Université de Paris Sud

Numéro d'ordre : 1169

MÉMOIRE SCIENTIFIQUE

présenté pour l'obtention d'une

HABILITATION À DIRIGER DES RECHERCHES

de l'Université Paris Sud

Spécialité Informatique

Data Mining

Exploration, Sélection, Compréhension

Vincent Lemaire

Soutenue le 1er décembre 2008 devant la commission d'examen

Organisme :
Orange Labs (France Telecom R&D) Lannion
Groupe "Traitement Statistique de l'Information" (TECH/EASY/TSI)

Coordonnées :
2 avenue Pierre Marzin, 22300 Lannion
Téléphone : 02 96 05 31 07
Email : vincent.lemaire@orange-ftgroup.com

Jury :

Présidente du jury	- Michèle Sebag	- Directrice de Recherche CNRS
Rapporteur	- Annie Morin	- Maître de Conférences habilitée à l'IFSIC, université de Rennes 1
Rapporteur	- Djamel A. Zighed	- Professeur de l'université de Lyon 2
Rapporteur	- Christian Pellegrini	- Professeur de l'université de Genève
Examinateur	- Isabelle Guyon	- Ingenieur conseil en analyse de données - Société ClopiNet
Examinateur	- Samy Bengio	- Research Scientist in Machine Learning - Google Research

Laboratoire de Recherche en Informatique UMR 8623

.

Membres du jury :

Rapporteur - **Annie Morin - Maître de Conférences habilitée à l'IFSIC, université de Rennes 1**

IFSIC/IRISA, Université de Rennes 1 - Campus de Beaulieu -Rennes Cedex 35042
tel : +33 (0)299 847 222, mailto :annie.morin@irisa.fr

Annie Morin is an associate professor at the computer Science department of the university of Rennes since 1979. Her areas of expertise include data analysis, clustering and classification, and text mining. More precisely, she works on image and text retrieval using factorial analysis methods. She is on the programme committee of several conferences ITI, EGC and manages several cooperation programs with Japan, Croatia (university of Zagreb) and Slovenia (university of Ljubljana) on multimedia indexing.

Rapporteur - **Djamel A. Zighed - Professeur 1er classe de l'université de Lyon 2**

Laboratoire ERIC - Université Lyon 2 - 5, avenue Pierre Mendès-France, Bât L., 69600 Bron France
tel : +33 478 772 376, mailto :abdelkader.zighed@univ-lyon2.fr, http ://eric.univ-lyon2.fr/ zighed

Président de l'association Extraction et Gestion connaissances et responsable du master Extraction des Connaissances à partir des Données. Il effectue ses recherches dans le domaine de la fouille de données et plus particulièrement la fouille de données complexes.

Rapporteur - **Christian Pellegrini - Professeur de l'université de Genève**

CUI - University of Geneva, Battelle, bâtiment A - 7 route de Drize - CH-1227 CAROUGE/Switzerland
tel : +41 22 379 02 20, mailto :Christian.Pellegrini@cui.unige.ch, http ://cui.unige.ch/AI-group/home.html

Christian Pellegrini is Director of the Computer Science Department of the University of Geneva. After obtaining an MSc in high-energy physics and a PhD in Computer Science from the University of Geneva, he worked as an IBM post-doctoral fellow at the T.J. Watson Research Center in Yorktown Heights, NY. He is Full Professor at the University of Geneva since 1982. He has taught a variety of computer science courses such as the architecture and technology of computers, image processing, computer graphics, and artificial intelligence. He is member of several expert committees for the Swiss National Research Foundation and for other international research institutions. He has also been a member of numerous programme committees for national and international conferences. His current research interests include imachine learning, knowledge extraction and data mining, high performance and distributed computing and artificial intelligence.

Examinateur - **Michèle Sebag - Directrice de Recherche CNRS**

Bat 490, Laboratoire de Recherche en Informatique, Université Paris-Sud Orsay, 91405 Orsay Cedex
tel. - , mailto :michele.Sbebag@lri.fr, http ://www.lri.fr/ẽbag/

Michele Sebag, (PhD Université Paris-Dauphine 1990 ; HdR Université Paris-Sud 1997) est Directrice de Recherche CNRS et co-responsable de l'équipe Inférence et Apprentissage au LRI de Paris-Sud, et de l'équipe TAO, INRIA-Ile de France. Ses recherches concernent l'apprentissage relationel, l'apprentissage à grande échelle, l'optimisation stochastique, et les systèmes complexes. Elle est membre du comité de pilotage du réseau d'excellence PASCAL, présidente de l'association francaise pour l'Intelligence Artificielle, membre (senior) du comité de programme de la plupart des conférences internationales d'apprentissage et d'evolution artificiels (ICML, ECML, NIPS, GECCO, PPSN, ECAI, IJCAI).

