmes de clustering dédiés uniquement aux données catégorielles. Dans les modèles topologiques, nous avons l'approche proposée dans [IC95]. Il s'agit d'une méthode appelée KACM, qui permet de classer les modalités des variables catégorielles suivant un codage spécifique lorsque la dimension de ces variables est supérieure ou égale à deux. Cette approche a l'inconvénient dans certains cas de fournir des prototypes en contradiction avec les données en entrée. Nous avons aussi un algorithme que j'ai publié en 2000 [CI-20] qui consiste à définir une nouvelle fonction de coût pour les cartes topologiques en utilisant la distance de Hamming sur des données catégorielles codées en binaire. Ainsi, j'ai introduit la notion du centre médian comme prototype. D'autres algorithmes existent [DLSW98, DLW+98, AAW06, SZC02, Li06, ZPAS07].

Avec l'accroissement des données contenant des variables mixtes, des algorithmes spécifiques sont apparus. Les auteurs de [LB02] présentent un algorithme de clustering hiérarchique (SBAC) en proposant une nouvelle mesure de similarité adaptée. D'après la littérature, cet algorithme fonctionne bien avec des variables mixtes continues et catégorielles, mais le calcul est très coûteux. Dans [Hua97b] les auteurs proposent une fonction de coût qui prend en compte les variables continues et catégorielles séparément. La fonction de coût manipule des données mixtes et calcule la distance entre une observation et un centre ou un prototype en fonction de deux valeurs de distance – une pour les variables continues et l'autre pour les variables catégorielles. Dans [HNRL05] les auteurs proposent une méthode de clustering dédiée aux données mixtes. Dans cette méthode, les pondérations des variables sont automatiquement calculées et elles sont les mêmes pour tous les clusters. [AD07a] proposent un algorithme de type k-means qui surmonte les faiblesses de la fonction de coût proposée par [HNRL05]. Concernant les modèles à base des cartes topologiques, j'ai proposé le premier modèle en 2005 [CI-17], qui est similaire au travail développé dans [HNRL05]. L'idée principale consiste d'utiliser une mesure adaptée aux données mixtes. Tous les algorithmes cités ci-dessus sont des algorithmes appartenant à la famille "déterministe".

Des modèles probabilistes spécifiques aux données qualitatives ont été proposés [VVK05, JN07, NG98b]. Différentes approches ont été envisagées, pour différents types de données, reposant sur des formalismes probabilistes. Dans [VVK05], les auteurs proposent une généralisation probabiliste des cartes auto-organisatrices SOMM qui maximisent l'énergie variationnelle et qui additionnent la log-vraisemblance des données et la divergence de Kullback-Leibler entre une fonction de voisinage normalisée et la distribution postérieure sur les données pour les composants. Nous avons également STVQ (Soft Topographic Vector Quantization), qui emploie une certaine mesure de divergence entre les données élémentaires et les référents afin de minimiser une nouvelle fonction d'erreur [Hes01, GBO98]. Il existe aussi un modèle original qui permet de créer un graphe entre les prototypes en se basant sur les modèles de mélanges et les graphes de Delaunay [Aup05]. Ce modèle est dédié uniquement aux données continues. Un autre modèle, appelé GTM (Generative Topographic Map), est souvent présenté comme la version probabiliste des cartes auto-organisatrices [BSI98, KG01]. Cependant, la façon dont GTM atteint l'organisation topologique est très différente de celle utilisée dans les modèles des cartes topologiques. Dans GTM le mélange des composantes est paramétré par une combinaison linéaire de fonctions non linéaires des positions des cellules de la carte dans l'espace

latent (GTM à l'origine a été conçu pour les données quantitatives uniquement). Des extensions de ce modèle à un modèle dédié aux données discrètes et binaires ont été proposées [Gir01, PN06, PNG08]. A notre connaissance, il est inconnu comment étendre le modèle GTM à des mélanges de densités en dehors de la famille des distributions exponentielles. A notre connaissance, concernant le traitement des données mixtes, l'algorithme le plus avancé à base de modèles de mélanges est ce lui proposé par [JH96, HJ03]

Avant de présenter les modèles de mélanges topologiques à variable cachée "latente", je vais présenter l'algorithme EM qui est solution générale au problème de l'estimation dans le cadre des modèles de mélanges et constitue un point de passage obligatoire dans ce domaine.

5.2 Algorithme EM

L'algorithme EM [DLR77, MK97] est un algorithme itératif qui permet de trouver un maximum local de la fonction de vraisemblance des observations, lorsque chaque observation contient une partie cachée (ou non observée). Ainsi, on suppose que chaque donnée est un couple de types $(\mathbf{x}, \mathbf{z})$ où $\mathbf{x}$ est sa partie observable et $\mathbf{z}$ sa partie cachée (non observable). Nous supposons connue d'une manière explicite la forme de la fonction densité jointe $p(\mathbf{x}, \mathbf{z}; \theta)$ où θ est l'ensemble des paramètres du modèle à estimer. On suppose que l'on dispose d'une série de données indépendantes : $(\mathbf{x}_1, \mathbf{z}_1), (\mathbf{x}_2, \mathbf{z}_2), ..., (\mathbf{x}_N, \mathbf{z}_N)$, pour lesquelles $\mathbf{x}_i$ sont les parties qu'on a réellement observées et les $\mathbf{z}_i$ sont les parties cachées. Nous souhaitons par la suite maximiser le logarithme de la vraisemblance des parties, des données réellement observées $\mathcal{A} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ dont le logarithme est égal à :

$$\ln V(\mathcal{A}; \theta) = \ln V(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N; \theta) = \sum_{i=1}^{N} \ln p(\mathbf{x}_i; \theta) \tag{5.1}$$

où $p(\mathbf{x}; \theta)$ est la fonction densité de la partie observée $\mathbf{x}$. En pratique, $p(\mathbf{x}; \theta)$ est calculable en marginalisant la fonction densité $p(\mathbf{x}, \mathbf{z}; \theta)$ ($p(\mathbf{x}; \theta) = \int p(\mathbf{x}, \mathbf{z}; \theta) d\mathbf{z}$), ce qui donne souvent une fonction log-vraisemblance $\ln V(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N; \theta)$ qui n'est pas simple à optimiser.

L'algorithme EM, proposé par Dempster et al [DLR77], maximise l'expression (5.1) en utilisant le log-vraisemblance des données $\ln V(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N, \mathbf{z}_1, ..., \mathbf{z}_N; \theta)$. On désigne par $\mathbf{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_N\}$ l'ensemble des parties correspondantes et non observées. Chaque itération de l'algorithme EM comporte deux étapes :

- L'étape d'Estimation (Expectation step), dite aussi étape "E"
- L'étape de Maximisation (Maximization step), dite étape "M"

Ainsi à l'itération t, ces deux étapes se présentent de la manière suivante :

- **Etape E (Expectation step)**
  On suppose, à cette étape, que la fonction densité de la partie cachée conditionnée par la partie observée $(\mathbf{z}/\mathbf{x})$ correspond à la valeur du paramètre θ^{t-1} calculée à l'itération précédente (ou égale à l'initialisation θ^0 si $t = 1$) ; cette fonction densité s'écrit donc $(p(\mathbf{z}/\mathbf{x}, \theta^{t-1}))$. On calcule alors l'espérance :

$$Q(\theta, \theta^{t-1}) \quad = \quad E\left[\ln V(\mathcal{A}, \mathbf{Z}/\theta)/\mathcal{A}, \theta^{t-1}\right] \tag{5.2}$$

Cette expression qui est parfois appelée "la vraisemblance relative" se justifie "intuitivement". En effet, étant donné qu'on ne connaît pas les valeurs des variables cachées $\mathbf{z}_i$ associées aux observations $\mathbf{x}_i \in \mathcal{A}$, on calcule l'espérance du log-vraisemblance relativement aux variables cachées.

– **Etape M (Maximization step)**
Ayant calculé $Q(\theta, \theta^{t-1})$ à l'étape E, il s'agit à cette étape de maximiser cette expression par rapport à θ. On prend alors :

$$\theta' = \arg\max_{\theta} Q(\theta, \theta^{t-1})$$

Il est démontré alors que chaque itération (EM) fait croître la fonction log-vraisemblance (5.1) : $\ln V(\mathcal{A}, \theta^t) \geq \ln V(\mathcal{A}, \theta^{t-1})$, [DLR77]. Cette formulation générale mène à différents algorithmes dédiés à l'apprentissage des paramètres des modèles de mélange. Ainsi, EM peut être vu comme un principe de conception d'algorithme plutôt qu'un simple algorithme.

5.3 Modèle de mélange et carte topologique

L'introduction d'un modèle de mélange qui permet d'exprimer la densité des observations en fonction de lois de paramètres amène tout naturellement à maximiser la vraisemblance de l'ensemble des observations et à déterminer par cette maximisation les paramètres optimaux du mélange représenté par la carte. Plusieurs méthodes peuvent êtres utilisées pour atteindre l'optimum : formalisme des nuées dynamiques, algorithme du gradient, algorithme d'Estimation-Maximisation. Un algorithme utilisant la modélisation par des modèles de mélange a été proposé pour les données réelles, $\Re^n$. Cet algorithme appelé PrSOM (PRobabilistic Self-Organizing Map (PrSOM)), [ABT97, ABT98], suppose que la distribution de probabilité $p(\mathbf{x}/c)$ associée à chaque cellule de la carte prend une forme analytique qui est représentée par une loi gaussienne sphérique. Chaque fonction densité est définie par son vecteur moyen qui est le référent $\mathbf{w}_c = (w_c^1, ..., w_c^n)$ ainsi que par l'écart type σ_c qui varie maintenant avec chaque cellule c. Les mélanges des fonctions considérées sont donc des mélanges de fonctions gaussiennes. Les paramètres à estimer dans ce cas représentent, $\theta = (\mathcal{W}, \Sigma)$, l'ensemble des référents $\mathcal{W}$ et l'ensemble des écarts-types $\Sigma = \{\sigma_c, c = 1..K\}$.

Dans le cas du modèle PrSOM, l'estimation se fait par maximisation de la vraisemblance classifiante. Nous ne présentons pas ici ce modèle, mais nous détaillerons dans la suite de ce chapitre le modèle qui sera utilisé pour estimer les paramètres de la carte topologique probabiliste dédiée aux données catégorielles, binaires et mixtes (continues et binaires).

Pour définir le modèle des cartes topologiques à base de modèle de mélange, on associe à chaque cellule c de la carte $\mathcal{C}$ une fonction densité $p(\mathbf{x}/\theta_c)$ dont les paramètres seront notés θ. Pour définir le mélange de densités des cartes topologiques, nous utiliserons le formalisme bayésien introduit par Luttrel [Lut94]. Ce formalisme suppose que les observations $\mathbf{x}$ sont générées de la manière suivante : on commence par choisir une cellule centrale c^* de la carte $\mathcal{C}$ qui permet de déterminer un voisinage de cellule suivant la probabilité conditionnelle $p(c/c^*)$ qui est supposée connue. Ainsi, l'observation $\mathbf{x}$ est générée suivant la probabilité $p(\mathbf{x}/c)$. Ce processus permet de modéliser les différentes étapes de propagation de l'information entre les différentes cellules $c \in \mathcal{C}$ et $c^* \in \mathcal{C}$. Ce formalisme nous amène à définir le générateur des

données $p(\mathbf{x})$ par un mélange de probabilités défini comme suit :

$$p(\mathbf{x}) = \sum_{c^* \in \mathcal{C}} p(c^*) p_{c^*}(\mathbf{x}) \tag{5.3}$$

avec

$$p_{c^*}(\mathbf{x}) = p(\mathbf{x}/c^*) = \sum_{c \in \mathcal{C}} p(c/c^*) p(\mathbf{x}/c). \tag{5.4}$$

Ainsi, $p(\mathbf{x})$ apparaît comme un mélange des probabilités $p_{c^*}(\mathbf{x})$. L'observation $\mathbf{x}$ s'obtient premièrement par la sélection de la cellule c^* puis de la cellule c de la carte $\mathcal{C}$, ensuite par la génération de $\mathbf{x}$ avec la probabilité $p(\mathbf{x}/c)$. Les coefficients du mélange sont les probabilités a priori $p(c^*)$ et les fonctions densités relatives à chaque élément du mélange $p_{c^*}(\mathbf{x})$ (eq.(5.4)). Ce formalisme montre qu'on peut calculer $p(\mathbf{x})$ à condition de connaître pour chaque cellule c la fonction de densité $p(\mathbf{x}/c)$ et la probabilité conditionnelle $p(c/c^*)$ d'activation de la cellule $c \in \mathcal{C}$ connaissant la cellule $c^* \in \mathcal{C}$.

Afin d'introduire la relation de voisinage topologique dans le formalisme probabiliste, nous supposons que chaque cellule c de la carte $\mathcal{C}$ est d'autant plus active qu'elle est proche de la cellule choisie c^*, (eq.(5.3), eq.(5.4)). Ceci nous permet de définir la probabilité $p(c/c^*)$ en fonction de la fonction de voisinage $\mathcal{K}^T(.)$:

$$p(c/c^*) = \frac{\mathcal{K}^T(\delta(c, c^*))}{T_{c^*}} \tag{5.5}$$

où $T_{c^*} = \sum_{r \in C} \mathcal{K}^T(\delta(r, c^*))$ est un terme normalisant pour obtenir des probabilités et $\delta(c, c^*)$ est la longueur du chemin le plus court qui sépare les deux cellules c et c^*.

La figure 5.1 un exemple de carte indiquant les la distribution de probabilité autour d'un point central c^*. Pour définir complètement $p(\mathbf{x})$, il reste à définir les coefficients du mélange $p(c^*)$ et les paramètres de la densité $p(\mathbf{x}/c)$. Ce modèle, qui généralise le modèle classique des cartes topologiques introduit par Kohonen permet d'obtenir une quantification de l'espace des données, mais aussi une estimation des densités locales.

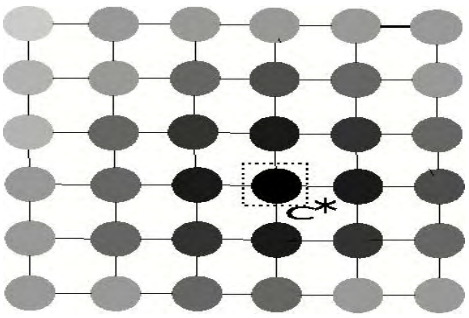


FIG. 5.1 – Une grille bi-dimensionnelle avec $\mathcal{C} = 6 \times 6$ cellules utilisées par le modèle générateur. Le niveau de gris indique les cellules déterminées par la probabilité $p(c/c^*)$

Dans le cadre du modèle des cartes topologiques, l'usage de l'algorithme EM s'explique par l'existence d'une variable cachée notée $\mathbf{z}$, constituée par le couple de cellules c et c^*,

$\mathbf{z} = (c, c^*)$, responsable de la génération d'une donnée observée $\mathbf{x}$. En effet, la variable cachée $\mathbf{z} = (c, c^*)$ apparaît lorsqu'on écrit :

$$p(\mathbf{x}) = \sum_{z \in \mathbf{Z}} p(\mathbf{x}, \mathbf{z}) = \sum_{c \in \mathcal{C}, c^* \in \mathcal{C}} p(\mathbf{x}/c) p(c/c^*) p(c^*)$$

car $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}, c, c^*) = p(c^*) p(c/c^*) p(\mathbf{x}/c)$

Ainsi, à chaque donnée réellement observée $\mathbf{x}$, il lui correspond une donnée catégorielle disjonctive non observée $\mathbf{z}$ qui appartient à $\mathcal{C} \times \mathcal{C}$. Si l'on code cette variable par le codage binaire disjonctif, on obtient un vecteur binaire $\mathbf{z}$ de dimension $K \times K$ dont les composantes $z_{(c,c^*)}$ sont définies par :

$$z^i_{(c,c^*)} = \begin{cases} 1 & \text{pour } \mathbf{z}_i = (c, c^*) \\ 0 & \text{sinon} \end{cases} \tag{5.6}$$

Par la suite, à chaque donnée réellement observée $\mathbf{x}_i$ de l'ensemble d'apprentissage $\mathcal{A}$, nous supposons qu'il lui correspond une donnée non observée $\mathbf{z}_i \in \mathcal{C} \times \mathcal{C}$ qui sera codée par le vecteur binaire appartenant à $\{0,1\}^{K \times K}$ dont toutes les composantes sont nulles sauf la composante d'indice $z^i_{(c,c^*)}$ qui vaut 1. On a donc la vraisemblance des données qui s'écrit par :

$$V^T(\mathcal{A}, \mathbf{Z}; \theta) = \prod_{i=1}^{N} \prod_{c^* \in \mathcal{C}} \prod_{c \in \mathcal{C}} \left[p(c^*) \frac{\mathcal{K}^T(\delta(c^*, c))}{T_{c^*}} p(\mathbf{x}_i/c) \right]^{z^i_{(c,c^*)}}$$

Ainsi, le log-vraisemblance s'écrit :

$$\ln V^T(\mathcal{A}, \mathbf{Z}; \theta) = \sum_{\mathbf{x}_i \in \mathcal{A}} \sum_{c^* \in \mathcal{C}} \sum_{c \in \mathcal{C}} z^i_{(c,c^*)} \left[\ln p(c^*) + \ln \left(\frac{\mathcal{K}^T(\delta(c^*, c))}{T_{c^*}} \right) + \ln(p(\mathbf{x}_i/c)) \right].$$

L'application de l'algorithme EM pour la maximisation de la vraisemblance des données observées nécessite d'une part l'estimation de l'espérance $Q^T(\theta, \theta^t)$ (eq. 5.2) et, d'autre part, sa maximisation par rapport à l'ensemble des paramètres θ. On sait que :

$$Q^T(\theta, \theta^t) = E\left[\ln V^T(\mathcal{A}, \mathbf{Z}; \theta) / \mathcal{A}, \theta^t \right]$$

$$Q^T(\theta, \theta^t) = \sum_{\mathbf{x}_i \in \mathcal{A}} \sum_{c^* \in \mathcal{C}} \sum_{c \in \mathcal{C}} E(z^i_{(c,c^*)}/\mathbf{x}_i, \theta^t) \left[\ln(\theta_{c^*}) + \ln \left(\frac{\mathcal{K}^T(\delta(c^*, c))}{T_{c^*}} \right) + \ln(p(\mathbf{x}_i/c)) \right]$$

où θ^t est l'ensemble des paramètres estimés à la t^{e} itération de l'algorithme, et θ l'ensemble des paramètres recherchés. La variable $z^i_{(c,c^*)}$ étant une variable de Bernoulli,

$$E(z^i_{(c,c^*)}/\mathbf{x}_i, \theta^t) = p(z^i_{(c,c^*)} = 1/\mathbf{x}_i, \theta^t) = p(c, c^*/\mathbf{x}, \theta^t) = \frac{p(c^*) p(c/c^*) p(\mathbf{x}/c)}{\sum_{r \in \mathcal{C}} p(r) p_r(\mathbf{x})} \tag{5.7}$$

Ainsi, la fonction $Q^T(\theta, \theta^t)$ s'écrit :

$$\begin{aligned} Q^T(\theta, \theta^t) &= \sum_{\mathbf{x}_i \in \mathcal{A}} \sum_{c^* \in \mathcal{C}} \sum_{c \in \mathcal{C}} p(c, c^*/\mathbf{x}_i, \theta^t) \ln(p(\mathbf{x}_i/c)) \rightarrow Q_1^T(\theta^C, \theta^t) \\ &+ \sum_{\mathbf{x}_i \in \mathcal{A}} \sum_{c^* \in \mathcal{C}} \sum_{c \in \mathcal{C}} p(c, c^*/\mathbf{x}_i, \theta^t) \ln(p(c^*)) \rightarrow Q_2^T(\theta^{C^*}, \theta^t) \\ &+ \sum_{\mathbf{x}_i \in \mathcal{A}} \sum_{c^* \in \mathcal{C}} \sum_{c \in \mathcal{C}} p(c, c^*/\mathbf{x}_i, \theta^t) \ln \left(\frac{\mathcal{K}^T(\delta(c^*, c))}{T_{c^*}} \right) \rightarrow Q_3^T(\theta^t) \end{aligned} \tag{5.8}$$

On observe que la fonction $Q^T(\theta, \theta^t)$ se décompose en une somme de trois termes. Le premier terme $Q_1^T(\theta^C, \theta^t)$ ne dépend que des paramètres θ^C (paramètres de la loi de probabilité associés à chaque cellule). Le second terme $Q_2^T(\theta^{C^*}, \theta^t)$ ne dépend que de θ^{C^*} (l'ensemble des probabilités a priori). Le troisième terme $Q_3^T(\theta^t)$ est constant par rapport à l'ensemble des paramètres θ et ne dépend que de θ^t. Maximiser la fonction auxiliaire (eq.5.8) par rapport à θ revient à maximiser les deux premiers termes de cette fonction séparément $Q_2^T(\theta^{C^*}, \theta^t)$ par rapport à θ^{C^*} et la fonction $Q_1^T(\theta^C, \theta^t)$ par rapport θ^C.