Examinateur - **Isabelle Guyon - Société ClopiNet**

Ingenieur de l'Ecole de Physique et Chimie Industrielles de la Ville de Paris, Docteur de l'Universite P&M Curie

Clopinet - 955 Creston Road - Berkeley CA 94708 - USA
tel : +1 (510) 524-6211, mailto :isabelle@clopinet.com, http ://www.clopinet.com

Isabelle Guyon is an independent engineering consultant, specialized in statistical data analysis, pattern recognition and machine learning techniques. Her areas of expertise include handwriting recognition, biometrics, and bioinformatics. She has recently consulted for several biotech companies developing new medical instruments including DNA microarray data, antibody arrays, and mass-spectrometers. This work included designing experiments for repeatability and reproducibility and data collection, developing noise models, constructing and selecting features and developing predictive models. Prior to starting her consulting practice in 1996, Isabelle Guyon was a researcher at AT&T Bell Laboratories, where she pioneered applications of neural networks to pen computer interfaces and invented Support Vector Machines (in collaboration with B. Boser and V. Vapnik), which is now a textbook method and is widely used. Isabelle Guyon holds a Ph.D. degree in Physical Sciences of the University Pierre and Marie Curie of Paris, France. She is vice-president of the Unipen foundation, action editor of the Journal of Machine Learning Research, area chair at the NIPS conference and competition chair of the IJCNN conference. She received in 2005 an award from the US National Science Foundation to organize a competition on model selection and in 2007 to organize a competition on causal discovery.

Examinateur - **Samy Bengio - Google Research**

Research Scientist in Machine Learning

Google, 1600 Amphitheatre Pkwy, B1350-138B, Mountain View, CA 94043, USA
tel : +1 (650) 253-2563, mailto :bengio@google.com, http ://bengio.abracadoudou.com

Samy Bengio (PhD in computer science, University of Montreal, 1993) is a research scientist at Google since 2007. Before that, he was senior researcher in statistical machine learning at IDIAP Research Institute since 1999, where he supervised PhD students and postdoctoral fellows working on many areas of machine learning such as support vector machines, time series prediction, mixture models, large-scale problems, speech recognition, multi channel and asynchronous sequence processing, multi-modal (face and voice) person authentication, brain computer interfaces, text mining, and many more. He is Associate Editor of

the Journal of Computational Statistics, has been general chair of the Workshops on Machine Learning for Multimodal Interactions (MLMI'2004, 2005 and 2006), programme chair of the IEEE Workshop on Neural Networks for Signal Processing (NNSP'2002), and on the programme committee of several international conferences such as NIPS and ICML.

Merci...

L'élaboration de ce mémoire m'a amené à jauger le chemin parcouru depuis l'époque où j'ai soutenu ma thèse et je tiens au terme de ce travail à remercier les personnes qui m'ont accompagné. J'ai considéré ma thèse comme la conclusion d'un cycle d'études, la conclusion de mon bref métier de professeur de lycée. Elle a été ma transition vers un métier lié à la recherche.

J'ai été touché par l'amabilité de tous les membres de mon jury qui me fait l'honneur de leur présence pour cette habiliation : Michèle Sebag, Annie Morin, Christian Pellegrini, Isabelle Guyon, Samy Bengio, Djamel A. Zighed. Merci à eux d'avoir pris de leur temps pour évaluer mon travail et donner leur avis. Un merci particulier à Michèle Sebag qui m'a incité à préparer cette habilitation et qui a été ma marraine du LRI.

Parmi les personnes qui m'ont accompagné depuis neuf ans, je pense à l'ensemble de mes coauteurs, tant pour la collaboration fructueuse, que pour tous les aspects méthodologiques qu'ils m'ont apportés mais aussi pour les liens de sympathie tissés entre nous. Je voudrais citer plus particulièrement ceux avec qui je travaille depuis plusieurs années, et je l'espère pour encore de nombreuses années, à savoir principalement Raphael Féraud, Fabrice Clérot, Francoise Fessant ... Je n'oublie pas parmi eux ceux de mon équipe ou de celle d'à coté : Marc Boullé, Olivier Bernier, Alexis Bondu avec lesquels j'ai eu de riches conversations scientifiques.

Je remercie aussi celles et ceux qui à la direction de la recherche de France Télécom m'ont fait confiance soit pour prendre la charge de l'encadrement d'activités de recherche, soit pour avoir écouter mon point de vue lors de commision. Je citerai à titre d'exemple Alain Léger, Christelle Sorin et Adam Ouorou.

L'environnement de travail de l'équipe TSI est aussi un moteur extraordinaire que je tiens à souligner en remerciant tous ses membres passés et actuels. Je pense tout particulièrement à une de nos figures incontournables, Daniel Collobert qui me mis le pied à l'étrier en 1996.

Je pense enfin à ma chérie et à Thomas. Leur amour, leur soutien, leur confiance, sont essentiels.

Table des matières

Introduction

"Exploratory Analysis using Kohonen Maps"

"Curriculum Vitae"

Introduction

Chapitre 1

Data Mining : Exploration, Sélection, Compréhension

1.1 Domaine de recherche

1.1.1 La fouille de données

La fouille de données est connue sous différentes appellations suivant que le processus évoqué précédemment est appliqué à des données qui se présentent sous les formes de la figure 1.1. Les différentes étapes du processus d'**analyse de données**, de data mining, peuvent être décrites selon le modèle CRISP-DM qui se propose de découper tout processus Data Mining en 6 phases :

1. La phase de recueil des besoins (de l'anglais business understanding), fixe les objectifs industriels et les critères de succès, évalue les ressources, les contraintes et les hypothèses nécessaires à la réalisation des objectifs, traduit les objectifs et critères industriels en objectifs et critères techniques, et décrit un plan de résolution afin d'atteindre les objectifs techniques.
2. La phase de compréhension des données (de l'anglais data understanding), réalise la collecte initiale des données, en produit une description, étudie éventuellement quelques hypothèses à l'aide de visualisations et vérifie le niveau de qualité des données.
3. La phase de préparation des données, consiste en la construction d'une table de données pour modélisation. Nous nous y intéressons plus particulièrement par la suite.
4. La phase de modélisation, procède à la sélection de techniques de modélisation, met en place un protocole de test de la qualité des modèles obtenus, construit les modèles et les évalue selon le protocole de test.
5. La phase d'évaluation estime si les objectifs industriels ont été atteints, s'assure que le processus a bien suivi le déroulement escompté et détermine la phase suivante : retour en arrière ou déploiement.
6. La phase de déploiement industrialise l'utilisation du modèle en situation opérationnelle, définit un plan de contrôle et de maintenance, produit un rapport final et effectue une revue de projet.

Le modèle CRISP-DM est essentiellement un guide méthodologique pour la conduite d'un projet Data Mining. Les phases initiales et finales s'apparentent à des activités d'expertise en organisation, consulting, bases de données et développement informatique. Elles supposent une implication humaine importante. Seules les phases centrales sont partiellement automatisables.

	Stream Mining	Text Mining	Data Mining	Web Mining	Graph Mining
Compréhension	x	x	x	x	x
Préparation	x	x	x	x	x
Modélisation	x	x	x	x	x
Interprétation	x	x	x	x	x
Déploiement	x	x	x	x	x

FIG. 1.1 – Domaine d'activité

Les principaux domaines d'application (liste non exhaustive en dehors du data mining [Ext-38][1] ou l'on considère que les données sont numériques et sous forme de table instances x variables) du data mining mentionnés tableau 1.1 sont :

– le **Web Mining** [Ext-39] : C'est l'application des techniques d'extraction de données pour découvrir des "modèles", "structures", à partir du web. On parle généralement de :
  * Web Content Mining : analyse du contenu des pages ;
  * Web Structure Mining : analyse de la structure des pages, du web ;
  * User Content Mining : analyse des parcours, des profils, des utilisateurs, ...

– le **Texte Mining** [Ext-40] : Il se réfère généralement au processus d'extraction d'information de données textuelles. On parle de clustering de texte, de catégorisation de texte, d'extraction de concept, de création automatique de résumés, ...

– le **Graph Mining** [Ext-41] : L'utilisation croissance de données structurées (exemple XML) a crée un besoin en data mining. Les algorithmes d'extraction d'information qui avaient été pensés pour des données tabulaires ont du être redéveloppés. Une première phase à constitué à utiliser des données relationnelles mais des limitations restaient. Des algorithmes alors capables de travailler directement sous la forme native (structurées en graphes) des données ont été développés.