Estimation des probabilités a priori $\theta_{c^*} = p(c^*)$
Afin d'estimer les probabilités a priori θ^{C^*}, il faut maintenant donner une forme explicite aux probabilités et maximiser le terme $Q_2^T(\theta^{C^*}, \theta^t)$. On suppose pour le modèle topologique que les probabilités a priori θ_{c^*} s'expriment par une relation du type softmax : $\theta_{c^*} = \frac{e^{\lambda_{c^*}}}{\sum_{r \in \mathcal{C}} e^{\lambda_r}}$. On réécrit alors l'expression $Q_2^T(\theta^C, \theta^t)$. Ainsi, pour $\frac{\partial Q_2^T(\theta^{C^*}, \theta^t)}{\partial \lambda_{c^*}} = 0$, on obtient l'expression permettant d'estimer le paramètre θ_{c^*} :

$$\theta_{c^*} = \frac{\sum_{\mathbf{x}_i \in \mathcal{A}} p(c^*/\mathbf{x}_i, \theta^t)}{N} \tag{5.9}$$

avec $p(c^*/\mathbf{x}_i, \theta^t) = \sum_{c \in \mathcal{C}} p(c, c^*/\mathbf{x}_i, \theta^t)$.

5.3.1 Modèle dédié aux données catégorielles

Dans le cas où les variables observées $\mathbf{x}$ sont constituées de composantes catégorielles, nous supposons que la probabilité $p(\mathbf{x}/c)$ se présente sous forme d'une table de probabilités unidimensionnelle. Nous parlerons dans ce cas de CPrSOM (Categorical Probabilistic Self-Organising Map) pour le modèle. Afin de simplifier ce modèle, nous supposons par la suite que les n composantes catégorielles de $\mathbf{x} = (x^1, x^2, ..., x^k, ..., x^n)$ sont indépendantes, ce qui nous permet d'écrire $p(\mathbf{x}) = \prod_{k=1}^n p(\mathbf{x}^k)$. Durant l'apprentissage, il faudra donc calculer les valeurs associées aux n tables de probabilités unidimensionnelles des composantes de $\mathbf{x}$. Nous associons dans la suite à chaque cellule c de la carte n tables unidimensionnelles de probabilités. Le modèle de carte topologique propose donc de présenter des mélanges de tables de probabilités. La table 5.1 schématise la table de probabilité unidimensionnelle associée à la variable x^k et à la cellule c. Cette table sera notée $\theta^{k,c}$.

On rappelle que $x^k \in M_k$ qui est l'ensemble fini formé par l'énumération des m_k modalités $\{y_1^k, y_2^k, ..., y_{m_k}^k\}$ de la k^{e} composante x^k. Dans ce cas, $\mathbf{x} \in \mathcal{D} \subset M_1 \times M_2 \times .... \times M_n$. L'hypothèse de l'indépendance des composantes de $\mathbf{x}$ permet d'écrire :

$$p(\mathbf{x}/c) = \prod_{k=1}^{n} p(x^k/c)$$

où $p(x^k/c)$ représente une table unidimensionnelle de probabilités (de dimension m_k) contenant les probabilités des m_k modalités de la variable x^k. Cette table de probabilités sera notée par la suite $\theta^{k,c}$:

$$\theta^{k,c} = \{\theta_j^{k,c}, j = 1 \dots m_k\} \text{ avec } \theta_j^{k,c} = p(x^k = y_j^k/c).$$

$\theta^{k,c}$
$p(x^k = y_1^k/c)$
$p(x^k = y_2^k/c)$
...
$\theta_j^{k,c} = p(x^k = y_j^k/c)$
...
$p(x^k = y_{m_k}^k/c)$

TAB. 5.1 – Table de probabilités $p(x^k/c)$ notée par $\theta^{k,c}$ avec m_k composantes (modalités), associée à la variable catégorielles x^k à la cellule c.

On suppose que les probabilités $\theta_j^{k,c}$ constituant les tables de probabilités θ^c de la cellule c s'expriment par une relation du type softmax :

$$\theta_j^{k,c} = p(x^k = y_j^k/c) = \frac{e^{\lambda_j^{k,c}}}{\sum_{q=1..m_k} e^{\lambda_q^{k,c}}},$$

D'autre part, étant donné que nous avons fait l'hypothèse de l'indépendance des composantes x^k de l'observation $\mathbf{x}$, on a $p(\mathbf{x}/c) = \prod_{k=1}^{n} p(x^k/c)$. D'où la réécriture du terme $Q_1^T(\theta^C, \theta^t)$:

$$Q_1^T(\theta^C, \theta^t) = \sum_{k=1..n} \sum_{\mathbf{x}_i \in \mathcal{A}} \sum_{c \in \mathcal{C}} p(c/\mathbf{x}_i, \theta^t) \ln(p(x_i^k/c))$$

Si on note par $\tau_{k,j} = \{\mathbf{x}_i \in \mathcal{A}; x_i^k = y_j^k\}$ l'ensemble des individus $\mathbf{x}_i$ qui ont répondu par la modalité j à la k^e composante, l'expression $Q_1^T(\theta^C, \theta^t)$ s'exprime alors par :

$$Q_1^T(\theta^C, \theta^t) = \sum_{k=1..n} \sum_{j=1..m_k} \sum_{\mathbf{x}_i \in \tau_{k,j}} \sum_{c \in \mathcal{C}} p(c/\mathbf{x}_i, \theta^t) \left[\lambda_j^{k,c} - \ln(\sum_{q=1..m_k} e^{\lambda_q^{k,c}}) \right]$$

Définir l'ensemble des paramètres θ^C de la carte $\mathcal{C}$ revient à déterminer pour chaque cellule c le paramètre $\theta_j^{k,c}$ correspondant à la modalité j de la variable k. Après le calcul de la dérivée de la fonction $Q_1^T(\theta^C, \theta^t)$ par rapport à $\lambda_j^{k,c}$ on obtient ainsi :

$$\theta_j^{k,c} = \frac{\sum_{\mathbf{x}_i \in \tau_{k,j}} p(c/\mathbf{x}_i, \theta^t)}{\sum_{\mathbf{x}_i \in \mathcal{A}} p(c/\mathbf{x}_i, \theta^t)} \tag{5.10}$$

Avec $p(c/\mathbf{x}_i, \theta^t) = \sum_{c^* \in \mathcal{C}} p(c, c^*/\mathbf{x}, \theta^t)$

Exemple d'application

Cette expérience concerne une base de données composées de chiffres manuscrits ("0"−"9") extraits à partir d'une collection de cartes des services hollandaises [AN07]. On a 200 exemples pour chaque caractère, ainsi, on a au total 2000 exemples. Chaque exemple est une imagette binaire (pixel "noir" ou "blanc") de dimension 15×16. L'ensemble de données forme une matrice binaire de dimension 2000×240. Chaque variable qualitative est un pixel à deux

$p(x^{pixel} = \text{On}/c)$
$p(x^{pixel} = \text{Off}/c)$

TAB. 5.2 – Table de probabilités $\theta^{pixel,c}$ avec 2 composantes associées à la variable catégorielle *pixel*.

valeurs possibles "On=1" and "Off=0".

Chaque cellule de la carte apprise avec l'algorithme CPrSOM à une table de probabilité de dimension 240 égale au nombre de variables catégorielles. Chaque composante de la table est constituée d'une table de probabilités de deux composantes $\theta^{pixel,c}$ (table 5.2), $(p(x^{pixel} = \text{On}/c), p(x^{pixel} = \text{Off}/c))$. La figure 5.2 montre les probabilités $p(x^{pixel} = \text{On}/c)$ pour chaque pixel sur toute la carte 16×16. Il est clair que l'ordre topologiques est préservé.

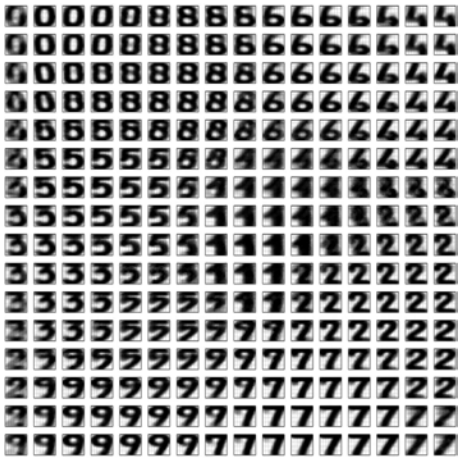


FIG. 5.2 – CPrSOM : La probabilité $p(x^{pixel} = \text{On}/c))$ associée pour chaque table de probabilité de la carte 16×16. Le niveau de gris est proportionnel à la probabilité que le pixel soit en mode "ON".

BeSOM-ε_c	$\epsilon = \varepsilon_c$	$\mathbf{w}_c = (w_c^1, ..., w_c^k, ..., w_c^n)$
BeSOM-ϵ_c	$\epsilon_c = (\varepsilon_c^1, ..., \varepsilon_c^k, ..., \varepsilon_c^n)$	$\mathbf{w}_c = (w_c^1, ..., w_c^k, ..., w_c^n)$

TAB. 5.3 – Les paramètres des trois versions BeSOM

5.3.2 Modèle dédié aux données binaires

Nous présentons dans ce paragraphe l'adaptation du modèle aux données binaires. Nous parlerons dans ce cas de BeSOM (Bernoulli Self-Organising Map) pour le modèle ou pour l'algorithme d'apprentissage permettant d'estimer les paramètres. L'algorithme BeSOM est ainsi une application directe de l'algorithme EM. Nous avons développé trois versions de BeSOM (tableau 5.3). La première est développée sous l'hypothèse que le paramètre de probabilité ϵ de loi de Bernoulli dépend uniquement de la cellule ($\epsilon = \varepsilon_c$). La deuxième est le cas général où le paramètre de probabilité dépend de la cellule et de la variable $\epsilon_c = (\varepsilon_c^1, ..., \varepsilon_c^k, ..., \varepsilon_c^n)$. La troisième version estime un paramètre qui dépend seulement de la carte ($\epsilon = \varepsilon$). Nous ne présentons ici que les deux premières versions.

Le formalisme des cartes topologiques probabilistes suppose que les données observées sont générées suivant la loi de mélange définie par les équations (eq.(5.3), eq.(5.4)). Les coefficients à estimer sont les paramètres de probabilités $p(c^*)$ et les paramètres des fonctions de densités relatives à chaque élément du mélange qui sont données par $p(\mathbf{x}/c)$. Dans la suite, nous allons donner une forme particulière à ces fonctions qui permettent explicitement de prendre en compte les dépendances entre les modalités d'une variable qualitative codée en binaire.

5.3.2.1 Modèle de mélange de Bernoulli : BeSOM-ϵ

Dans le cas où les variables observées $\mathbf{x} = (x^1, ..., x^k, ..., x^n)$ sont constituées de composantes x qualitatives codées en binaire, nous supposons que la probabilité $p(\mathbf{x}/c)$ se présente sous forme de loi de Bernouilli de paramètre $\epsilon_c = (\varepsilon_c^1, ..., \varepsilon_c^k, ..., \varepsilon_c^n)$ composé de n valeurs ε_c^k dépendant de chaque variable binaire x^k. Ainsi, la loi de probabilité est écrite [NG98a] :

$$p(\mathbf{x}/\epsilon_c, \mathbf{w}_c) = \prod_{k=1}^{n} (\varepsilon_c^k)^{|x^k - w_c^k|} (1 - \varepsilon_c^k)^{1-|x^k - w_c^k|}$$

où $\varepsilon^k \in]0\ 1/2[$ et $w^k \in \beta = \{0, 1\}$.

Cette formulation de la loi de probabilité [CG91b, NG98a] suppose que la loi de probabilité génère des observations $\mathbf{x}$ de β^n autour d'un vecteur fixe $\mathbf{w}_c$ de telle façon que tous les composants soient indépendants et non corrélés et que chaque composante de $\mathbf{x}$ reproduise la composante correspondante de $\mathbf{w}_c$ avec la même probabilité $(1 - \varepsilon_c^k)$. Ainsi, cette formulation introduit la notion de vecteur référent puisqu'elle suppose que la loi de probabilité considérée $p(\mathbf{x}/\mathbf{w}_c, \varepsilon_c)$ génère des données plus ou moins similaires à $\mathbf{w}_c$.

L'ensemble des paramètres permettant de définir la loi de Bernoulli d'une cellule c de la carte $\mathcal{C}$ est constitué de l'union de tous les couples $\theta_c = (\mathbf{w}_c, \epsilon_c)$ qui constituent les paramètres de la loi de Bernoulli génératrice au niveau de chaque cellule c, $\theta^C = \cup_{c=1}^{K} \theta_c$. Ainsi, les paramètres sont $\theta = \theta^C \cup \theta^{C^*}$, qui permettent de définir le modèle probabiliste de la carte

topologique de taille K où $\theta^C = \cup_{c=1}^{K} \theta_c$, où $\theta_c = (\varepsilon_c, \mathbf{w}_c)$. Pour la maximisation, le premier terme $Q_1^T(\theta^C, \theta^t)$ est réécrit de la manière suivante :

$$Q_1^T(\theta^C, \theta^t) = \sum_{\mathbf{x}_i \in \mathcal{A}} \sum_{c \in \mathcal{C}} \sum_{k=1}^{n} p(c/\mathbf{x}_i, \theta^t) \left[-\ln\left(\frac{1-\varepsilon_c^k}{\varepsilon_c^k}\right) |x_i^k - w_c^k| + \ln(1-\varepsilon_c^k) \right]$$

Comme dans la modélisation précédente, la maximisation de la fonction $Q_1^T(\theta^C, \theta^t)$ s'effectue en deux étapes : la première consiste à fixer ε_c^k et la maximiser par rapport à w_c^k ; à l'inverse, dans la deuxième étape, on fixe w_c^k et on maximise par rapport à ε_c^k.

- **Etape 1 :** Pour une valeur fixe de ε_c^k de $]0, 1/2[$, le deuxième terme de la fonction $Q_1^T(\theta^C, \theta^t)$ est constant et positif, donc la maximisation de $Q_1^T(\theta^C, \theta^t)$ est équivalente à la minimisation de "l'inertie" $\mathcal{I}_c^k(w_c^k)$ au niveau de chaque cellule c et par rapport à chaque variable k :

$$\mathcal{I}_c^k(w_c^k) = \sum_{\mathbf{x}_i \in \mathcal{A}} p(c/\mathbf{x}_i, \theta^t) \ln\left(\frac{1-\varepsilon_c^k}{\varepsilon_c^k}\right) |x_i^k - w_c^k| \tag{5.11}$$

  Le point qui minimise cette inertie est la valeur médiane de l'ensemble des observations de la variable $\mathbf{x}_i^k$ appartenant au sous-ensemble $\mathcal{A}$. Chaque composante de $\mathbf{w}_c = (w_c^1, ..., w_c^k, ..., w_c^n)$ est calculée de la manière suivante :

$$w_c^k = \begin{cases} 0 & \text{si } \left[\sum_{\mathbf{x}_i \in \mathcal{A}} p(c/\mathbf{x}_i, \theta^{t-1})(1 - x_i^k)\right] \geq \\ & \left[\sum_{\mathbf{x}_i \in \mathcal{A}} p(c/\mathbf{x}_i, \theta^{t-1}) x_i^k\right] \\ 1 & \text{sinon} \end{cases} \tag{5.12}$$

- **Etape 2 :** Dans le cas où l'on suppose que les centres médians $\mathbf{w}_c$ sont connus et fixés, la maximisation du critère $Q_1^T(\theta^C, \theta^t)$ par rapport à ε_c^k revient à résoudre l'équation $\frac{\partial Q_1^T(\theta^C, \theta^t)}{\partial \varepsilon_c^k} = 0$, ce qui donne les formules suivantes :

$$\varepsilon_c^k = \frac{\sum_{\mathbf{x}_i \in \mathcal{A}} p(c/\mathbf{x}_i, \theta^t) |x_i^k - w_{c1}^k|}{\sum_{\mathbf{x}_i \in \mathcal{A}} p(c/\mathbf{x}_i, \theta^t)} \tag{5.13}$$

5.3.2.2 Modèle de mélange de Bernoulli : BeSOM-ε_c

On se place dans les mêmes conditions que dans la modélisation précédente en remplaçant le vecteur $\epsilon_c = (\varepsilon_c^1, ..., \varepsilon_c^k, ..., \varepsilon_c^n)$ composé de n valeurs par une seule valeur ε_c dépendant uniquement de la cellule. Ainsi, la loi de probabilité est réécrite :

$$p(x^k/\varepsilon_c, w_c^k) = \varepsilon_c^{|x^k - w_c^k|} (1 - \varepsilon_c)^{1 - |x^k - w_c^k|}$$

Ainsi, la loi de probabilité qui définit le vecteur aléatoire $\mathbf{x} = (x^1, x^2, ..., x^k, ..., x^n)$ s'écrit :

$$p(\mathbf{x}/\mathbf{w}_c, \varepsilon_c) = \varepsilon_c^{\mathcal{H}(\mathbf{x}, \mathbf{w}_c)} (1 - \varepsilon_c)^{n - \mathcal{H}(\mathbf{x}, \mathbf{w}_c)} \tag{5.14}$$

Où $\mathcal{H}$ indique la distance de Hamming qui calcule le nombre de différences entre les deux vecteurs binaires $\mathbf{x} \in \beta^n$ et $\mathbf{w}_c \in \beta^n$.

On suppose que le paramètre θc constitue le paramètre de la loi de Bernoulli associée à la cellule c. D'autre part, étant donné qu'on a fait l'hypothèse de l'indépendance des composantes x^k de l'observation $\mathbf{x}$, on a $p(\mathbf{x}/c) = \prod_{k=1}^{n} p(x^k/c)$. D'où la réécriture du premier terme :

$$Q_1^T(\theta^C, \theta^t) = \sum_{\mathbf{x}_i \in \mathcal{A}} \sum_{c \in \mathcal{C}} \left[-\ln\left(\frac{1-\varepsilon_c}{\varepsilon_c}\right) p(c/\mathbf{x}_i, \theta^t)\mathcal{H}(\mathbf{x}_i, \mathbf{w}_c) + np(c/\mathbf{x}_i, \theta^t)\ln(1-\varepsilon_c) \right]$$

Puisque $\theta_c = (\varepsilon_c, \mathbf{w}_c)$, de la même manière que le modèle précédent, la maximisation de la fonction $Q_1^T(\theta^C, \theta^t)$ s'effectue en deux étapes : la première consiste à fixer ε_c et la maximiser par rapport à $\mathbf{w}_c$; à l'inverse, dans la deuxième étape, on fixe $\mathbf{w}_c$ et on maximise par rapport à ε_c. Ce processus de maximisation permet de définir les formules suivantes :

Chaque composante du vecteur $\mathbf{w}_c = (w_c^1, ..., w_c^k, ..., w_c^n)$ est calculée comme :

$$w_c^k = \begin{cases} 0 & \text{si } \left[\sum_{\mathbf{x}_i \in \mathcal{A}} p(c/\mathbf{x}_i, \theta^t)(1 - x_i^k)\right] \geq \\ & \left[\sum_{\mathbf{x}_i \in \mathcal{A}} p(c/\mathbf{x}_i, \theta^t) x_i^k\right] \\ 1 & \text{sinon} \end{cases}, \tag{5.15}$$

$$\varepsilon_c = \frac{\sum_{\mathbf{x}_i \in \mathcal{A}} p(c/\mathbf{x}_i, \theta^t)\mathcal{H}(\mathbf{x}_i, \mathbf{w}_c)}{\sum_{\mathbf{x}_i \in \mathcal{A}} np(c/\mathbf{x}_i, \theta^t)} \tag{5.16}$$

n est la dimension du vecteur binaire $\mathbf{x}_i$.

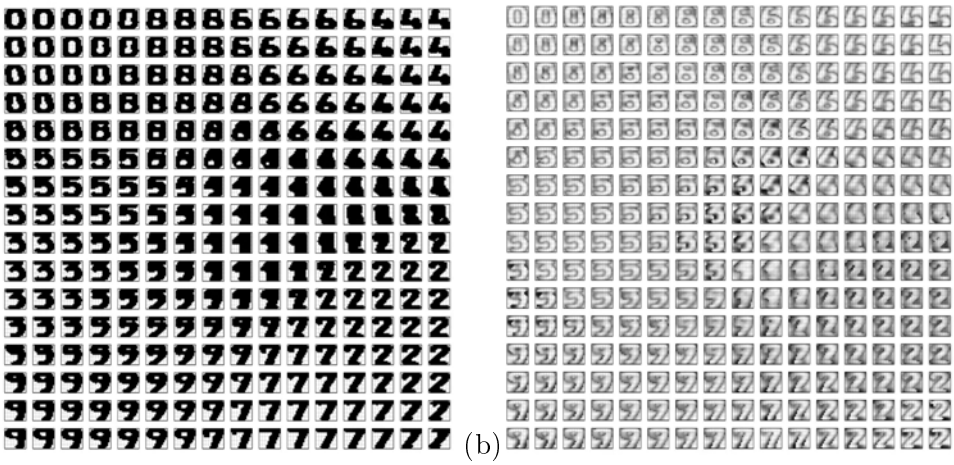
Exemple d'application

Dans le cas où les variables catégorielles sont codées en binaires, nous avons appliqué les deux versions de l'algorithme BeSOM. Avec le modèle qu'on propose, on peut obtenir différents niveaux de pertinences. En fonction de l'hypothèse sur la probabilité ε, on peut avoir deux cas :

1) La probabilité ε dépend de la cellule c et de la variable k : $\epsilon_c = (\varepsilon_c^1, \varepsilon_c^2, ..., \varepsilon_c^k, ..., \varepsilon_c^n)$ figure 5.3(a) (cas général)
2) La probabilité ϵ dépend d'une cellule seulement $\epsilon = \varepsilon_c$, figure 5.4(b)

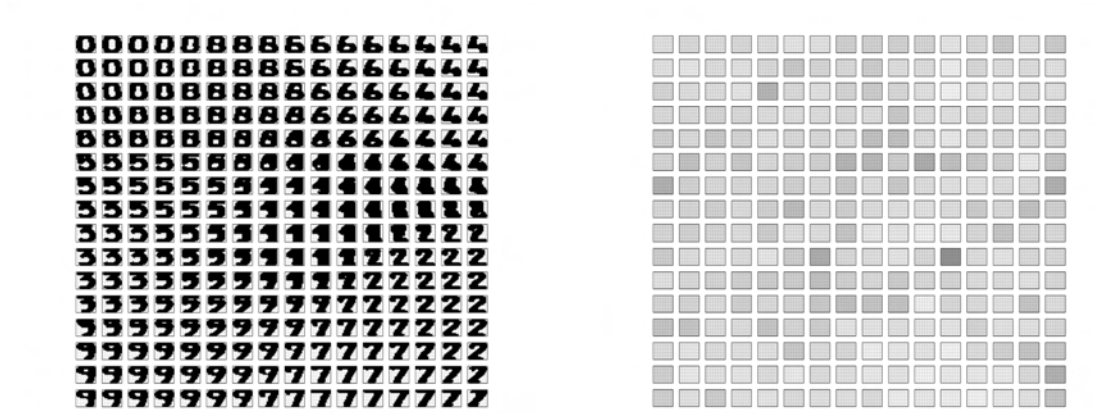
La figure 5.3 montre les paramètres estimés par notre modèle BeSOM. La figure 5.3(a) montre le vecteur prototype $\mathbf{w}_c \in \beta^n$ qui est représenté comme une imagette de dimension 15×16. La figure 5.3(b) nous montre le vecteur paramètre ϵ_c comme une imagette de la même dimension 15×16. Ainsi, chaque pixel définit la probabilité ε_c^k d'être différent de la variable binaire w_c^k associée au vecteur prototype binaire $\mathbf{w}_c$ présenté dans la figure 5.3(a). La nuance de gris de chaque pixel est proportionnelle à la probabilité d'être différent du prototype estimé $\mathbf{w}_c$. On constate que les composantes ou les probabilités ε_c^k, qui correspondent au contour de l'image, sont élevées. En observant les deux figures à la fois, on constate une organisation topologique des prototypes assez claire sur toute la carte.

La figure 5.4 nous montre les paramètres estimés par le modèle BeSOM-ε. La figure 5.4(a) montre le vecteur prototype $\mathbf{w}_c \in \beta^n$ qui est représenté aussi comme une imagette de dimension 15×16. On observe une organisation topologique des prototypes assez nette. La figure



(a) (b)

FIG. 5.3 – La carte BeSOM−ϵ de dimension 16×16 **(a)**. Chaque image représente le vecteur prototype $\mathbf{w}_c = (w_c^1, w_c^2, ...w_c^k, ..., w_c^{240})$ qui est un vecteur binaire. Le pixel blanc indique "0" et le pixel noir indique "1" **(b)**.Chaque image représente un vecteur de probabilité $\epsilon_c = (\varepsilon_c^1, \varepsilon_c^2, ..., \varepsilon_c^k, ..., \varepsilon_c^{240})$.



(a) (b)

FIG. 5.4 – La carte de BeSOM-ε de dimension 16×16 . Dans la figure (a), chaque image représente le vecteur prototype $\mathbf{w}_c = (w_c^1, w_c^2, ...w_c^k, ..., w_c^{240})$ qui est un vecteur binaire. Le pixel blanc indique "0" et le pixel noir indique "1". Dans la figure (b), chaque image représente la probabilité ε_c. Le niveau de gris de chaque cellule est proportionnel à la probabilité d'être différent du prototype estimé $\mathbf{w}_c$ (figure(b))

5.4(b) représente le paramètre ε_c estimé pour chaque cellule $c \in \mathcal{C}$. Ainsi, la nuance de gris de chaque cellule est proportionnelle à la probabilité ε_c d'être différente du vecteur prototype binaire $\mathbf{w}_c$ représenté dans la figure 5.4(a).

Evidemment, on constate que le modèle général BeSOM−ϵ présenté dans la figure 5.3 fournit plus d'informations que le modèle réduit BeSOM-ε_c (figure 5.4). Le modèle BeSOM-ε_c est intéressant quand on manipule de grandes bases de données. Nous observons, dans le pre-

mier cas (figure 5.3(b)), que les valeurs des probabilités permettent de détecter les variables pertinentes. Nous arrivons visuellement à reconnaître les chiffres. Par conséquent, il est possible d'utiliser ces valeurs pour caractériser les groupes. Dans le deuxième cas (figure 5.4(b)), il n'est pas possible de caractériser la forme des chiffres, mais nous pouvons caractériser les groupes. Ainsi, il est possible de sélectionner des groupes d'individus et de réaliser un échantillonnage optimisé.

5.3.3 Modèle dédié aux données mixtes

On s'intéresse dans cette section au traitement des variables mixtes (qualitatives et continues). Nous appellerons ce modèle PrMTM (Probabilistic Mixed Topological Map). On cherche en particulier des algorithmes à base de cartes auto-organisatrices permettant de traiter les particularités attachées à ces données. Dans le cas des variables mixtes, on cherche à mettre en évidence la typologie des modalités avec les variables continues et on essaye de faire ressortir les relations qui existent entre les différentes variables. Le modèle est aussi une extension du modèle BeSOM dédié aux données binaires. Pour définir notre modèle probabiliste manipulant des données mixtes, nous avons modélisé la distribution par un mélange de mélange de lois de Bernoulli et de lois gaussienne. Nous nous sommes inspirés du modèle *Multimix* [JH96, HJ99]. Les auteurs de ces travaux, définissent leur modèle de "clustering" comme une généralisation commune tant des modèles "latents" que des modèles de mélange de distributions normales. Ils ont proposé une approche basée sur une forme d'indépendance locale en partitionnant les variables en plusieurs sous-vecteurs.

Nous supposons par la suite que les données $\mathbf{x}$ sont des vecteurs à d composantes composées de deux parties : une partie quantitative $\mathbf{x}_i^{r[.]} = (x_i^{r[1]}, x_i^{r[2]}, ..., x_i^{r[n]})$ $(\mathbf{x}_i^{r[.]} \in \Re^n)$ et une partie qualitative codée en binaire $\mathbf{x}_i^{b[.]} = (x_i^{b[1]}, x_i^{b[2]}, ..., x_i^{b[k]}, ..., x_i^{b[m]})$ où la k^e composante $x_i^{b[k]}$ est une variable binaire $(x_i^{b[k]} \in \beta = \{0, 1\})$. Chaque observation $\mathbf{x}_i$ est ainsi la réalisation d'une variable aléatoire appartenant à $\Re^n \times \beta^m$. Avec ces notations, une observation particulière $\mathbf{x}_i = (\mathbf{x}_i^{r[.]}, \mathbf{x}_i^{b[.]})$ est un vecteur composé d'une partie quantitative et d'une partie binaire avec la dimension $d = n + m$. La partie binaire peut représenter une partie de variables catégorielles codées en binaire. Afin de simplifier ce modèle, nous supposons que la composante quantitative est indépendante de la composante binaire. Ainsi, la probabilité conditionnelle peut être réécrite comme le produit de deux termes :

$$p(\mathbf{x}/c) = p(\mathbf{x}^{r[.]}/c) \times p(\mathbf{x}^{b[.]}/c)$$

Nous prenons aussi une hypothèse nécessaire, qui est celle de considérer que les n composantes quantitatives et les m binaires sont indépendantes sachant une cellule c :

$$p(\mathbf{x}^{b[.]}/c) = \prod_{k=1}^{m} p(x^{b[k]}/c) \text{ et } p(\mathbf{x}^{r[.]}/c) = \prod_{k=1}^{n} p(x^{r[k]}/c)$$

Nous associons dans la suite à chaque cellule $c \in \mathcal{C}$ de la grille une probabilité conditionnelle qui décrit la génération d'observation par une cellule c avec les deux parties :

- La fonction densité de la partie quantitative qui est une gaussienne sphérique $p(\mathbf{x}^{r[.]}/c) = \mathcal{N}(\mathbf{x}^{r[.]}, \mathbf{w}^{r[.]}, \sigma_c^2 I)$, définie par un vecteur référent $\mathbf{w}^{r[.]} = (w^{r[1]}, ..., w^{r[k]}, w^{r[n]})$, la matrice covariance, définie par $\sigma_c^2 I$ où σ_c est l'écart-type et I la matrice identité,

$$\mathcal{N}(\mathbf{x}^{r[.]}, \mathbf{w}^{r[.]}, \sigma_c^2 I) = \frac{1}{(2\pi\sigma_c)^{\frac{n}{2}}} \exp\left[-\frac{||\mathbf{w}_c^{r[.]} - \mathbf{x}_i^{r[.]}||^2}{2\sigma_c^2}\right] \tag{5.17}$$

- La fonction densité de la partie binaire qui est la distribution de Bernoulli $p(\mathbf{x}^{b[.]}/c) = p(\mathbf{x}^{b[.]}/\varepsilon_c, \mathbf{w}_c^{b[.]}) = f_c(\mathbf{x}^{b[.]}, \mathbf{w}_c^{b[.]}, \varepsilon_c)$ avec les paramètres $\mathbf{w}_c^{b[.]} = (w_c^{b[1]}, w_c^{b[2]}, ..., w_c^{b[k]}, ..., w_c^{b[m]}) \in \beta^m$ avec la probabilité $\varepsilon_c \in]0, \frac{1}{2}[$ associée à $\mathbf{w}_c^{b[.]} \in$

$\{0,1\}^m$ (§5.3.2). Le paramètre ε_c définit la probabilité d'être différent du prototype $\mathbf{w}_c^{b[.]}$. La distribution associée à chaque cellule $c \in \mathcal{C}$ est celle utilisée par le modèle BeSOM qui est définie comme suit :

$$f_c(\mathbf{x}^{b[.]}, \mathbf{w}_c^{b[.]}, \varepsilon_c) = \varepsilon_c^{\mathcal{H}(\mathbf{x}^{b[.]}, \mathbf{w}_c^{b[.]})}(1-\varepsilon_c)^{m-\mathcal{H}(\mathbf{x}^{b[.]}, \mathbf{w}_c^{b[.]})} \tag{5.18}$$

où la distance $\mathcal{H}$ mesure le nombre de composantes binaires différentes entre les deux observations $\mathbf{x}_i^{b[.]}$ et $\mathbf{x}_j^{b[.]}$.

Les paramètres $\theta = \theta^{\mathcal{C}} \cup \theta^{\mathcal{C}^*}$ qui définissent le modèle générateur de mélange sont constitués à la fois des paramètres de la distribution gaussienne et Bernoulli ($\theta^{\mathcal{C}} = \{\theta^c, c = 1..K\}$, où $\theta^c = (\mathbf{w}_c = (\mathbf{w}_c^{r[.]}, \mathbf{w}_c^{b[.]}), \sigma_c^2, \varepsilon_c)$), et la probabilité a priori aussi appellée coefficient de la mixture ($\theta^{\mathcal{C}^*} = \{\theta_{c^*}, c^* = 1..K\}$, où $\theta_{c^*} = p(c^*)$).

L'objectif maintenant est de redéfinir la fonction de coût ainsi que l'algorithme d'apprentissage associé qui permettra d'estimer ces paramètres. De la même manière que précédemment, l'algorithme d'apprentissage consiste à maximiser la vraisemblance des observations en appliquant l'algorithme EM. L'usage de l'algorithme EM s'explique par l'existence d'une variable cachée notée $\mathbf{z}$, constituée par le couple de cellules c et c^*, $\mathbf{z} = (c, c^*)$, impliquées dans la génération d'une donnée mixte observée $\mathbf{x}$.

Donc, de la même manière que le modèle dédié aux données binaires, nous pouvons définir la vraisemblance des données de la façon suivante :

$$V^T(\mathcal{A}, \mathbf{Z}; \theta) = \prod_{i=1}^{K} \prod_{c^* \in \mathcal{C}} \prod_{c \in \mathcal{C}} \left[p(c^*) p(c/c^*) \mathcal{N}(\mathbf{x}_i^{r[.]}, \mathbf{w}^{r[.]}, \sigma_{c_1}^2 I) \times f_c(\mathbf{x}_i^{b[.]}, \mathbf{w}_c^{b[.]}, \varepsilon_c) \right]^{z_i^{(c,c^*)}} \tag{5.19}$$

L'application de l'algorithme EM pour la maximisation de la vraisemblance des données observées nécessite d'une part l'estimation de $Q^T(\theta, \theta^t) = E\left[\ln V^T(\mathcal{A}, \mathbf{Z}; \theta)/\mathcal{A}, \theta^t\right]$, où θ^t est l'ensemble des paramètres estimés à la t^{e} itération de l'algorithme, et θ l'ensemble des paramètres recherchés.

L'étape "E (Estimation)" calcule l'espérance du log-vraisemblance par rapport aux variables cachées en prenant en considération les paramètres θ^{t-1}. A l'issue de l'étape "M (Maximisation)", la fonction $Q^T(\theta^t, \theta^{t-1})$ est maximisée par rapport à θ^t, ($\theta^t = \arg\max_\theta(Q^T(\theta, \theta^{t-1}))$). Ces deux étapes maximisent la fonction objective $Q^T(\theta^t, \theta^{t-1})$ où $\ln V^T(\mathcal{A}, \theta^t) \geq \ln V^T(\mathcal{A}, \theta^{t-1})$.

Il est clair que le modèle dédié aux données mixtes PrMTM est une généralisation du modèle BeSOM. Nous avons montré que la fonction objective est composée de deux termes correspondant aux fonctions objectives de la version probabiliste dédiée aux données continues et à la fonction objective du modèle BeSOM-ε_c. Les expressions permettant d'estimer les paramètres sont identiques respectivement au modèle BeSOM, présenté dans la section 5.3.2.2, et les formules mettant à jour la partie des variables continues :

$$\mathbf{w}_c^{r[.]} = \frac{\sum_{\mathbf{x}_i \in \mathcal{A}} \mathbf{x}_i^{r[.]} p(c/\mathbf{x}_i)}{\sum_{\mathbf{x}_i \in \mathcal{A}} p(c/\mathbf{x}_i)} \tag{5.20}$$

$$\sigma_c^2 = \frac{\sum_{\mathbf{x}_i \in \mathcal{A}} ||\mathbf{w}_c^{r[.]} - \mathbf{x}_i^{r[.]}||^2 p(c/\mathbf{x}_i)}{n \sum_{\mathbf{x} \in \mathcal{A}} p(c/\mathbf{x}_i)} \tag{5.21}$$

Exemple d'application

Nous illustrons dans cette partie l'application du modèle PrMTM sur une base de données artificielles composée de 1500 observations bidimensionnelles générées sous forme d'une structure de deux spirales. Dans cette base, la variable binaire dénote l'appartenance d'un point à l'une des deux spirales. L'objectif de cet exemple est de montrer que le modèle permet de construire des cartes auto-organisées en utilisant les deux types de variables.

Pour illustrer la phase d'apprentissage de notre modèle, nous avons entraîné trois cartes de dimensions différentes (1×5, 1×10 et 10×10). Comme attendu, les résultats de la carte prennent en compte les étiquettes et la structure spatiale des données. La figure 5.5 permet de s'assurer que la composante qualitative codée en binaire est cohérente avec les composantes quantitatives qui représentent les coordonnées des points.

Nous observons sur cet exemple "jouet" que notre modèle de cartes probabilistes respecte l'ordre topologique des données malgré le mélange de lois de probabilité. En effet, avec la carte 10×10, les prototypes ont une petite probabilité ε_c d'être différents de l'étiquette de la classe. Par contre, les ficelles 1×5 et 1×10 suivent la structure de la carte, mais avec une confiance faible exprimée par des probabilités fortes (le rayon du cercle est proportionnel à la probabilité ε).

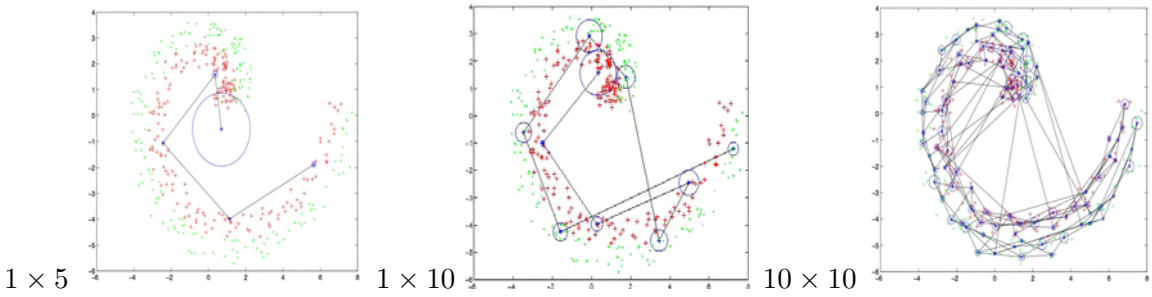


FIG. 5.5 – Carte PrMTM (1×5, 1×10, 10×10). Nous visualisons la partie des données quantitatives $\mathbf{w}_c^{r[1,2]}$ et la probabilité ε_c d'être différent de l'étiquette $\mathbf{w}_c^{b[3]}$. Le rayon du cercle est proportionnel à la probabilité ε_c.

5.3.4 Algorithme d'apprentissage

Les algorithmes d'apprentissage de CPrSOM, BeSOM et PrMTM sont l'application de l'algorithme EM aux modèles des cartes topologiques et auto-organisatrices probabilistes adaptées aux différents types de données. Cet algorithme permet d'estimer les coefficients du modèle de mélange. En supposant que les probabilités $p(c^*/c)$ sont fixées pour une valeur donnée de T, on aura alors l'algorithme d'apprentissage de l'algorithme à T fixé présenté ci-dessous. L'algorithme consiste à réitérer les deux étapes du formalisme EM. Si on note θ^t le vecteur de paramètres, estimé lors de la t^e itération et θ^{t+1} la mise à jour de ce vecteur, L'algorithme d'apprentissage pour les différents modèles pour une température T fixé se présente de la manière suivante :

---

1) **Initialisation**
   Partant d'une position θ^0, et d'un nombre d'itérations N_{iter},
2) **Itération de base**$(t \geq 1)$
   Ayant estimé les paramètres θ^t à l'itération précédente, l'itération en cours estime les nouveaux paramètres θ^{t+1}, et ceci, en appliquant les formules ci-dessous :

3) Calculer la probabilité a priori $p(c^*)$ selon l'expression (eq.5.9)
4) Selon la nature des variables, calculer les expressions suivantes :
   - CPrSOM : expression (eq.5.10)
   - BeSOM-ϵ : expressions (eq.5.12) et (eq.5.13).
   - BeSOM-ε : expressions (eq.5.15) et (eq.5.16).
   - PrMTM : expressions (eq.5.15), (eq.5.16), (eq.5.20) et (eq.5.21)

5) **Répéter** l'itération de base jusqu'à $t \geq N_{iter}$

Dans cet algorithme, nous avons présenté l'algorithme d'apprentissage utilisé pour estimer les paramètres maximisant la fonction log-vraisemblance à température fixe T. Par analogie à l'algorithme des cartes auto-organisatrices, on fait maintenant décroître la valeur de T entre deux valeur T_{max} et T_{min}. Pour chaque valeur de T, on obtient une fonction log-vraisemblance $Q^T(\theta, \theta^t)$. Ainsi, la fonction $Q^T(\theta, \theta^t)$ varie au cours des itérations avec la température T. Dans ce cas, on obtient deux catégories de fonctions :

- La première catégorie correspond aux grandes valeurs de T. Ces fonctions ont tendance à faire participer un très grand nombre d'observations à l'estimation des paramètres du modèle. Les valeurs obtenues représentent dans ce cas les paramètres des recouvrements de sous-ensembles définis par le voisinage d'influence de chaque cellule.