– le **Stream Mining** [Ext-42] : C'est le processus d'extraction d'information à partir de données représentées sous forme de flux. Les enregistrements contenus dans le flux ne peuvent être lus qu'une fois (ou un nombre très restreint de fois), en utilisant des capacités limitées de calcul et de mémoire. Les analyses doivent alors être menées en une passe ce qui demande des techniques bien différentes de celles employées sur des données "mémorisables".

1.1.2 L'apprentissage automatique

L'apprentissage automatique, dans une définition très générale, consiste en l'élaboration de programmes qui s'améliorent avec l'expérience. On utilisera l'un des termes de la figure 1.2.

Les principaux types d'apprentissage statistique mentionnés tableau 1.2 sont :

– **Apprentissage non supervisé** [Ext-43] : Il n'y a pas de tuteur, plus exactement le "modèle" ne reçoit aucune information de l'environnement lui indiquant quelles devraient être ses sorties ou même si celles-ci sont correctes. Le système doit donc découvrir par lui-même les corrélations existant entre les patrons d'apprentissage. On cherche à dégager un certain degré d'organisation. L'objectif est de

[1] Les références bibliographiques données dans cette section sont des références qui normalement permettent une appréhension globale et "facile" du domaine".

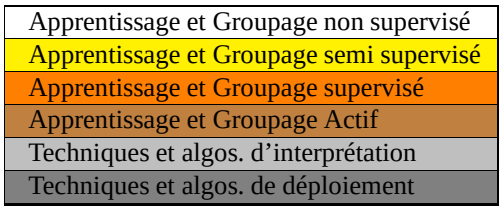

FIG. 1.2 – Domaine d'apprentissage et algos

générer une taxonomie des données sans connaissances préalables (groupage, clustering, ...).

– **Apprentissage semi-supervisé** [Ext-44] : Il s'agit d'utiliser [pour des tâches d'accès à l'information] une petite quantité de données étiquetées conjointement à une masse importante de données non-étiquetées. Cela correspond à une situation de plus de plus fréquente en recherche d'information.

– **Apprentissage supervisé** [Ext-45] : L'apprentissage supervisé est une technique d'apprentissage automatique où l'on cherche à produire automatiquement des règles à partir d'une base de données d'apprentissage contenant des exemples de cas déjà traités. Plus précisément, la base de données d'apprentissage est un ensemble de couples entrée-sortie, que l'on considère être tirés selon une loi inconnue. Le but de la méthode d'apprentissage supervisé est alors d'utiliser cette base d'apprentissage afin de déterminer une représentation compacte et donc de généraliser pour des entrées inconnues ce qu'il a pu « apprendre » grâce aux données déjà traitées par des experts, ceci de façon « raisonnable ». On distingue généralement deux types de problèmes que l'on cherche à résoudre avec une méthode d'apprentissage automatique supervisée : (1) lorsque la sortie que l'on cherche à associer à une entrée est une valeur dans un ensemble continu de réels, on parle d'un problème de régression ; (2) lorsque l'ensemble des valeurs de sortie est de cardinal fini, on parle d'un problème de classification car le but est en fait d'attribuer une étiquette à une entrée.

– **Apprentissage en interaction** :

  * **Apprentissage par renforcement** [Ext-46] : L'apprentissage par renforcement fait référence à une classe de problèmes d'apprentissage automatique, dont le but est d'apprendre, à partir d'expériences, ce qu'il convient de faire en différentes situations, de façon à optimiser une récompense numérique au cours du temps. Un paradigme classique pour présenter les problèmes d'apprentissage par renforcement consiste à considérer un agent autonome, plongé au sein d'un environnement, et qui doit prendre des décisions en fonction de son état courant. En retour, l'environnement procure à l'agent une récompense, qui peut être positive ou négative. L'agent cherche, au travers d'expériences itérées, un comportement décisionnel (appelé stratégie ou politique, et qui est une fonction associant à l'état courant l'action à exécuter) optimal.

  * **Apprentissage Actif** [MP-1] : L'idée est de chercher à chaque pas les exemples les plus informatifs. Il consiste à combiner la construction de modèles à partir de données issues d'expérience avec un système visant à produire de nouvelles données de manière à accélérer l'apprentissage. Il peut donc choisir judicieusement les exemples que l'utilisateur (l'expert) doit étiqueter de façon à minimiser l'effort qui lui est demandé ou le temps d'apprentissage.