- La deuxième catégorie de fonctions correspond à des petites valeurs de T. Plus T est petit, plus le nombre d'observations est restreint, ainsi, la maximisation est locale.

L'algorithme général d'apprentissage pour une fonction particulière de décroissance de T se présente de la manière suivante :

Phase d'initialisation
Effectuer l'algorithme d'apprentissage pour la valeur de T constante égale à T_{max}, $t = 0$.

Etape itérative
L'ensemble des paramètres θ^t de l'étape précédente est connu. Calculer la nouvelle valeur de T en appliquant la formule suivante :

$$T = T_{max} \left(\frac{T_{min}}{T_{max}} \right)^{\frac{t}{N_{iter}-1}}$$

Pour cette valeur du paramètre T, calculer θ^{t+1} à l'aide des expressions correspondant au CPrSOM, BeSOM, PrMTM.

Répéter l'étape itérative jusqu'à atteindre $T = T_{min}$

5.4 Lien entre le modèle déterministe et le modèle de mélange

J'ai eu l'occasion de développer des modèles déterministes pour les données mixtes (MTM : Mixed Topological Map), [CI-17]. Récemment, j'ai proposé une généralisation du modèle déterministe en prenant en compte la pondération des variables [CN-3, RI-2]. Ce travail se base aussi sur une collaboration avec Nistor Grozavu, au cours de sa thèse, sur un modèle de pondération des variables continues [CI-7, CHN-1, CN-7, CN-9]. Etant donné que pour des vecteurs à composantes binaires la distance euclidienne n'est que la distance de Hamming $\mathcal{H}$, alors la distance euclidienne pour les données mixtes peut être réécrite comme suit :

$$||\mathbf{x} - \mathbf{w}_j||^2 = ||\mathbf{x}^{r[.]} - \mathbf{w}_j^{r[.]}||^2 + \mathcal{H}(\mathbf{x}^{b[.]}, \mathbf{w}_j^{b[.]}).$$

Utilisant cette expression, la fonction de coût de l'algorithme classique de Kohonen peut être exprimée comme suit :

$$\begin{aligned} \mathcal{G}(\phi, \mathcal{W}, \Pi) &= \sum_{\mathbf{x}\in\mathcal{A}} \sum_{j\in\mathcal{C}} \mathcal{K}(\delta(\phi(\mathbf{x}), j))(\pi_j^{r[.]})^\tau \mathcal{D}_{euc}(\mathbf{x}^{r[.]}, \mathbf{w}_j^{r[.]}) \rightarrow \mathcal{G}_{som}(\phi, \mathcal{W}, \Pi) \\ &+ \sum_{\mathbf{x}\in\mathcal{A}} \sum_{j\in\mathcal{C}} \mathcal{K}(\delta(\phi(\mathbf{x}), j))(\pi_j^{c[.]})^\tau \mathcal{H}(\mathbf{x}^{b[.]}, \mathbf{w}_j^{b[.]})] \rightarrow \mathcal{G}_{bin}(\phi, \mathcal{W}, \Pi) \end{aligned} \quad (5.22)$$

où $\mathcal{G}_{som}(\phi, \mathcal{W}, \Pi)$ représente la fonction de coût amélioré dans le travail présenté [CI-7] et $\mathcal{G}_{bin}(\phi, \mathcal{W}, \Pi)$ la fonction de coût dédiée aux données binaires. Cette dernière améliore la fonction de coût proposée au cours de ma thèse [CI-20].

La fonction de coût (eq.5.22) est minimisée en utilisant un processus itératif à trois étapes (affectation, quantification, pondération). Je ne vais pas détailler ici les trois étapes, mais je fournis uniquement les formules nécessaires à l'étape de quantification, qui vont nous servir à faire le lien avec le modèle probabiliste. Ainsi, supposant que ϕ et Π sont fixés, cette étape minimise $\mathcal{G}(\phi, \mathcal{W}, \Pi)$ par rapport à $\mathcal{W}$ dans l'espace $\Re^n \times \beta^m$. La minimisation de la fonction de coût (eq.5.22) nous mène à minimiser les deux termes de la fonction respectivement dans $\Re^n$ et dans β^m. Ces deux minimisations nous permettent de définir les formules suivantes :

- **la partie continue** $\mathbf{w}_j^{r[.]}$ du vecteur référent $\mathbf{w}_j$ comme le vecteur "moyenne" de la manière suivante :

$$\mathbf{w}_j^{r[.]} = \frac{\sum_{i\in\mathcal{C}} \mathcal{K}(\delta(i,j)) \sum_{\mathbf{x}\in\mathcal{A}, \phi(\mathbf{x})=r} \mathbf{x}^{r[.]}}{\sum_{i\in\mathcal{C}} \mathcal{K}^T(\delta(i,j)) n_i},$$

  où n_i représente le nombre correspondant d'observations affectées.
- **la partie binaire** $\mathbf{w}_j^{b[.]}$ du vecteur référent $\mathbf{w}_j$ comme le centre médian de la partie binaire des observations $\mathbf{x} \in \mathcal{A}$ pondérées par $\mathcal{K}(\delta(j, \phi(\mathbf{x})))$. Chaque composante $\mathbf{w}_j^{b[.]} = (w_j^{b[1]}, ..., w_j^{b[l]}, ..., w_j^{b[m]})$ est ensuite calculée de la manière suivante :

$$w_j^{b[l]} = \begin{cases} 0 & \text{si } \left[\sum_{\mathbf{x}\in\mathcal{A}} \mathcal{K}(\delta(j, \phi(\mathbf{x})))(1 - \mathbf{x}^{b[l]})\right] \geq \\ & \left[\sum_{\mathbf{x}\in\mathcal{A}} \mathcal{K}(\delta(j, \phi(\mathbf{x})))\mathbf{x}^{b[l]}\right] \\ 1 & \text{sinon} \end{cases},$$

L'analyse de ces formules indique que la mise à jour pour le centre médian $\mathbf{w}_j^{b[.]}$ et la partie réelle $\mathbf{w}_j^{r[.]}$ de notre modèle PrMTM coïncide avec le modèle déterministe et le modèle batch "nuées dynamiques" pour lesquels chaque observation $\mathbf{x}$ est pondérée proportionnellement par la fonction de voisinage ou une probabilité centrée sur le prototype gagnant.

Dans les deux modèles probabiliste et déterministe, nous minimisons l'inertie des observations $\mathbf{x}$ dans l'espace des variables continues et binaires ($\Re^n \times \beta^m$). Dans le modèle probabiliste PrMTM, la définition du gagnant est différente de celle du modèle déterministe. Dans le cas probabiliste, l'affectation est effectuée à la fin de l'algorithme d'apprentissage à l'aide de la probabilité a postériori. Notons également qu'il existe un lien fort entre le modèle déterministe des cartes topologiques dédiées aux données mixtes, et le modèle de mélange PrMTM. A chaque pas d'apprentissage, nous calculons la probabilité a posteriori $p(j/\mathbf{x})$ qui est utilisée pour pondérer l'observation $\mathbf{x}$. Dans le modèle déterministe, l'affectation est simplifiée en minimisant $((\pi_j^{r[.]})^\tau ||\mathbf{x}^{r[.]} - \mathbf{w}_j^{r[.]}||^2 + (\pi_j^{c[.]})^\tau \mathcal{H}(\mathbf{x}^{b[.]}, \mathbf{w}_j^{b[.]}))$.

Ainsi, si nous supposons que ε et σ représentent le même paramètre pour toutes les cellules et pour toutes les variables (binaires et quantitatives), et si les probabilités a priori $p(j^*) = \frac{1}{K}$ sont égales, alors une distance faible implique une probabilité $p(\mathbf{x}/j) = p(\mathbf{x}^{r[.]}/j) \times p(\mathbf{x}^{b[.]}/j)$ élevée. Nous déduisons donc que les probabilités a posteriori $p(j/\mathbf{x})$ et $p(j^*/\mathbf{x})$ deviennent élevées aussi.

Dans ces conditions, la maximisation de la fonction de coût $Q^T(\theta^t, \theta^{t-1})$ est la même que la maximisation de la fonction de coût $Q_1^T(\theta^{\mathcal{C}}, \theta^{t-1})$ qui est décomposée en deux termes. Le premier terme dépend uniquement de $\mathbf{w}_c$ et le deuxième terme dépend de la probabilité ε et de l'écart σ qui dépendent de la carte. Ainsi, maximiser $Q_1^T(\theta^{\mathcal{C}}, \theta^{t-1})$ par rapport à $\mathbf{w}_c$ nécessite la minimisation de la fonction de coût simplifiée $G(\mathbf{w})$ qui est définie par :

$$\begin{aligned} G(\mathbf{w}) &= \sum_{j \in \mathcal{C}} \sum_{\mathbf{x}_i \in \mathcal{A}} p(j/\mathbf{x}_i, \theta^{t-1}) \mathcal{H}(\mathbf{x}_i^{b[.]}, \mathbf{w}_j^{b[.]}) \\ &\quad + \sum_{j \in \mathcal{C}} \sum_{\mathbf{x}_i \in \mathcal{A}} p(j/\mathbf{x}_i, \theta^{t-1}) ||\mathbf{x}_i^{r[.]} - \mathbf{w}_j^{r[.]}||^2 \end{aligned} \tag{5.23}$$

On déduit que le modèle déterministe a tendance à rendre toutes les cellules équivalentes. La détermination de la cellule responsable de la génération j^* avec une fonction d'affectation nous permet de déduire que la fonction de coût $\mathcal{G}(\phi, \mathbf{w})$ (eq.5.22) n'est qu'une simplification de la fonction de coût $G(\mathbf{w})$ (eq.5.23). Il est clair que le modèle PrMTM probabiliste fournit plus d'information que le modèle déterministe.

5.5 Synthèse des expérimentations

Les modèles BeSOM, CPrSOM, PrMTM ont été implémentés en C/C++, puis ont été appliqués à diverses bases de données. Sur des bases réelles telles que des données issues de la Sofres ou d'une compagnie d'assurance. Nous avons montré dans [RI-3] que l'utilisation du modèle de mélange de données catégorielles (sans aucun codage) donnait des résultats très intéressants. Nous avons aussi étudié les performances de notre modèle de mélange dédié aux données binaires : au lieu de traiter les données catégorielles directement, nous avons utilisé un codage binaire et par conséquent une loi de probabilité adaptée à cette nature [CI-13, RI-5]. Nous avons aussi comparé le pouvoir de classement des modèles. Là encore, les performances

obtenues sont très satisfaisantes, en particulier grâce à la loi de probabilité adaptée aux données binaires, catégorielles ou mixtes [CN-3, RI-2]. Le modèle déterministe dédié aux données mixtes avec pondération des variables a été aussi validé [RI-2]. Dans un premier temps, des travaux ont été consacrés aux données continues. Sur ce sujet, des expérimentations permettant de prouver le pouvoir de caractérisation des clusters de cette approche ont été réalisées [CI-7, CHN-1, CN-7, CN-9].

Nous avons aussi étudié le lien entre les différents modèles. Les expériences réalisées montrent la compétitivité du modèle de mélange. L'avantage principal semble résider dans l'adaptation à la nature des données et des variables. Ces modèles ont été utilisés dans le cadre de projets ANR ou autres. C'était le cas dans le projet Infomagic [CI-9, CHN-2] où un brevet sur la recherche d'information visuelle a été déposé [BR-1]. D'autres travaux ont permis de valider certaines approches sur des données réelles de pollution issues du Centre Scientifique et Technique du Bâtiment [COL-1]. Ces modèles sont toujours utilisés dans le cadre de projets internes (au laboratoire LIPN) ou externes (principalement le laboratoire LOCEAN-UPMC).

Remarques pratiques

Dans la pratique, la mise en œuvre d'un algorithme EM avec les cartes topologiques nécessite certaines précautions. Comme tout algorithme, le nôtre doit être arrêté, il est donc nécessaire de choisir un test d'arrêt. L'idéal est de surveiller l'évolution de la vraisemblance, mais généralement, on choisit un nombre d'itérations. Comme nos algorithmes convergent à un minimum local, l'initialisation est très importante. La meilleure stratégie, à mon avis, est d'initialiser avec un modèle déterministe et de faire suivre avec quelques itérations avec le modèle probabiliste. Ceci permet de réduire la complexité du temps de calcul et d'espérer que la convergence qui s'ensuit correspondra au maximum global. Enfin, une autre manière d'accélérer la convergence est d'introduire une étape de classification entre l'étape "E" et l'étape "M". Evidemment, cette modification de l'algorithme EM est connue sous le nom CEM [CG92].

La complexité de notre algorithme d'apprentissage est de $O(NK^2)$. C'est K fois moins rapide que l'algorithme déterministe. Il est sûr que, pour les grandes bases de données, un temps de calcul important est nécessaire. Nous pouvons réduire la complexité en temps de calcul en limitant le voisinage autour de la cellule c^* pour limiter le nombre de couples (c, c^*). Une autre manière de réduire cette complexité est d'utiliser la version CEM en introduisant une étape d'affectation à chaque itération au lieu de l'affecter à la fin d'apprentissage.

5.6 Conclusions et perspectives

Quelques remarques pour conclure ce chapitre sur l'apprentissage avec des données catégorielles, binaires et mixtes. Ce chapitre présente une nouvelle formalisation des cartes auto-organisatrices pour la classification topologique et probabiliste de données binaires, catégorielles et mixtes multivariées (continues et qualitatives). Nous avons utilisé l'algorithme EM pour maximiser la vraisemblance des données afin d'estimer les paramètres du modèle de mélange. Nous n'avons pas choisi d'utiliser des variantes de l'algorithme classique en les appliquant à un codage particulier des données, mais de concevoir de nouveaux algorithmes qui tiennent compte de la nature des données elles-mêmes.

La question qui s'impose, c'est celle de l'utilité d'avoir à disposition des modèles qui s'adaptent à la nature des variables manipulées. Effectivement, certaines variables qualitatives, particulièrement les variables ordinales, peuvent être traités comme des variables continues, mais il n'est pas possible de traiter certaines variables qualitatives dans l'espace euclidien. Les modèles obtenus peuvent très certainement être utiles dans le cas où le problème d'apprentissage traite des données qualitatives, binaires ou mixtes. L'intérêt pratique de ces algorithmes réside dans le fait que l'extension des cartes topologiques auto-organisatrices, aux données binaires et catégorielles et mixtes, permet d'étendre le champ d'application de celles-ci à des domaines très variés : l'analyse sensorielle, la recherche d'informations et tout autre champ d'application dont les individus sont caractérisés par des variables binaires ou mixtes.

La deuxième question est le choix entre le modèle déterministe et le modèle probabiliste. Dans mes travaux de recherche, j'ai essayé de développer les deux approches, car tout dépend du problème à traiter. Souvent, un partitionnement rapide et plus fin est cherché ; ainsi, la faible complexité du modèle déterministe lui permet de fournir des résultats rapidement tout en respectant la nature des données. Dans d'autres problèmes, on a besoin de modèles plus robustes pour décrire la base étudiée ; ainsi, il est clair que le pouvoir de classification des modèles probabilistes est plus intéressant que les modèles déterministes.

Une approche sur laquelle je n'ai pas travaillé pour l'instant, mais qui semble digne d'intérêt, est celle de la classification croisée "co-clustering" lorsque les variables sont de nature différente. Nous pouvons envisager éventuellement l'extension des modèles BeSOM, CPrSOM et PrMTM à la classification croisée en utilisant le formalisme probabiliste des cartes topologiques. Des résultats élégants existent sur le sujet [NG10, NG08], mais à ma connaissance, il n'existe pas de co-clustering pour des bases de données ayant des variables mixtes. C'est une des pistes que je souhaite poursuivre pour cet axe en encadrant des étudiants.

Chapitre 6

Apprentissage non supervisé et données séquentielles

6.1 Introduction

Dans les approches présentées dans le chapitre précédent, une hypothèse est toujours faite qui est celle de considérer que les observations sont indépendantes et identiquement distribuées (i.i.d "independent and identically distributed"). La méthode la plus facile pour traiter les données séquentielles serait tout simplement d'ignorer l'aspect séquentiel et de traiter les observations comme des données i.i.d. Cependant, souvent, dans une situation réelle, des données sont collectées comme des observations de processus séquentiels, où l'hypothèse d'indépendance par rapport au temps ou les observations précédentes n'est pas toujours raisonnable. Par conséquent, développer un modèle non supervisé offrant la possibilité aussi de visualisation pour des données séquentielles est une perspective attrayante pour les applications de fouille de données complexes.

Les données séquentielles pourraient être générées à partir d'un grand nombre de sources telles que : le traitement de la parole, la fouille de texte, le séquençage de l'ADN, le diagnostic médical, les transactions des clients, la fouille de données web, l'analyse des capteurs en robotique. Mis à part le clustering qui peut se voir comme une instance de cette problématique, je n'ai pas eu encore le temps d'approfondir d'autres applications comme la prévision ou la détection de changement brusque dans les données séquentielles. Je projette néanmoins d'y consacrer une partie de mon activité future. Généralement, les algorithmes pour le clustering des données séquentielles sont essentiellement répartis en trois catégories.

- Similarité de séquence : La première catégorie est basée sur la mesure de la distance (ou similarité) entre chaque paire de séquences. Ensuite, l'utilisation des algorithmes de clustering, soit hiérarchique soit de partitionnement classique, peut être envisagée [KL83, BHR00].
- Partitionnement de séquences indirect : La seconde approche emploie une stratégie indirecte, qui commence par l'extraction d'un ensemble de caractéristiques à partir des séquences. Toutes les séquences sont alors projetées dans l'espace des variables transformées, où des algorithmes classiques de clustering peuvent être utilisés pour former des groupes. De toute évidence, l'extraction de caractéristiques devient le facteur essentiel qui détermine l'efficacité de ces algorithmes [GK01].

– Partitionnement statistique des séquences : Souvent, les deux premières approches sont utilisées pour traiter des données séquentielles composées d'alphabets, tandis que le troisième paradigme vise à construire des modèles statistiques pour décrire la dynamique de chaque groupe de séquences, et peuvent être appliqués aux séquences numériques ou catégorielles. La méthode la plus importante est celle basée sur les modèles de Markov cachés (HMM) [OFC01, OAB97, Smy97]. En effet, Les chaines de Markov cachées (HMMs) proposent une solution à ce problème en introduisant, pour chaque état, un processus stochastique sous-jacent qui n'est pas connu (caché), mais qui pourrait être déduit par les observations qu'il génère.

Etre capable d'apprendre d'une manière non supervisée les données séquentielles tout en fournissant des modèles de classification et de visualisation est un sujet de recherche extrêmement riche, et ceci est motivé par de nombreuses situations réelles. Mes travaux de recherche dans ce domaine se situent dans le troisième paradigme. Le formalisme développé dans ce chapitre se base sur les modèles de mélange et les HMM.

Je propose dans ce chapitre un modèles dédié à la classification et à la visualisation de données séquentielles. L'idée de base de l'approche est de considérer que les données séquentielles sont générées de la même manière que les HMMs en tenant compte de la "topologie" des états cachés. Ainsi, une relation spatio-séquentielle entre les états cachés est introduite. Les modèles de Markov cachés [Bau72] figurent parmi les meilleures approches adaptées aux traitements des séquences, étant donné leur capacité à traiter des séquences de longueurs variables et leur pouvoir à modéliser la dynamique d'un phénomène décrit par des suites d'événements. L'objectif de cette approche est de construire un nouveau modèle auto-organisé génératif d'un ensemble de données de séquences. La topologie ou l'auto-organisation introduite est inspirée des modèles de mélanges introduits au chapitre 5. Dans notre modèle, la génération d'une observation à un instant donné du temps est conditionnée par les états voisins au même instant du temps. Ainsi, une grande proximité implique une grande probabilité pour la contribution de la génération. Cette proximité est quantifiée en utilisant la fonction de voisinage. L'approche proposée est appelée Probabilistic Self-Organizing Map for Sequential data (PrSOMS). Le modèle aurait pu s'appeler self-organizing HMMs. En effet, ce modèle peut être vu sous deux angles différents. A notre connaissance, les modèles liant séquences, réduction de dimension et visualisation tout en manipulant les données dans l'espace multidimensionnel, sont rares. Les seuls travaux qui nous paraissent avancés dans ce sens et le modèle GTM, qui est souvent présenté comme la version probabiliste de la carte d'auto-organisation et qui a été étendu au traitement de séries chronologiques univariées (GTM through time) [BHS97, OV08] et aux données structurées [BMS10]. Récemment, dans [Yam10], l'auteur propose l'extension de l'algorithme SOMM (Self-Organizing Mixture Model, [VVK05]) pour les séries chronologiques multivariées (SOHMMs : self-organizing hidden Markov models). Cependant, la manière dont GTM et SOMM réalisent l'organisation topographique est tout à fait différente de celles utilisées dans nos approches.

Travaux similaires

Plusieurs techniques de classification automatique des données séquentielles ont été développées ces dernières années. Elles ont été appliquées dans différents domaines tels que la reconnaissance des caractères manuscrits [PMM+09], la reconnaissance de la parole [Rab89],

l'étude de la mobilité des objets dans les vidéos [BSK04] et l'analyse de séquences biologiques [OSJM09].

Les cartes auto-organisatrices présentées précédemment sont intéressantes de par leurs apports topologiques à la classification non supervisée et leurs capacités à résumer de manière simple un ensemble de données multidimensionnelles. Malheureusement, les paradigmes de l'auto-organisation ne peuvent pas être facilement transférés à des données non i.i.d. Différentes approches ont été développées pour intégrer l'information temporelle dans la carte d'auto-organisation. La façon la plus directe consiste à inclure les versions temporisées en entrée ou à ajouter un prétraitement pour la capture spatiale dynamique [Kan90, Wie03]. Une variété de modèles existe pour les cartes auto-récurrentes : la carte de Kohonen temporelle (TKM), la SOM récurrente (RSOM), la SOM récursive (RecSOM), et SOM pour les données structurées (SOMSD), [SH03, CT93, HSTM03, VMH97, Voe02]. Dans TKM et RSOM, chaque cellule est associée à un vecteur de poids de même dimension que les entrées. Dans chaque cellule, une observation est traitée à chaque pas de temps t dans le contexte donné par le passé des activations cette cellule. Quand une nouvelle entrée est présentée, les cellules ne perdent pas leurs activités passées immédiatement comme dans le SOM traditionnel, mais les informations du contexte se perdent progressivement. La décroissance progressive est contrôlée par un paramètre qui varie entre 0 et 1. Toutefois, RSOM modifie TKM en additionnant l'écart des poids, par opposition à des distances. Dans TKM, seule l'activité précédente de la cellule compte pour le calcul de son activité courante.

Il existe d'autres approches utilisant une distorsion temporelle dynamique. La DTW [Sak78] est un algorithme de programmation dynamique, où une séquence d'observations est souvent comparée à une autre séquence référence. L'association de telles références aux cellules de la carte et l'utilisation de la DTW dans la phase d'affectation ont donné lieu au modèle DTW-SOM [SK99, Som00]. D'autres modèles ont été proposés dans la littérature [HMSS04, AZ03, VHLM01]. Parmi ces modèles, nous pouvons citer Recursive SOM [Voe02] ; le modèle SOMSD (SOM for Structured Data) qui inclut l'index de la cellule gagnante précédemment dans le calcul du référent associé à chaque cellule ; le modèle MSOM (Merge SOM) qui inclut le contenu de la cellule gagnante précédemment dans le calcul du référent de toutes les cellules.

D'autres approches ont été proposées consistant à combiner les HMMs et les cartes auto-organisatrices (Self-Organizing Map) pour obtenir des modèles hybrides qui combinent le pouvoir de partitionnement des cartes auto-organisatrices et qui modélisent la notion de dynamique dans les séquences avec les HMMs. Dans [FS08b], les auteurs proposent un modèle hybride SOM-HMM, dans lequel chaque cellule de la carte représente un HMM. Cependant, le processus d'organisation n'est pas intégré explicitement dans l'approche HMM. D'autres approches ont été proposées pour combiner les HMMs et les cartes auto-organisatrices (Sel-Organizing Map) pour obtenir des modèles hybrides.

En fait, la modélisation graphique probabiliste motive différentes structures graphiques basées sur les HMMs [BF94, BB96, GJ97]. On trouve aussi dans la littérature plusieurs méthodes hybrides, composées des HMM avec les réseaux de neurones et plus particulièrement les cartes auto-organisatrices [BM93, MJ95]. Evidemment, il existe plusieurs structures probabilistes possibles qui peuvent être construites par rapport aux besoins des applications particulières. Les modèles graphiques fournissent un formalisme général pour décrire et analyser de telles structures. Afin de surmonter les limites des HMMs, des travaux récents dans [BT06, Bou08] proposent un nouveau paradigme supervisé, appelé *topological HMM*, qui

manipule les nœuds du graphe associé au HMM et ses transitions dans un espace euclidien. Cette approche modélise la structure locale d'un HMM et extrait leur forme en définissant une unité d'information comme une forme composée d'un groupe de symboles d'une séquence.

Les modèles auto-organisés pour le traitement des séquences utilisés dans la littérature peuvent être grossièrement divisés en trois groupes. Dans le premier groupe sont les méthodes trace d'affectations des cartes traditionnelles et des HMM. Ces deux modèles génèrent une séquence d'états, ou une séquence de distributions de probabilité d'état, qui peut être utilisée comme entrée pour un traitement ultérieur. Dans le second groupe, nous avons le RecSOM, TKM, RSOM et DTW-SOM, les opérateurs dans les cellules sont modifiés pour l'entrée séquentielle. Dans ces modèles, la structure des cartes est principalement utilisée pour partager des données entre les cellules concurrentes. En d'autres termes, l'espace des séquences est quantifié par les cartes et le processus de quantification est régularisé par le voisinage. Le troisième groupe comprend les modèles combinés, qui sont des hybridations des deux premiers groupes.

6.2 Le modèle probabiliste dédié aux données séquentielles (non i.i.d)

Nous supposons une séquence d'observations $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n, ..., \mathbf{x}_N\}$ telles que $\mathbf{x}_n$ est un élément de la séquence de taille N. La principale problématique est d'estimer les paramètres du modèle d'apprentissage PrSOMS. Le modèle proposé est l'extension du modèle de classification topologique probabiliste dédiée aux des données i.i.d (chapitre 5). On suppose que l'architecture de la grille modélise aussi un HMM qui est représenté par un treillis $\mathcal{C}$, qui a une topologie discrète définie par un graphe non orienté qui peut être en 2D ou 1D. On notera le nombre des cellules (nœuds, états) de $\mathcal{C}$ par K. Pour chaque paire d'états (c,r) dans le graphe, la distance $\delta(c, r)$ est définie comme la longueur de la plus courte chaîne qui lie les cellules c et r.

En s'inspirant des modèles probabilistes présentés au chapitre 5, on suppose que chaque élément $\mathbf{x}_n$ d'une séquence d'observations $\mathbf{X}$ est généré par le processus suivant : on commence par associer à chaque état $c \in \mathcal{C}$ une probabilité $p(\mathbf{x}_n/c)$. Par la suite, on sélectionne une cellule c^* de la carte $\mathcal{C}$ selon une probabilité a priori $p(c^*)$ pour laquelle on sélectionne une cellule $c \in \mathcal{C}$ selon la probabilité conditionnelle $p(c/c^*)$. Toutes les cellules $c \in \mathcal{C}$ et au même instant n, contribuent à la génération d'un élément $\mathbf{x}_n$ avec $p(\mathbf{x}_n/c)$ selon la proximité à la cellule c^* décrite par la probabilité $p(c/c^*)$.

Nous avons introduit deux variables binaires aléatoires comme variables cachées $\mathbf{z}_n$ et $\mathbf{z}_n^*$ de dimension K, dans lesquelles un élément particulier z_{nr} et $z^*_{nc^*}$ est égal à 1 et tous les autres éléments sont égaux à 0. Les deux composantes $z^*_{nc^*}$ et z_{nr} indiquent un couple d'états responsable de la génération d'un élément de la séquence. Utilisant cette notation, on peut réécrire la probabilité $p(\mathbf{x}_n/c)$ comme suit :

$$p(\mathbf{x}_n/c) \equiv p(\mathbf{x}_n/z_{nc} = 1) \equiv p(\mathbf{x}_n/\mathbf{z}_n)$$

et

$$p(c/c^*) = p(z_{nc} = 1/z^*_{nc^*} = 1) \equiv p(z_{nc}/z^*_{nc^*}) \equiv p(\mathbf{z}_n/\mathbf{z}_n^*)$$

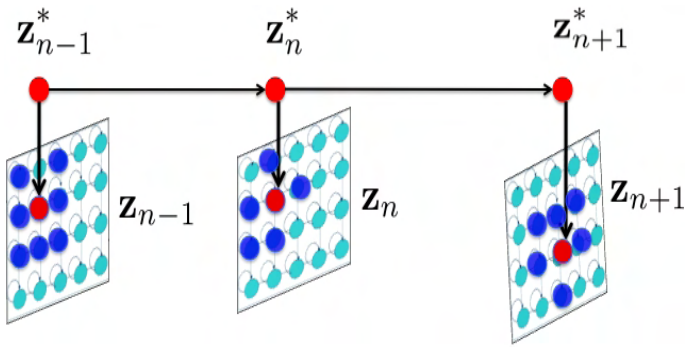


FIG. 6.1 – Un fragment du graphe représentant le modèle PrSOMS.

Pour introduire le processus d'auto-organisation dans l'apprentissage du modèle de mélange, on suppose que $p(z_{nc}/z^*_{nc^*})$ peut être définie de la même manière que les modèles des cartes probabilistes :

$$p(z_{nc}/z^*_{nc^*}) = \frac{\mathcal{K}^T(\delta(c,c^*))}{\sum_{r\in\mathcal{C}}\mathcal{K}^T(\delta(r,c^*))},$$

où $\mathcal{K}^T$ est la fonction de voisinage qui dépend du paramètre T (appelé température) : $\mathcal{K}^T(\delta) = \mathcal{K}(\delta/T)$. Ainsi, $\mathcal{K}$ définit pour chaque état de la chaîne de Markov $z^*_{nc^*}$ une région de voisinage dans le graphe $\mathcal{C}$. Le paramètre T permet de contrôler la taille du voisinage qui influence une cellule donnée de la carte $\mathcal{C}$ au même instant. Comme dans le cas de l'algorithme pour les données i.i.d, la valeur de T varie entre deux valeurs T_{max} et T_{min}.

On note l'ensemble de toutes les variables cachées par $\mathbf{Z}^*$ et $\mathbf{Z}$, où chaque ligne $\mathbf{z}^*_n$ et $\mathbf{z}_n$ est associée à chaque élément de la séquence $\mathbf{x}_n$. Chaque observation de la séquence en $\mathbf{X}$ est associée à un couple de variables cachées $\mathbf{Z}$ et $\mathbf{Z}^*$ responsables de la génération. On note par $\{\mathbf{X},\mathbf{Z},\mathbf{Z}^*\}$ l'ensemble complet des données, et on se réfère aux données observables $\mathbf{X}$ comme incomplètes. Ainsi, le modèle générateur d'une séquence est défini de la manière suivante :

$$p(\mathbf{X};\theta) = \sum_{\mathbf{Z}^*}\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z},\mathbf{Z}^*;\theta) \tag{6.1}$$

où θ est l'ensemble des paramètres.

Puisque la distribution $p(\mathbf{X},\mathbf{Z},\mathbf{Z}^*;\theta)$ ne peut pas se simplifier, une caractéristique importante pour les distributions des probabilités sur des variables multiples est celle de l'indépendance conditionnelle [Lut94]. On suppose que la distribution conditionnelle de $\mathbf{X}$, sachant $\mathbf{Z}^*$ et $\mathbf{Z}$, ne dépend pas de la variable cachée $\mathbf{Z}^*$. Souvent, cette hypothèse est utilisée pour les modèles graphiques, ainsi, $p(\mathbf{X}/\mathbf{Z},\mathbf{Z}^*) = p(\mathbf{X}/\mathbf{Z})$. Dans ce cas, la distribution jointe des observations de la séquence est égale à :

$$p(\mathbf{X},\mathbf{Z}^*,\mathbf{Z}) = p(\mathbf{Z}^*)p(\mathbf{Z}/\mathbf{Z}^*)p(\mathbf{X}/\mathbf{Z})$$

et on peut réécrire la distribution marginale comme :

$$p(\mathbf{X};\theta) = \sum_{\mathbf{Z}^*} p(\mathbf{Z}^*)\sum_{\mathbf{Z}} p(\mathbf{Z}/\mathbf{Z}^*)p(\mathbf{X}/\mathbf{Z}) \tag{6.2}$$

avec

$$p(\mathbf{X}/\mathbf{Z}^*) = \sum_{\mathbf{Z}} p(\mathbf{Z}/\mathbf{Z}^*)p(\mathbf{X}/\mathbf{Z}) \tag{6.3}$$

Paramètres du modèle topologique markovien

Considérant que la carte $\mathcal{C}$ représente un modèle de Markov, ainsi, la distribution à l'état $\mathbf{z}_n^*$ dépend de l'état de la variable latente précédente $\mathbf{z}_{n-1}^*$. Cette dépendance est représentée avec la probabilité conditionnelle $p(\mathbf{z}_n^*|\mathbf{z}_{n-1}^*)$. Puisque les variables latentes sont des variables binaires de dimension K, cette distribution conditionnelle correspond à une table de probabilité qu'on note par $\mathbf{A}$. Les éléments de $\mathbf{A}$ sont connus comme des probabilités de transition notées par

$$A_{jk} = p(z_{nk}^* = 1/z_{n-1,j}^* = 1) \text{ avec } \sum_k A_{jk} = 1$$

La matrice $\mathbf{A}$ a au maximum $K(K-1)$ paramètres indépendants. Dans notre cas, le nombre de transitions est limité par les cellules de la grille. On peut écrire la distribution conditionnelle explicitement sous cette forme :

$$p(\mathbf{z}_n^*/\mathbf{z}_{n-1}^*, \mathbf{A}) = \prod_{k=1}^{K}\prod_{j=1}^{K} A_{jk}^{z_{n-1,j}^* z_{nk}^*}$$

Toutes les distributions conditionnelles qui manipulent les variables cachées partagent les mêmes paramètres $\mathbf{A}$. L'état initial $\mathbf{z}_1^*$ est un cas particulier puisqu'il n'a pas de cellule parente, et ainsi il a une distribution marginale $p(\mathbf{z}_1^*)$ représentée par un vecteur de probabilités π avec les éléments $\pi_k = p(\mathbf{z}_{1k}^* = 1)$, ainsi que $p(\mathbf{z}_1^*|\pi) = \prod_{k=1}^{K} \pi^{z_{1k}^*}$, où $\sum_k \pi_k = 1$.

Les paramètres du modèle sont complétés en définissant les distributions conditionnelles des variables observées $p(\mathbf{x}_n/\mathbf{z}_n; \phi)$, où ϕ est un ensemble de paramètres qui définissent la distribution qui est connue comme des probabilités d'émission dans les modèles HMMs. Puisque $\mathbf{x}_n$ est observable, la distribution $p(\mathbf{x}_n/\mathbf{z}_n, \phi)$ consiste, pour une valeur donnée de ϕ, en un vecteur de K composantes qui correspondent aux K états possibles du vecteur binaire $\mathbf{z}_n$. On peut représenter les probabilités d'émission sous la forme suivante :

$$p(\mathbf{x}_n/\mathbf{z}_n; \phi) = \prod_{k=1}^{K} p(\mathbf{x}_n; \phi_k)^{z_{nk}}$$

La probabilité jointe des variables observables et les deux variables latentes $\mathbf{Z}$ et $\mathbf{Z}^*$ est exprimée par :

$$p(\mathbf{X}, \mathbf{Z}^*, \mathbf{Z}; \theta) = p(\mathbf{Z}^*; \mathbf{A}) \times p(\mathbf{Z}/\mathbf{Z}^*) \times p(\mathbf{X}/\mathbf{Z}; \phi)$$

$$p(\mathbf{X}, \mathbf{Z}^*, \mathbf{Z}; \theta) = \left[p(\mathbf{z}_1^*|\pi) \prod_{n=2}^{N} p(\mathbf{z}_n^*/\mathbf{z}_{n-1}^*; \mathbf{A})\right] \times \left[\prod_{i=1}^{N} p(\mathbf{z}_i/\mathbf{z}_i^*)\right] \times \left[\prod_{m=1}^{N} p(\mathbf{x}_m/\mathbf{z}_m; \phi)\right] \tag{6.4}$$

où $\theta = \{\pi, \mathbf{A}, \phi\}$ décrit l'ensemble des paramètres qui manipulent le modèle.

Il n'est pas évident de maximiser la fonction de vraisemblance, à cause de la complexité de l'expression. C'est pour cela qu'on utilise l'algorithme EM pour trouver les paramètres qui

maximisent la fonction de vraisemblance. L'algorithme EM commence avec quelques sélections initiales pour les paramètres du modèle, qu'on note par θ^{old}. Dans l'étape E (Estimation), on prend les valeurs des paramètres et on trouve la distribution a posteriori des variables latentes $p(\mathbf{Z}^*, \mathbf{Z}/\mathbf{X}, \theta^{old})$. Ensuite, on utilise cette distribution a posteriori pour évaluer l'espérance du logarithme de la vraisemblance des séquences complètes des données (eq.6.4), en fonction des paramètres θ, pour obtenir la fonction objective $Q(\theta, \theta^{old})$ définie par :

$$Q(\theta, \theta^{old}) = \sum_{\mathbf{Z}^*} \sum_{\mathbf{Z}} p(\mathbf{Z}^*, \mathbf{Z}/\mathbf{X}; \theta^{old}) \ln p(\mathbf{X}, \mathbf{Z}^*, \mathbf{Z}; \theta)$$

Après développement, on obtient la fonction suivante :

$$\begin{aligned} Q(\theta, \theta^{old}) &= \sum_{k=1}^{K} \gamma(z_{1k}^*) \ln \pi_k \rightarrow Q_1(\pi, \theta^{old}) \\ &+ \sum_{n=2}^{N} \sum_{k=1}^{K} \sum_{j=1}^{K} \xi(z_{n-1,j}^* z_n^*) \ln(A_{jk}) \rightarrow Q_2(\mathbf{A}, \theta^{old}) \\ &+ \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(\mathbf{x}_n; \phi_k) \rightarrow Q_3(\phi, \theta^{old}) \\ &+ \xi(z_{n-1,j}^* z_n^*) \ln p(\mathbf{Z}/\mathbf{Z}^*) \rightarrow Q_4 \end{aligned}$$

A cette étape, on va introduire quelques notations. On va utiliser $\gamma(\mathbf{z}_n^*, \mathbf{z}_n)$ pour noter la distribution marginale a posteriori des variables latentes $\mathbf{z}_n^*$ et $\mathbf{z}_n$, et $\xi(\mathbf{z}_{n-1}^*, \mathbf{z}_n^*) = p(\mathbf{z}_{n-1}^*, \mathbf{z}_n^*/\mathbf{X}, \theta^{old})$ pour noter la distribution a posteriori jointe des variables latentes successives. Nous définissons la distribution a posteriori de la variable latente $\mathbf{z}_n^*$ comme suit :

$$\gamma(\mathbf{z}_n^*) = \sum_{\mathbf{z}} p(\mathbf{z}_n^*, \mathbf{z}_n | \mathbf{X}; \theta^{old}) \text{ et } \gamma(\mathbf{z}_n) = \sum_{\mathbf{z}^*} p(\mathbf{z}_n^*, \mathbf{z}_n | \mathbf{X}; \theta^{old})$$

$$\gamma(z_{nk}^*) = \sum_{\mathbf{z}^*} \gamma(\mathbf{z}_n^*) z_n^k$$

On observe que la fonction objective $Q(\theta, \theta^{old})$ est définie comme une somme de quatre termes. Le premier terme $Q_1(\pi, \theta^{old})$ dépend des probabilités initiales ; le deuxième terme $Q_2(\mathbf{A}, \theta^{old})$ dépend des probabilités de transition $\mathbf{A}$; le troisième terme $Q_3(\phi, \theta^{old})$ dépend de ϕ qui est l'ensemble des paramètres de la probabilité d'émission, et le quatrième est une constante. La maximisation de $Q(\theta, \theta^{old})$ par rapport à $\theta = \{\pi, \mathbf{A}, \phi\}$ peut être effectuée séparément.

1) **La maximisation de $Q_1(\pi, \theta^{old})$: les probabilités initiales**
   De la même manière que les modèles probabilistes dédiés aux données i.i.d, une forme explicite de la distribution des probabilités initiales est utilisée. Ainsi, le paramètre de mise à jour est calculé de la manière suivante :

$$\pi_k = \frac{\gamma(z_{1k}^*)}{\sum_{j=1}^{K} \gamma(z_{1j}^*)} \tag{6.5}$$

2) **La maximisation de $Q_2(\mathbf{A}, \theta^{old})$: probabilités de transition**
Comme dans le cas des HMMs traditionnels, notre modèle utilise un état caché de valeur discrète avec une distribution multinomiale sachant les valeurs précédentes de l'état. Donc, notre modèle est un modèle du premier ordre. La mise à jour des paramètres est calculée de la manière suivante :

$$A_{jk} = \frac{\sum_{n=2}^{N} \xi(z^*_{n-1,j}, z^{*k}_n)}{\sum_{l=1}^{K}\sum_{n=2}^{N} \xi(z^*_{n-1,j}, z^*_{nl})} \tag{6.6}$$

où $\xi(z^*_{n-1,j}, z^*_{n,k}) = \mathbf{E}[z^*_{n-1,j} z^*_{n,k}] = \sum_{\mathbf{z}^*} \gamma(\mathbf{z}^*) z^*_{n-1,j} z^*_{n,k}$

3) **La maximisation de $Q_3(\phi, \theta^{old})$: les probabilités d'émission**
L'ensemble des paramètres ϕ dépend de la distribution utilisée. Nous présentons l'application en utilisant la loi gaussienne. Dans le cas des probabilités d'émission avec une densité sphérique gaussienne, on a $p(\mathbf{x}/\phi_k) = \mathcal{N}(\mathbf{x}; \mathbf{w}_k, \sigma_k)$, définie par son centre $\mathbf{w}_k$, qui a la même dimension que $\mathbf{x}$, et sa matrice de covariance, définie par $\sigma_k^2\mathbf{I}$ où σ_k est l'écart-type et $\mathbf{I}$ la matrice identité,

$$N(\mathbf{x}; \mathbf{w}_k, \sigma_k) = \frac{1}{(2\pi\sigma_k)^{\frac{d}{2}}} \exp\left[-\frac{||\mathbf{x} - \mathbf{w}_k||^2}{2\sigma_k^2}\right]$$

La maximisation de la fonction $Q_3(\phi, \theta^{old})$ fournit les expressions connues :

$$\mathbf{w}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{6.7}$$

$$\sigma_k^2 = \frac{\sum_{n=1}^{N} \gamma(z_{nk})||\mathbf{x}_n - \mathbf{w}_k||^2}{d\sum_{n=1}^{N} \gamma(z_{nk})} \tag{6.8}$$

où d est la dimension de l'élément $\mathbf{x}$.

De la même manière que le modèle de Markov caché, on va utiliser l'algorithme Baum-Welch [Bau72, Air07]. Dans notre cas, il peut être renommé par l'algorithme "Baum-Welch" topologique, puisqu'on utilise la structure du graphe pour organiser les données séquentielles d'une manière explicite. Dans la pratique, nous sommes aussi intéressés par trouver la séquence la plus probable d'états/cellules sur la carte PrSOMS, et cela peut être résolu efficacement en utilisant l'algorithme de Viterbi [Vit67, For73].

6.3 Analyse de l'auto-organisation

L'approche que nous proposons nous permet d'estimer les paramètres qui maximisent la fonction log-vraisemblance pour un paramètre T fixe. Nous pouvons constater clairement que sous l'hypothèse de données i.i.d notre modèle PrSOMS est équivalent au modèle présenté au chapitre 5. Ceci ce traduit par la disparition dans les calculs de la matrice de transition $\mathbf{A}$

puisque chaque état n'aura aucun état parent. Par conséquent sous cette condition la probabilité jointe (eq.6.4) sera équivalente à celle présenté dans le chapitre 5.

Comme dans le cas de l'algorithme de classification topologique probabiliste sous l'hypothèse de données i.i.d, on doit varier la valeur du paramètre T entre deux valeurs T_{max} et T_{min} pour contrôler l'influence du voisinage d'un état donné dans le graphe au même instant. Pour chaque valeur de T, on obtient une fonction de vraisemblance Q^T, et par conséquent l'expression varie avec T. Deux étapes sont définies :

- La première phase correspond à des valeurs élevées de T. Dans ce cas, l'influence du voisinage de chaque état $\mathbf{z}^*$ dans le graphe HMM, associé à notre approche, est importante et correspond à des valeurs plus élevées de la fonction $\mathcal{K}^T(\delta(c,r))$ et par conséquent des probabilités $p(\mathbf{z}/\mathbf{z}^*)$ forte. Ainsi, les formules décrites pour l'estimation des paramètres utilisent un grand nombre d'états au même instant et par conséquent utilisent un grand nombre d'observations pour estimer les paramètres du modèle. Cette phase permet d'obtenir l'ordre topologique du modèle de Markov.
- La deuxième étape correspond à des petites valeurs de T. A chaque instant, le nombre d'états est limité, donc, l'adaptation est très locale. Les paramètres sont calculés avec précision à partir de la densité des composantes. Dans ce cas, on peut considérer qu'on converge vers des HMMs traditionnels, puisque le voisinage est presque nul. En observant la probabilité jointe (eq.6.4), nous constatons que lorsque le voisinage d'un état $\mathbf{z}^*$ est réduit à lui même ($p(\mathbf{z}/\mathbf{z}^*) = 1$), la probabilité jointe coïncide avec la probabilité associé aux HMMs.

6.4 Synthèse des expérimentations

L'algorithme proposé et les visualisations ont été implémentés sous Matlab en utilisant la BNT toolbox [1] et la SOM toolbox [2]. L'approche a été testée sur des données réelles issues de l'INA. Le détail des résultats sera présenté dans la thèse de Rakia Jaziri. L'analyse détaillée des résultats montre que notre approche permet d'insérer une relation spatio-ordre (ou spatio-temporelle) entre les éléments d'une séquence [CI-2]. Dans [CI-1], nous avons présenté le modèle avec un exemple d'application réelle. Les données se composent de 2858 séquences d'écritures de lettre. Elles ont été capturées à l'aide d'une tablette WACOM, dont chaque point de la séquence est représenté en 3 dimensions. Cet exemple sera en partie détaillé par la suite. Les résultats présentés montrent que notre modèle fournit une généralisation des HMMs et une introduction claire de l'auto-organisation dans l'apprentissage du HMM. Nous sommes en train de préparer une soumission d'une revue détaillant le fonctionnement de notre modèle sur des séquences multivariables et sur les séries chronologiques unidimensionnelles. Une grande partie sera consacrée à l'application de notre modèle sur les échantillons issus de l'INA. Une nouvelle notion est aussi introduite et testée sur notre modèle PrSOMS. Dans [CI-2], nous avons présenté l'utilisation de la notion macro et micro HMM et le lien entre les deux approches. Chaque macro-état est composé de plusieurs micro-états. Cette notion de hiérarchie (macro et micro) dans les HMM sera abordée plus en détail dans la thèse de Rakia Jaziri.

[1] Bayes Net Toolbox for Matlab, http ://code.google.com/p/bnt/
[2] http ://www.cis.hut.fi/somtoolbox/

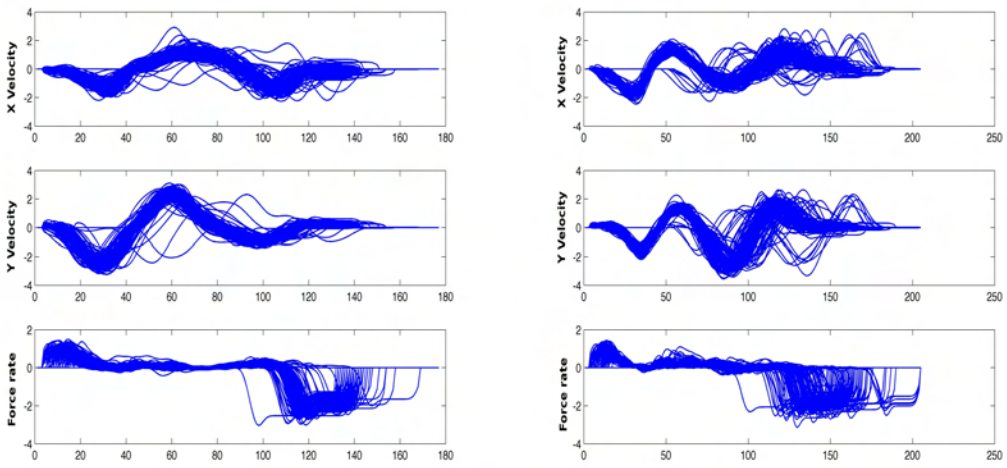
Afin d'illustrer notre modèle, j'ai choisi de montrer ici quelques résultats sur une base réelle issue du répertoire UCI [AN07] qui a été introduite par [Wil08].

Base des lettres manuscrites

Les données se composent de 2858 séquences. Elles ont été capturées à l'aide d'une tablette WACOM, où les 3 dimensions, x, y et la force de pointe du stylo, ont été conservées. Chaque caractère est une trajectoire de vitesse de pointe du stylo. Nous disposons pour chaque point de la vitesse en x, de la vitesse en y, et de la différence de la force du stylo. La séquence la plus probable est obtenue en utilisant l'algorithme de Viterbi.

Pour expliquer le fonctionnement de notre algorithme, nous l'avons appliquée séparément sur différents ensembles de données. Pour un souci de clarté, j'ai choisi volontairement de ne donner ici le détail des visualisations que sur un seul sous ensemble p. Je montre aussi quelques résultats sur l'ensemble q et $\{a, b, c\}$. La figure 6.2 montre une superposition de 131 échantillons de p (première colonne) et 124 échantillons de la lettre q (deuxième colonne) dans l'espace des vitesses. Toutes les séquences sont considérées dans un espace multi-variables. Chaque composante est représentée par trois variables dans l'espace des vitesses. Il est facile de retrouver les trois composantes dans l'espace de la tablette (coordonnées x et y et la force du stylo).

La figure 6.3 indique différentes visualisations qui peuvent êtres envisagées avec notre



(a) lettre p (b) lettre q

FIG. 6.2 – Superposition de 131 échantillons de p (première colonne) et 124 échantillons de la lettre q (deuxième colonne) dans l'espace des vitesses

modèle PrSOMS. L'apprentissage a été réalisé avec uniquement les échantillons de la séquence "p" avec une grille de dimension 12×12. La figure 6.3(a) présente la cardinalité associée à chaque cellule de la carte PrSOMS. Celle-ci est calculée après avoir affecté chaque séquence en utilisant l'algorithme Viterbi. Ainsi, notre modèle PrSOMS permet de faire un clustering des données en tenant compte de l'ordre ou de la propriété non-i.i.d. La taille du carré est

proportionnelle aux composantes des séquences captées. Dans la figure 6.3(b), on visualise les profils associés à chaque cellule/état de la carte. En fait, nous visualisons dans chaque cellule à la fois les trois composantes du profil (vitesse en x, y et la différence de force du stylo). Il est clair qu'il existe un ordre topologique ; en fait, chaque région de la carte est représentée par des profils similaires. La figure 6.3(c) représente la projection des échantillons "p" visualisés aussi dans l'espace des états latents ou cellules de la carte. Les points en bleu présentent les éléments composant les séquences originales. Nous observons une organisation topologique claire de la carte. Les observations proches dans l'espace des données sont aussi captées par des états proches sur la carte. Ces projections fournissent une visualisation topographique de l'ensemble des échantillons séquentiels. Cette propriété se confirme en visualisant aussi tous les chemins de Viterbi calculés avec toutes les lettres, figure 6.3(d). Nous pouvons clairement observer que l'ordre topologique des états ou des cellules respecte l'ordre topologique de l'ensemble des éléments de la séquence : des éléments proches sont représentés/captés par des états ou des cellules proches. En utilisant les profils trouvés avec les chemins de Viterbi les plus probables (figures 6.3(d) et 6.3(b)), nous pouvons reconstruire toutes les lettres p. La figure 6.3(e) présente en bleu la séquence originale et en rouge la séquence reconstruite à l'aide des profils trouvés avec l'algorithme de Viterbi.

D'autres analyses peuvent êtres réalisées avec la carte PrSOMS. La figure 6.4(a) présente dans chaque cellule en couleur rouge toutes les composantes des séquences captées et en bleu les autres composantes qui n'ont pas été captées. Ceci est dans l'objectif de visualiser la région ou la partie de la séquence qui est captée par chaque état. La figure 6.4(b) présente un agrandissement des cellules de numéro 1, 55, 144. La figure 6.4(c) représente les échantillons d'origine dans l'espace de la tablette en indiquant par une couleur le numéro de la cellule la plus probable obtenue dans le chemin de Viterbi (figure 6.3(d)). La figure 6.4(d) présente les même échantillons dans l'espace des vitesses (vitesses en x, y et la différence de force). Chaque couleur correspond au numéro de l'état le plus probable (de 1 à 144) fourni par l'algorithme de Viterbi. La figure 6.5(a) représente les trois séquences constituant un seul échantillon p. La couleur bleue indique la vitesse en x, la couleur verte indique la vitesse en y et la couleur rouge indique la différence de la force de pression du stylo. En pointillé, pour chaque couleur, on représente le signal reconstruit pour chaque composante. La figure 6.5(b) représente l'échantillon original et reconstruit dans l'espace de la tablette.

Pour montrer la performance du modèle génératif, nous avons reconstruit la lettre q en utilisant la carte p-PrSOMS apprise avec les séquences associées à la lettre q et vice versa. Les figures 6.6(a) et 6.6(b) montrent, dans les deux cas, que chaque modèle offre une bonne reconstruction de la partie commune des deux lettres p et q. En effet, l'approche PrSOMS arrive à détecter des régions communes lors de l'écriture des lettres.

La figure 6.7 indique 4 étapes de l'apprentissage de la carte PrSOMS avec l'ensemble $\{a, b, c\}$. Ces étapes illustrent les phases d'auto-organisations indiquées dans la section 6.3. La figure 6.8 montre des échantillons reconstruits en utilisant les mêmes paramètres à la fin de l'apprentissage. Pour chaque figure, sur la gauche sont visualisés les caractères originaux et sur la droite sont présentés les caractères reconstruits. Chaque caractère est représenté dans l'espace de la tablette (x, y et la pression du stylo). La couleur indique la valeur de la pression du stylo. La reconstruction en utilisant Viterbi produit une reconstruction très proche du jeu de données. Les caractères sont bien reconstruits, et facilement reconnaissables.

6.5 Conclusions et perspectives

L'apprentissage de données séquentielles est un domaine très abordé en recherche et reste ouvert sur certaines problématiques comme la prise en compte de la structure avec le temps ou l'ordre. Souvent, on appelle ce domaine par l'apprentissage de données non i.id. Construire des méthodes de partitionnement/clustering et de visualisation est d'une très grande importance pour les utilisateurs et les experts du domaine d'application. J'ai essayé d'apporter ma contribution avec la doctorante Jaziri Rakia en définissant un modèle générateur de données séquentielles en se basant sur les chaînes de Markov cachées. Ceci nous a permis de définir un modèle de HMM auto-organisé. La contrainte qui a toujours guidé mes recherches est celle d'avoir des prototypes ou des lois de probabilité adaptés aux données.

Je me suis concentré dans cet axe sur le développement de modèles de classification et de visualisation, car vu la quantité et la complexité croissante de données séquentielles qui posent des problèmes pour les experts, ils ne peuvent pas s'appuyer uniquement sur des techniques traditionnelles entièrement automatiques.

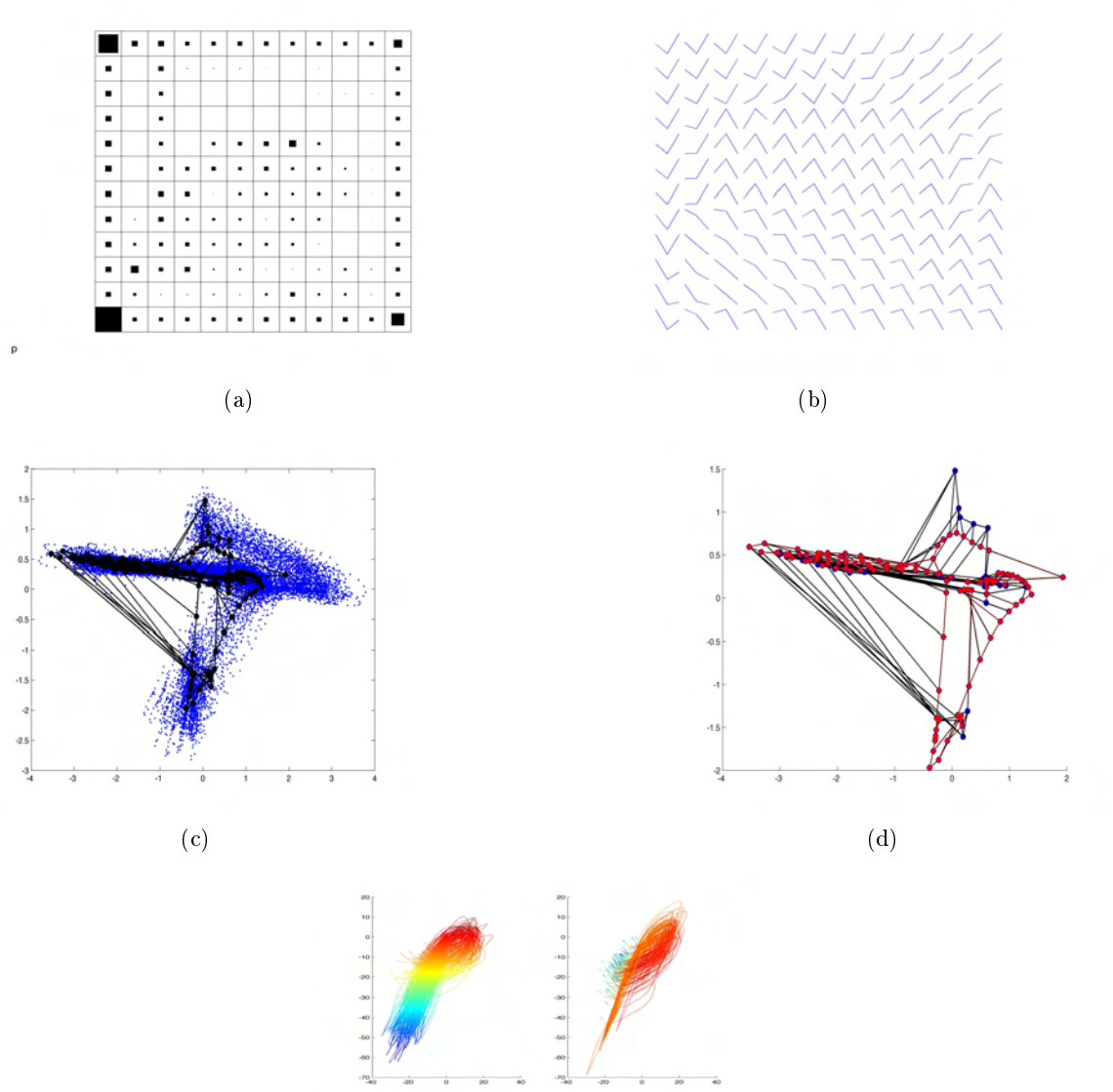
Par analogie aux modèles dédiés aux données i.i.d, on peut proposer de segmenter le premier résultat de classification pour construire une hiérarchie [VA00]. En effet, dans ce cas, nous introduisons la notion de macro-état qui est considéré comme un état fédérateur constitué de micro-état (les états de la carte). Cette démarche a l'avantage de proposer une architecture globale du HMM $\equiv$ PrSOMS appelée Macro-HMM. L'expert a la possibilité d'utiliser le Macro-HMM ou d'utiliser une information plus fine avec le micro-HMM [CI-2]. Cette procédure permet de hiérarchiser l'information et ceci semble d'une grande importance pour l'application de l'INA[3] abordée dans la thèse de Rakia (Apprentissage non supervisé de données structurées en séquences). Un autre point qui est abordé dans la thèse de Jaziri Rakia est la prise en compte du long terme. Nous sommes en train d'étudier la possibilité d'utiliser des modèles auto-regressifs afin de prendre en compte le temps dans le cas de l'application visée avec l'INA. Nous veillons toujours à ce que nos modèles soient les plus généraux possible afin de pouvoir les exploiter dans d'autres applications.

D'autres points reste ouverts comme la classification incrémentale dédiée aux séquences [FLBH05]. Ceci correspond à un système capable de recevoir et d'intégrer de nouveaux exemples sans devoir réaliser un apprentissage complet. Cet axe concerne le développement de modèles d'apprentissage non-supervisé visant à créer par apprentissage des groupes homogènes (ou typologies) en fonction de l'évolution temporelle ou d'ordre des données séquentielles. Ces modèles permettront de découvrir un espace topologique d'un ensemble de données.

Une question qui vient à l'esprit est : pourquoi traiter toutes les séquences ayant des composantes de nature différente de la même manière. Pourquoi ne pas apprendre sur des parties de la base décrites par des variables de même nature, puis utiliser des méthodes ensemblistes pour fusionner les résultats de classification. Concernant, les données i.i.d, la recherche est bien avancée [TJP05, TJP04, FPS02, SG02]. J'ai eu l'occasion de faire une tentative pour les données i.i.d (non séquentielles), et cela m'a convaincu qu'il faudrait collaborer entre les différentes classifications [CI-9, CHN-2]. Je n'ai pas eu l'occasion de développer cet axe, mais il me semble qu'il est nécessaire pour les données séquentielles, car vu l'augmentation des données, la nécessité de partager les calculs me paraît obligatoire. Des

[3] Institut National de l'Audiovisuel

proportions très intéressantes de modèles existent dans la littérature, en particulier pour les données séquentielles [DL04, Mac97]. Il me semble qu'il est nécessaire d'introduire la notion de collaboration ou de fusion. On peut supposer avoir deux modèles "corrects " qui peuvent collaborer pour améliorer l'estimation de leurs paramètres. Ceci sera ma priorité concernant cet axe dans le futur proche.

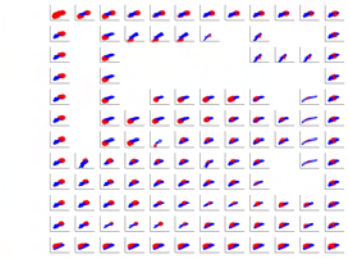


(a) (b)

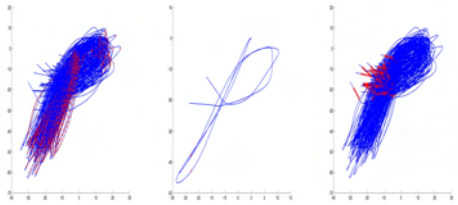
(c) (d)

(e)

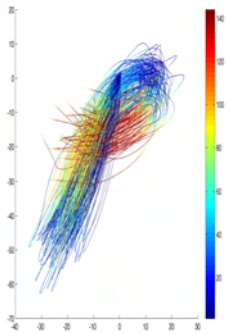
FIG. 6.3 – Apprentissage de la carte 12×12 avec la lettre p. 6.3(a) Cardinalité associée à la carte PrSOMS. La taille du carré est proportionnelle aux composantes captées en appliquant Viterbi. 6.3(b) Carte des profils (3 composantes : vitesses en x et y et la différence de force). 6.3(c) Projection ACP des échantillons et des profils associés à chaque cellule/état de la carte p-PrSOMS. 6.3(d) Le chemin de Viterbi (en rouge) correspondant à tous les échantillons. 6.3(e) Reconstruction de toutes les lettres p en utilisant les profils de la carte PrSOMS.



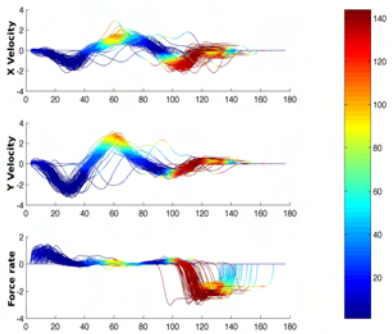
(a)



(b) Zoom : cellule 1, 55, 144

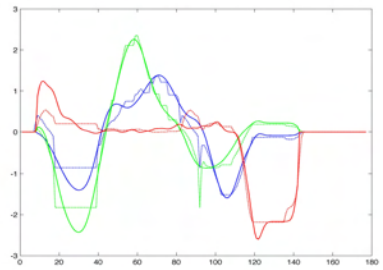


(c)

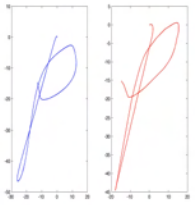


(d)

FIG. 6.4 – Apprentissage de la carte 12 × 12 avec la lettre p. 6.4(a) Chaque cellule visualise en rouge toutes les composantes captés et en bleu les autres composantes. 6.4(b) Zoom sur la cellule 1, 55, 144. 6.4(c) : Les échantillons originaires en indiquant par une couleur le numéro de la cellule affectée. 6.4(d) Les échantillons originaires dans l'espace des vitesses. Chaque couleur correspond au numéro de l'état le plus probable (de 1 à 144) fourni par l'algorithme de Viterbi

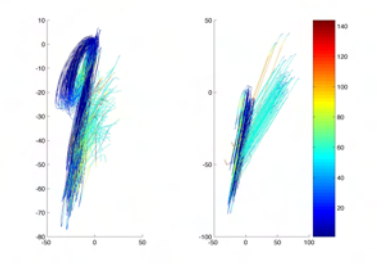


(a) bleu : la vitesse en x vert : la vitesse en y ; rouge : la différence de la force de pression

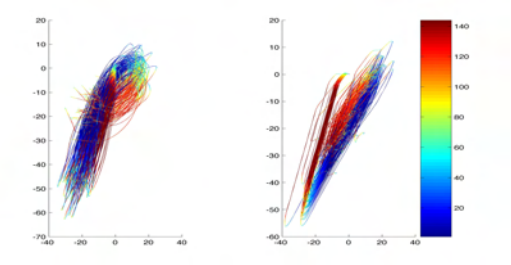


(b) bleu : originaire ; rouge : reconstruit

FIG. 6.5 – Reconstruction d'un seul exemple de la lettre p. 6.5(a) Les trois composantes de l'exemple p représentées dans l'espace des vitesses. En pointillé, pour chaque couleur, on représente le signal reconstruit. 6.5(b) présente la séquence originale et reconstruite avec les profils.

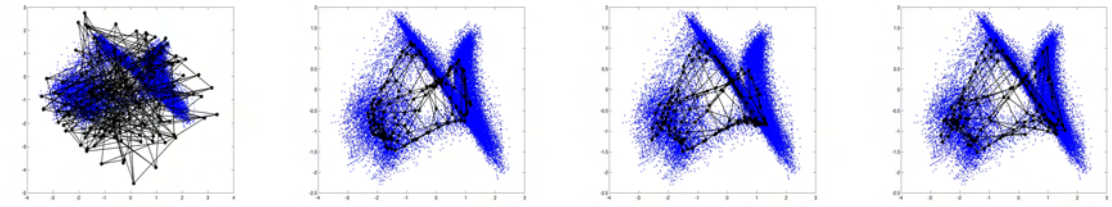


(a) Reconstruction de la lettre q en utilisant la carte p-PrSOMS



(b) Reconstruction de la lettre p en utilisant la carte q-PrSOMS

FIG. 6.6 – Reconstruction des échantillons en utilisant le modèle PrSOMS. Le niveau de couleur indique le numéro de la cellule fourni par l'algorithme de Viterbi ; à gauche de chaque sous-figure, nous avons la séquence originale.



(a) Random Initialization, $T = 5$

(b) iteration : 5 , $T = 2.57$

(c) iteration : 15 , $T = 1.52$

(d) iteration : 20 , $T = 1$

FIG. 6.7 – Configuration de la carte *abc*-PrSOMS 12×12 après 5 ; 15 ; 20 itérations.

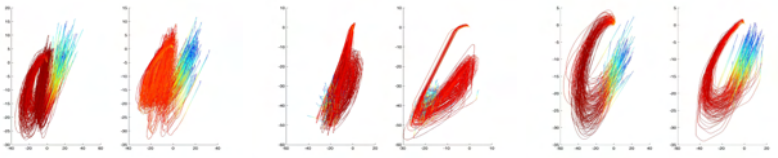


FIG. 6.8 – Reconstruction des caractères en utilisant la carte *abc*-PrSOMS 12×12. La couleur indique la valeur de la pression du stylo. Les caractères originaux sont indiqués à gauche de chaque figure ; à droite, les caractères reconstruits.

Bilan et perspectives

L'apprentissage non supervisé traite ou explore les données non étiquetées, soit par la construction de structure topologique, hiérarchique, soit en formant une partition de groupes ou de clusters. Ce processus comprend une série d'étapes, allant du prétraitement et de l'algorithme de traitement à l'évaluation. Chacune des étapes est étroitement liée aux autres et présentent des défis en apprentissage non supervisé. Dans ma recherche, j'ai mis l'accent sur les algorithmes de clustering/partitionnement/classification non supervisé et j'ai examiné particulièrement les modèles auto-organisés avec des approches déterministes et à base de modèles de mélanges. Les algorithmes proposés visent à résoudre des problèmes différents par le type de données traitées (données catégorielles, binaires, mixtes, séquences), et ont leurs propres avantages et inconvénients. Bien que avec les différents collaborateurs nous ayons déjà vu, nos algorithmes, en validation sur de nombreux exemples, il reste encore de nombreux problèmes ouverts en raison de la diversité et de la complexité des données qu'il reste à explorer. Les problèmes abordés dans ce mémoire ont déjà attiré et continueront d'attirer des efforts intenses en recherche. Je pense que certains chercheurs de cette discipline doivent avoir une vision assez précise des aspects algorithmiques et d'autres avoir une vision des aspects théoriques, mais je pense qu'il est important de s'intéresser à des problématiques motivées par des applications réelles. Beaucoup de problèmes réels vont au-delà du simple partitionnement, notamment, celui de travailler avec des téraoctets de données est en soi un grand défi.

L'encadrement de doctorants et d'étudiants en master 2 a été une expérience enrichissante d'un point de vue scientifique et humain. Participer à l'animation du groupe de travail "fouille de données complexes" m'a permis de me rendre compte de la multiplicité des données et des algorithmes utilisés en apprentissage non supervisé. Finalement, l'une de mes activités importantes, c'est l'enseignement et la transmission de connaissances.

Perspectives

Il n'existe pas d'algorithme de clustering/partitionnement qui peut être universellement utilisé pour résoudre tous les problèmes et tous les types de données. Par contre, on peut trouver des modèles adaptés à différents types de données. Habituellement, les algorithmes sont conçus avec certaines hypothèses et avec des objectifs bien précis : classement, classification ou partitionnement, prédiction, régression. En particulier pour le partitionnement, à mon avis, il n'existe pas de "meilleurs" algorithmes, bien que certaines comparaisons soient possibles. Ces comparaisons sont principalement basées sur certaines applications spécifiques, sous certaines conditions, et les résultats peuvent devenir tout à fait différents si les conditions changent. La

meilleure validation est de collaborer avec les experts du domaine des applications visées.

Mon activité de recherche s'articule autour de plusieurs axes, encouragés par des collaborations en cours, des résultats récents.

A court terme

Classification topologique de données séquentielles à base de modèles de mélanges.

Une question sur laquelle je travaille avec Rakia Jaziri, que j'encadre en thèse, est celle du développement de modèle de classification et de visualisation dédié aux données séquentielles. Nous travaillons sur des modèles HMMs capables de s'auto-organiser. Une version a été présentée dans ce mémoire : elle a été validée avec succès sur des données issues de l'INA. Nous pensons que cette version peut nous aider à proposer une architecture globale du HMM traditionnel. Ceci est possible en utilisant les modèles hiérarchiques.

Un problème qui nous intéresse tout particulièrement, est celui d'introduire la notion du long terme. Lors d'une conférence internationale, un chercheur japonais m'a fait la remarque que le modèle des HMMs tel qu'on le présente peut inclure d'une manière implicite le long terme. Il suffit de rajouter une contrainte dans la fonction de voisinage de telle manière à réduire le nombre de couples (c, c^*) ou le nombre de liens à la cellule centrale c^*. C'est une piste de recherche qui est intéressante, et nous disposons de tous les outils théoriques et pratiques pour la valider. Nous espérons aboutir à un algorithme d'apprentissage adaptatif respectant les objectifs que nous avons fixés.

Lors de son passage au LIPN en tant que ATER, M. Faicel Chamroukhi, aujourd'hui maître de conférences à Toulon, nous a proposé d'étendre le modèle GTM-TT (GTM Through Time, [BHS97, OV08]) qui fonctionnait sur les séries chronologiques aux séquences multidimensionnelles en utilisant le même principe que PrSOMS. Nous avons déjà commencé à définir le modèle théorique ; il nous reste à le programmer et à comparer aux résultats disponibles. Comme je l'avais souligné, ces travaux sont déjà liés à mon activité avec Rakia Jaziri sur le traitement des séquences.

Apprentissage non supervisé et graphes. Ces travaux ne sont qu'au début du chemin et des analyses approfondies de l'approche sont nécessaires particulièrement sur la convergence de l'algorithme et le passage à l'échelle. L'exploitation de la notion de "leader/hub" pour définir un espace de représentation fait partie aussi de mes priorités. J'avance dans cet axe en collaboration avec Hanane Azzag qui est maître de conférences dans mon équipe, avec laquelle j'encadre Nhat-Quang Doan sur ce sujet. L'aspect visualisation est très important dans ce domaine ; on ne peut pas discuter de grands graphes sans se poser la question de comment les visualiser. Ce travail de recherche est particulièrement intéressant puisqu'il soulève des questions algorithmiques, de manipulation de graphes et d'optimisation.

A long terme

Apprentissage de données non-i.i.d. Comme je l'avais signalé précédemment, le thème d'apprentissage sur des données non i.i.d est un thème sur lequel je vais concentrer une partie de ma recherche. Je vais évidemment continuer à travailler sur les données séquentielles, avec comme objectifs de les étendre aux données structurées en graphes dynamiques. De plus, mon

objectif est de réduire le nombre de paramètres des algorithmes proposés. C'est vraiment l'un des thèmes de recherche qui anime un certain nombre de workshop et de congrès internationaux traitant la nouvelle thématique appelée "Autonomius learning".

Lors de la phase de pré-traitement et de post-traitement, les fonctions de sélection/extraction (ainsi que la normalisation) et de validation sont aussi importantes que les algorithmes de clustering. Choisir les variables appropriées peut réduire considérablement la complexité du problème. Le compromis entre différents critères et les méthodes est toujours dépendant des applications. Les travaux antérieurs, que j'ai menés avec Nistor Grozavu maître de conférences dans mon équipe, sur la caractérisation et la sélection non supervisée des variables sur des données i.i.d, peuvent être transférés dans l'apprentissage de données structurées en séquences ou en graphes.

Apprentissage incrémental distribué nécessite d'avoir la capacité de traiter une nouvelle donnée sans réapprendre tout le modèle, ceci résoudra en partie le problème du traitement de grands volumes de données. Estimer le nombre de groupes/clusters et apprendre d'une manière incrémentale, c'est la tendance actuelle de tous les algorithmes. Il est clair que cet axe d'apprentissage est commun à tous les thèmes abordés dans ce mémoire. Je suis impliqué dans un projet ANR qui se donne le défi d'apprendre des modèles à partir de téraoctets de données avec la difficulté supplémentaire d'être des données avec des classes déséquilibrées. Une des pistes que je souhaiterais aborder à long terme est l'apprentissage distribué. A mon avis, ça fait partie des directions qu'il faudrait explorer pour pouvoir manipuler des quantités de données en forte croissance. Traiter un volume important de données représentées dans un espace à grande dimension avec des délais d'exécution acceptables ; à mon avis ce point sera un des challenges des nouveaux algorithmes. L'aspect apprentissage distribué ou collaboratif sera l'un des axes d'où certaines solutions pourront émerger.

Je veille à ce que tous les algorithmes que je développe, respectent la structure initiale des données et qu'ils fournissent aux utilisateurs des visualisations qui peuvent simplifier l'analyse ultérieure des données. Beaucoup de personnes travaillant en apprentissage automatique négligent l'aspect visualisation, qui est de plus en plus primordial en fouille de données.

Enfin, des problématiques pour des applications spécifiques ou des rencontres influent et influeront sûrement sur mes directions de recherche. L'enjeu dans ces moments-là, c'est de trouver des thématiques qui nous rassemblent pour identifier des problématiques qui nous intéressent.

Bibliographie

[AAW06] Bill Andreopoulos, Aijun An, and Xiaogang Wang. Bi-level clustering of mixed categorical and numerical biomedical data. *International Journal of Data Mining and Bioinformatics*, 1(1) :19 – 56, 2006.

[ABT97] Fatiha Anouar, Fouad Badran, and Sylvie Thiria. Self-organizing map, a probabilistic approach. In *Proceedings of WSOM'97-Workshop on Self-Organizing Maps, Espoo, Finland June 4-6*, pages 339–344, 1997.

[ABT98] Fatiha Anouar, Fouad Badran, and Sylvie Thiria. Probabilistic self-organizing map and radial basis function networks. *Neurocomputing*, 20(1-3) :83–96, 1998.

[AD07a] Amir Ahmad and Lipika Dey. A k-mean clustering algorithm for mixed numeric and categorical data. *Data Knowl. Eng.*, 63 :503–527, November 2007.

[AGV06a] Hanene Azzag, Christiane Guinot, and Gilles Venturini. Data and text mining with hierarchical clustering ants. In *Swarm Intelligence in Data Mining*, pages 153–189. 2006.

[Air07] Edoardo M Airoldi. Getting started in probabilistic graphical models. *PLoS Comput Biol*, 3(12) :e252, 12 2007.

[AN07] A. Asuncion and D.J. Newman. UCI machine learning repository. http ://www.ics.uci.edu/∼mlearn/MLRepository.html, 2007.

[And02] Péter András. Kernel-kohonen networks. *International Journal of Neural Systems*, 12 :117–135, 2002.

[Aup05] Michaël Aupetit. Learning topology with the generative gaussian graph and the em algorithm. In *NIPS*, pages 592–598, 2005.

[AZ03] Khalid Benabdeslem Arnaud Zeboulon, Younès Bennani. Hybrid connectionist approach for knowledge discovery from web navigation patterns. In *ACS/IEEE International Conference on Computer Systems and Applications*, pages 118–122, 2003.

[Bau72] L. E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3 :1–8, 1972.

[BB96] Samy Bengio and Yoshua Bengio. An EM algorithm for asynchronous input/output hidden Markov models. In L. Xu, editor, *International Conference On Neural Information Processing*, pages 328–334, Hong-Kong, 1996.

[BCL02] Daniel Barbara, Julia Couto, and Yi Li. Coolcat : an entropy-based algorithm for categorical clustering. In *In Proceedings of the eleventh international conference on Information and knowledge management*, pages 582–589. ACM Press, 2002.

[BF94] Yoshua Bengio and Paolo Frasconi. An input output hmm architecture. In *NIPS*, pages 427–434, 1994.

[BHR00] L. Bergroth, H. Hakonen, and T. Raita. A survey of longest common subsequence algorithms. In *Proceedings of the Seventh International Symposium on String Processing Information Retrieval (SPIRE'00)*, pages 39–, Washington, DC, USA, 2000. IEEE Computer Society.

[BHS97] Christopher M. Bishop, Geoffrey E. Hinton, and Iain G. D. Strachan. Gtm through time. In *In IEE Fifth International Conference on Artificial Neural Networks*, pages 111–116, 1997.

[BJRV08] Romain Boulet, Bertrand Jouve, Fabrice Rossi, and Nathalie Villa. Batch kernel som and related laplacian methods for social network analysis. *Neurocomputing*, 71(7–9) :1257–1273, March 2008.

[BM93] Herve A. Bourlard and Nelson Morgan. *Connectionist Speech Recognition : A Hybrid Approach*. Kluwer Academic Publishers, Norwell, MA, USA, 1993.

[BMS10] Davide Bacciu, Alessio Micheli, and Alessandro Sperduti. Compositional generative mapping of structured data. In *Proceedings of International Joint Conference on Neural Networks*, IJCNN'10, pages 1–8, 2010.

[Bou08] Djamel Bouchaffra. Embedding hmm's-based models in a euclidean space : The topological hidden markov models. In *ICPR08*, pages 1–4, 2008.

[BSI98] C. M. Bishop, M. Svensén, and C. K. I.Williams. GTM : The generative topographic mapping. *Neural Comput*, 10(1) :215–234, 1998.

[BSK04] Dan Buzan, Stan Sclaroff, and George Kollios. Extraction and clustering of motion trajectories in video. In *In International Conference on Pattern Recognition*, pages 521–524, 2004.

[BT04] D. Bouchaffra and J. Tan. Introduction to the concept of structural hmm : Application to mining customers' preferences in automotive design. In *ICPR '04 : Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 2*, pages 493–496, Washington, DC, USA, 2004. IEEE Computer Society.

[BT06] D. Bouchaffra and J. Tan. Structural hidden markov models : An application to handwritten numeral recognition. *Intell. Data Anal.*, 10(1) :67–79, 2006.

[CG91b] Gilles Celeux and Gérard Govaert. Clustering criteria for discrete data and latent class models. *Journal of classification*, (8) :157–176, 1991.

[CG92] Gilles Celeux and Gérard Govaert. A classification em algorithm for clustering and two stochastic versions. *Comput. Stat. Data Anal.*, 14(3) :315–332, 1992.

[Chu97] Fan R. K. Chung. *Spectral Graph Theory (CBMS Regional Conference Series in Mathematics, No. 92)*. American Mathematical Society, February 1997.

[Cox70] D R. Cox. The analysis of binary data. Chapman and Hall, 1970.

[CT93] Geoffrey J. Chappell and John G. Taylor. The temporal kohonen map. *Neural Netw.*, 6 :441–445, March 1993.

[DH73] R. DUDA and P. HART. Pattern classification and scene analysis. 1973.

[DL04] Richard I. A. Davis and Brian C. Lovell. Comparing and evaluating hmm ensemble training algorithms using train and test and condition number criteria., 2004.

[DLR77] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Roy. Statist. Soc*, 39(1) :1–38, 1977.

[DLSW98] Sara Dolnicar, Friedrich Leisch, Gottfried Steiner, and Andreas Weingessel. A comparison of several cluster algorithms on artificial binary data scenarios from tourism marketing : Part 2. Working Paper 19, SFB "Adaptive Information Systems and Modeling in Economics and Management Science", September 1998.

[DLW+98] Sara Dolnicar, Friedrich Leisch, Andreas Weingessel, Christian Buchta, and Evgenia Dimitriadou. A comparison of several cluster algorithms on artificial binary data scenarios from tourism marketing. Working Paper 7, SFB "Adaptive Information Systems and Modeling in Economics and Management Science", April 1998.

[FLBH05] German Florez-Larrahondo, Susan Bridges, and Eric A. Hansen. Incremental estimation of discrete hidden markov models based on a new backward procedure. In *Proceedings of the 20th national conference on Artificial intelligence - Volume 2*, pages 758–763. AAAI Press, 2005.

[For73] G. D. Forney. The viterbi algorithm. *Proceedings of The IEEE*, 61 :268–278, 1973.

[FPS02] Dimitrios S. Frossyniotis, Minas Pertselakis, and Andreas Stafylopatis. A multi-clustering fusion algorithm. In *SETN '02 : Proceedings of the Second Hellenic Conference on AI*, pages 225–236, London, UK, 2002. Springer-Verlag.

[Fri95] Bernd Fritzke. A growing neural gas network learns topologies. In *Advances in Neural Information Processing Systems 7*, pages 625–632. MIT Press, 1995.

[FS08a] C. Ferles and A. Stafylopatis. Sequence clustering with the self-organizing hidden markov model map. *BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on*, pages 1–7, Oct. 2008.

[FS08b] Christos Ferles and Andreas Stafylopatis. A hybrid self-organizing model for sequence analysis. In *ICTAI '08 : Proceedings of the 2008 20th IEEE International Conference on Tools with Artificial Intelligence*, pages 105–112, Washington, DC, USA, 2008. IEEE Computer Society.

[GBO98] T. Graepel, M. Burger, and K. Obermayer. Self-organizing maps : generalizations and new optimization techniques. *Neurocomputing*, 21 :173–190, 1998.

[GGR99] Venkatesh Ganti, Johannes Gehrke, and Raghu Ramakrishnan. Cactusclustering categorical data using summaries. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 73–83, New York, NY, USA, 1999. ACM.

[Gir01] M. Girolami. The topographic organisation and visualisation of binary data using mutivariate-bernoulli latent variable models. *I.E.E.E Transactions on Neural Networks*, 12(6) :1367–1374, 2001.

[GJ97] Zoubin Ghahramani and Michael I. Jordan. Factorial hidden markov models. *Mach. Learn.*, 29(2-3) :245–273, 1997.

[GK01] Valerie Guralnik and George Karypis. A scalable algorithm for clustering sequential data. In *IEEE International Conference on Data Mining*, pages 179–186, 2001.

[Gov90a] G. Govaert. Classification binaire et modèles. *Revue de Statistique Appliquée*, XXXVIII(1) :67–81, 1990.

[GRS99] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Rock : A robust clustering algorithm for categorical attributes. In *Proceedings of the 15th International Conference on Data Engineering*, ICDE '99, pages 512–, Washington, DC, USA, 1999. IEEE Computer Society.

[Hes01] T. Heskes. Self-organizing maps, vector quantization, and mixture modeling. *IEEE Trans. Neural Networks*, 12 :1299–1305, 2001.

[HH08] Chung-Chian Hsu and Yan-Ping Huang. Incremental clustering of mixed data based on distance hierarchy. *Expert Syst. Appl.*, 35 :1177–1185, October 2008.

[HJ99] Lynette Hunt and Murray Jorgensen. Mixture model clustering using the multimix program. *Austral. & New Zealand J Statistics*, 41(2) :153–171, 1999.

[HJ03] Lynette Hunt and Murray Jorgensen. Mixture model clustering for mixed data with missing information. *Comput. Stat. Data Anal.*, 41(3-4) :429–440, 2003.

[HMSS04] Barbara Hammer, Alessio Micheli, Alessandro Sperduti, and Marc Strickert. A general framework for unsupervised processing of structured data. *Neurocomputing*, 57 :3–35, March 2004.

[HNRL05] Joshua Zhexue Huang, Michael K. Ng, Hongqiang Rong, and Zichen Li. Automated variable weighting in k-means type clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27 :657–668, May 2005.

[HSTM03] Markus Hagenbuchner, Ro Sperduti, Ah Chung Tsoi, and Senior Member. A self-organizing map for adaptive processing of structured data. *IEEE Transactions on Neural Networks*, 14 :491–505, 2003.

[Hsu06] Chung-Chian Hsu. Generalizing self-organizing map for categorical data. *Neural Networks, IEEE Transactions on*, 17(2) :294–304, 2006.

[Hua97a] Zhexue Huang. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. In *In Research Issues on Data Mining and Knowledge Discovery*, pages 1–8, 1997.

[Hua97b] Zhexue Huang. Clustering large data sets with mixed numeric and categorical values. In *In The First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 21–34, 1997.

[HW06] Chung-Chian Hsu and Sheng-Hsuan Wang. An integrated framework for visualized and exploratory pattern discovery in mixed data. *IEEE Trans. on Knowl. and Data Eng.*, 18 :161–173, February 2006.

[IC95] S. Ibbou and M. Cottrell. Multiple correspondance analysis crosstabulation matrix using the kohonen algorithm. In *Verlaeysen, M. Editor proc of ESANN'95*, pages 27–32. Dfacto Bruxelles, 1995.

[JH96] Murray Jorgensen and Lynette Hunt. Mixture model clustering of data sets with categorical and continuous variables. In Eds. D. L. Dowe K. B. Korb

J. J. Oliver World Scientific, editor, *Information, Statistics and Induction in Science,ISIS 96*, pages 375–384, Australia, 1996. MIT Press.

[JN07] François-Xavier Jollois and Mohamed Nadif. Speed-up for the expectation-maximization algorithm for clustering categorical data. *Journal of Global Optimization*, 37(4) :513–525, April 2007.

[Kan90] J Kangas. Time-delayed self-organizing maps. In *Proceedings of International Joint Conference on Neural Networks*, IJCNN'90, pages 331 – 336, San Diego, CA , USA, 1990.

[KG01] A. Kaban and M. Girolami. A combined latent class and trait model for the analysis and visualization of discrete data. *IEEE Trans. Pattern Anal. Mach. Intell*, 23 :859–872, 2001.

[KKL97] Teuvo Kohonen, Samuel Kaski, and Harri Lappalainen. Self-organized formation of various invariant-feature filters in the adaptive-subspace som. *Neural Comput.*, 9(6) :1321–1344, 1997.

[KL83] J. Kruskal and M. Liberman. The symmetric time-warping problem : from continuous to discrete. 1983.

[Koh01b] T. Kohonen. *Self-organizing Maps*. Springer Berlin, Vol. 30, Springer, Berlin, Heidelberg, New York, 1995, 1997, 2001. Third Extended Edition, 501 pages, 2001.

[KS02] Teuvo Kohonen and Panu Somervuo. How to make large self-organizing maps for nonvectorial data. *Neural Networks*, 15(8-9) :945–952, 2002.

[LB02] Cen Li and Gautam Biswas. Unsupervised learning with mixed numeric and nominal data. *IEEE Trans. on Knowl. and Data Eng.*, 14 :673–690, July 2002.

[Li06] Tao Li. A unified view on clustering binary data. *Machine Learning*, 62(3) :199–215, 2006.

[Lut94] S. P Luttrel. A bayesian analysis of self-organizing maps. *Neural Computing*, 6 :767 – 794, 1994.

[Mac67] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[Mac97] David J.C. MacKay. Ensemble learning for hidden markov models. Technical report, 1997.

[Mar89] F. Marchetti. Contribution à la classification de données binaires et qualitatives. In *thèse de l'université de Metz*, 1989.

[MF00] Donald Macdonald and Colin Fyfe. The kernel self-organising map. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 317–320, 2000.

[MJ95] Marina Meila and Michael I. Jordan. Learning fine motion by markov mixtures of experts. Technical report, Cambridge, MA, USA, 1995.

[MK97] G. McLachlan and T. Krishman. *The EM algorithm and Extensions*. Wiley, New York, 1997.

[Moh91] Bojan Mohar. The laplacian spectrum of graphs. In *Graph Theory, Combinatorics, and Applications*, pages 871–898. Wiley, 1991.

[MP00] Geoffrey Mclachlan and David Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley-Interscience, October 2000.

[MS91] T. Martinetz and K. Schulten. A "neural-gas" network learns topologies. *Artificial Neural Networks*, I :397–402, 1991.

[NG98a] M. Nadif and G. Govaert. Clustering for binary data and mixture models : Choice of the model. *Applied Stochastic Models and Data Analysis*, 13 :269–278, 1998.

[NG98b] M. Nadif and G. Govaert. Clustering for binary data and mixture models : Choice of the model. *Applied Stochastic Models and Data Analysis*, 13 :269–278, 1998.

[NG08] Mohamed Nadif and Gérard Govaert. Algorithms for model-based block gaussian clustering. In Robert Stahlbock, Sven F. Crone, and Stefan Lessmann, editors, *DMIN, Proceedings of The 2008 International Conference on Data Mining, DMIN 2008, July 14-17, 2008, Las Vegas, USA, 2 Volumes*, pages 536–542. CSREA Press, 2008.

[NG10] Mohamed Nadif and Gerard Govaert. Model-based co-clustering for continuous data. *Machine Learning and Applications, Fourth International Conference on*, 0 :175–180, 2010.

[NJW01] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering : Analysis and an algorithm. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 849–856. MIT Press, 2001.

[OAB97] L. M. D. Owsley, L. E. Atlas, and G. D. Bernard. Self-organizing feature maps and hidden markov models for machine-tool monitoring. *IEEE Transactions on Signal Processing*, 45 :2787–2798, 1997.

[OFC01] Tim Oates, Laura Firoiu, and Paul Cohen. Using dynamic time warping to bootstrap hmm-based clustering of time series. In Ron Sun and C. Giles, editors, *Sequence Learning*, volume 1828 of *Lecture Notes in Computer Science*, pages 35–52. Springer Berlin / Heidelberg, 2001.

[OR06] Jörg Ontrup and Helge Ritter. Large-scale data exploration with the hierarchically growing hyperbolic som. *Neural Netw.*, 19 :751–761, July 2006.

[OSJM09] Timothy F. Oliver, Bertil Schmidt, Yanto Jakop, and Douglas L. Maskell. High speed biological sequence analysis with hidden markov models on reconfigurable platforms. *IEEE Transactions on Information Technology in Biomedicine*, 13 :740–746, 2009.

[OV08] Iván Olier and Alfredo Vellido. Advances in clustering and visualization of time series using gtm through time. *Neural Netw.*, 21 :904–913, September 2008.

[PMM+09] Federico Prat, Andrés Marzal, Sergio Martín, Rafael Ramos-garijo, and María José Castro Bleda. A template-based recognition system for on-line handwritten characters. *Journal of Information Science and Engineering*, 25 :779–791, 2009.

[PN06] Rodolphe Priam and Mohamed Nadif. Carte auto-organisatrice probabiliste sur données binaires. In *EGC*, pages 445–456, 2006.

[PNG08] Rodolphe Priam, Mohamed Nadif, and Gérard Govaert. Binary block gtm : Carte auto-organisatrice probabiliste pour les grands tableaux binaires. In *Extraction et gestion des connaissances (EGC'2008), Actes des 8èmes journées Extraction et Gestion des Connaissances*, Revue des Nouvelles Technologies de l'Information, pages 265–272, Sophia-Antipolis, France„ 29 janvier au 1er février 2008. Cépaduès-Éditions.

[Rab89] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) :257–286, 1989.

[RMD02] Andreas Rauber, Dieter Merkl, and Michael Dittenbach. The growing hierarchical self-organizing map : Exploratory analysis of high-dimensional data. *IEEE Transactions on Neural Networks*, 13 :1331–1341, 2002.

[Sak78] Hiroaki Sakoe. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26 :43–49, 1978.

[SG02] Alexander Strehl and Joydeep Ghosh. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research (JMLR)*, 3 :583–617, December 2002.

[SH03] Marc Strickert and Barbara Hammer. Neural gas for sequences. In *Proceedings of the Workshop on Self-Organizing Networks (WSOM), Kyushu Institute of Technology*, pages 53–57, 2003.

[SK99] Panu Somervuo and Teuvo Kohonen. Self-organizing maps and learning vector quantization forfeature sequences. *Neural Process. Lett.*, 10(2) :151–159, 1999.

[Smy97] Padhraic Smyth. Clustering sequences with hidden markov models. In *Advances in Neural Information Processing Systems*, pages 648–654. MIT Press, 1997.

[Som00] Panu Somervuo. Competing hidden markov models on the self-organizing map. *Neural Networks, IEEE - INNS - ENNS International Joint Conference on*, 3 :3169, 2000.

[SZC02] Ying Sun, Qiuming Zhu, and Zhengxin Chen. An iterative initial-points refinement algorithm for categorical data clustering. *Pattern Recogn. Lett.*, 23(7) :875–884, 2002.

[TJP04] Alexander P. Topchy, Anil K. Jain, and William F. Punch. A mixture model for clustering ensembles. In *SDM,proceedings of the Fourth SIAM International Conference on Data Mining, Lake Buena Vista, Florida, USA, April 22-24*, 2004.

[TJP05] Member-Alexander Topchy, Fellow-Anil K. Jain, and William Punch. Clustering ensembles : Models of consensus and weak partitions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(12) :1866–1881, 2005.

[VA00] J. Vesanto and E. Alhoniemi. Clustering of the self-organizing map. *Neural Networks, IEEE Transactions on*, 11(3) :586–600, May 2000.

[VHLM01] Markus Varsta, Jukka Heikkonen, Jouko Lampinen, and José Del R. Millán. Temporal kohonen map and the recurrent self-organizing map : Analytical and experimental comparison. *Neural Process. Lett.*, 13 :237–251, July 2001.

[Vit67] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13 :260–269, 1967.

[VMH97] Markus Varsta, José Del R. Millán, and Jukka Heikkonen. A recurrent self-organizing map for temporal sequence processing. In *Proceedings of the 7th International Conference on Artificial Neural Networks*, ICANN '97, pages 421–426, London, UK, 1997. Springer-Verlag.

[Voe02] Thomas Voegtlin. Recursive self-organizing maps. *Neural Netw.*, 15 :979–991, October 2002.

[VVK05] J.J. Verbeek, N. Vlassis, and B.J.A. Krose. Self-organizing mixture models. *Neurocomputing*, 63 :99–123, 2005.

[War63] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of The ACM*, 1963.

[Wie03] Jan C. Wiemer. The time-organized map algorithm : extending the self-organizing map to spatiotemporal signals. *Neural Comput.*, 15 :1143–1171, May 2003.

[Wil08] Ben H Williams. *Extracting Motion Primitives from Natural Handwriting Data*. PhD thesis, Institute for Adaptive and Neural Computation School of Informatics, 2008.

[Yam10] Nobuhiko Yamaguchi. Self-organizing hidden markov models. In Kok Wong, B. Mendis, and Abdesselam Bouzerdoum, editors, *Neural Information Processing. Models and Applications*, volume 6444 of *Lecture Notes in Computer Science*, pages 454–461. Springer Berlin / Heidelberg, 2010.

[ZPAS07] Mohammed J. Zaki, Markus Peters, Ira Assent, and Thomas Seidl. Clicks : An effective algorithm for mining subspace clusters in categorical datasets. *Data Knowl. Eng.*, 60(1) :51–70, 2007.

Résumé

Ce mémoire de synthèse est consacré à l'analyse des données complexes pour lesquelles la représentation des variables qui est toujours numérique rencontre des limites. L'ensemble des travaux présentés dans ce mémoire s'inscrit dans le cadre de l'apprentissage non supervisé dont la problématique consiste à construire des représentations simplifiées de données sans connaissance a priori des classes. Il existe actuellement un nombre conséquent de méthodes de partitionnement, mais elles ne s'adaptent pas toujours aux particularités de certains types de données (binaires, mixtes, séquences). On peut distinguer deux grandes familles de modèles de classification non supervisée : les modèles probabilistes et les modèles déterministes ou tout simplement les modèles de quantification. Dans ce mémoire, une importance particulière est accordée aux modèles des cartes topologiques auto-organisatrices. Deux modèles sont proposés pour le traitement des données mixtes (continues et qualitatives). Dans le premier modèle des modifications de la distance sont apportées pour prendre en compte le type de variables. Dans le deuxième modèle, des cartes topologiques dédiées aux données binaires et mixtes sont proposées, utilisant la distribution gaussienne et de Bernoulli. Un autre axe étudié dans ce mémoire est celui de l'apprentissage de données structurées en séquences (non i.i.d). Un lien étroit est montré entre les chaînes de Markov cachées et les cartes à base de modèles de mélanges. Enfin, un bilan des travaux est présenté tout en fournissant des perspectives générales.

Mots clés : apprentissage non-supérvisé, cartes auto-organisatrices, modèles de mélanges, chaînes de Markov, données binaires, données catégorielles, données mixtes, données séquentielles.

Abstact

The research presented in this thesis concerns the analysis of complex data for which the representation of numeric variables always encounters limits. All the approaches presented in this document are part of the unsupervised learning method. There is currently a significant number of clustering methods, but they do not take into account certain types of data (binary, mixed, sequences). For each type of data we propose an adapted unsupervised learning algorithm. There are two main families of clustering models : probabilistic models and deterministic models. In this thesis, a particular emphasis is given to models of self-organizing maps. Two models are proposed for the clustering of mixed data (continuous and categorical). In the first model, we propose to modify the distance in order to take into account the variables type. The second model, described in this work, is a new learning algorithm of topological map dedicated to binary data and mixed data using the Gaussian distribution and Bernoulli. This approach allows probability map interpretation and offers the possibility to take advantage of local distribution associated with continuous and categorical variables. Another field studied in this thesis is the learning with sequence data (not iid). A close link is shown between the hidden Markov chains and self-organized maps based on mixture models. Finally, a review of the work is presented and general perspectives are provided.

Key words : unsupervised learning, self-organised-maps, mixture models, hidden Markov models, categorical data, binary data, mixed data, sequential data