1.1.3 Un domaine de recherche à l'intersection

En concaténant les deux figures ci-dessus on obtient alors la figure 1.3 qui présentent une adéquation entre les phases d'un processus de data mining et l'une des formes d'apprentissage. Cette activité de recherche représente "L'apprentissage automatique" au service de "l'analyse de données".

	Stream Mining	Text Mining	Data Mining	Web Mining	Graph Mining
Compréhension	Apprentissage et Groupage non supervisé				
Préparation	Apprentissage et Groupage non supervisé				
	Apprentissage et Groupage actif				
	Apprentissage et Groupage semi supervisé				
Modélisation	Apprentissage et Groupage semi supervisé				
	Apprentissage et Groupage supervisé				
	Apprentissage et Groupage actif				
Interprétation	Techniques et algos. d'interprétation				
Déploiement	Techniques et algos. de déploiement				

FIG. 1.3 – Domaine d'activité (associations principales)

Mes activités m'ont aussi amené à cotoyer des domaines d'application variés :
(*) Son : spatialisation sonore ;
(*) Social Media : apprentissage de profil, clustering de contenus ;
(*) Connaissance Client : churn, fraudes, etc
(*) Images : Détection de visages, réglage d'invariant de luminosité
(*) Web Mining : Agencement de résultats de recherche, optimisation de taux de clics.

1.1.4 Enjeux et Questions

Intérêt - Les progrès technologiques favorisent l'accès à des données de plus en plus importantes. Leur exploitation à des fins de décision par exemple constitue aujourd'hui un défi majeur pour le scientifique. La complexité des données, leur taille croissante, suscitent de nombreux problèmes de recherche et d'applications comme en témoignent la vitalité et le nombre de conférences, symposiums sur le sujet (Knowledge Discovery, Data Mining, Learning, Web Intelligence ...).

La fouille de données s'intéresse au développement de méthodes et de systèmes permettant d'organiser, de prévoir, d'interpréter des données. Ces données peuvent être structurées, comporter de l'imprécision et être de taille importante. De telles données s'accumulent de plus en plus dans des bases de données réparties sur un réseau et suscitent les questions suivantes auxquelles on cherche à répondre : comment représenter des données et des connaissances pour pouvoir extraire des régularités ? Comment organiser les observations en classes ou concepts facilement interprétables ? Comment trouver des fonctions permettant de décider de l'appartenance d'une observation à des classes connues à priori ? Comment traiter des flux de données en temps réel ("à la volée") ?

Enjeux - Le problème est celui de l'extraction et de la représentation de connaissances issues de données collectées par une organisation ou une entreprise (France Telecom étant un cas particulier) au sein de ses systèmes d'information (trafic (fixe ou mobile), profil, logs..), et répond à plusieurs enjeux majeurs pour FT. L'environnement concurrentiel et la multiplication des offres imposent d'aider les décideurs à limiter des risques d'erreurs et à maximiser les opportunités. A cet effet, la fouille de données (KDD) recouvre un ensemble de techniques permettant :

– de classifier, de découvrir ou d'apprendre à partir des données, et donc d'aider à la décision ;

- d'analyser les comportements d'un réseau ou d'un client (limitation les risques comme la fraude, les intrusions, les dénis de service, les passages à la concurrence, réalisation et actualisation de segmentations) ;
- d'exploiter l'accès à des données de plus en plus importantes ;
- d'organiser, d'interpréter les gisements de données.

Verrous technologiques et questions dures - Il résulte des paragraphes ci-dessus une liste non exhaustive de verrous technologiques :

- Adaptation des technologies pour supporter les très grands volumes de données, stockables ou non - le passage à l'échelle ;
- Sélection d'attributs multivariés (identification d'attributs plus significatifs en groupe qu'individuellement) ;
- Complexité de l'apprentissage sur des flux temps réels (streaming) ;
- Construction guidée d'agrégats ;
- Techniques d'explication simple de résultats (interprétabilité).

Exemple de questions dures de recherche

1. Volume (données)/ Abondance (info)
   - Comment introduire explicitement le compromis ressource / performance dans l'arsenal du data mining pour le rendre robuste à l'explosion des volumétries ?
2. Acquisition et Apprentissage
   - Comment acquérir rapidement suffisamment de connaissances (d'information ? de données ?) pour traiter un problème représentatif ?
   - Comment associer observation, exploration, innovation et curiosité pour apprendre dans un univers dynamique ?
   - Quels outils pour l'apprentissage en milieu complexe et non stationnaire ?
   - Comment associer, utiliser des connaissances acquises ou apprises sur les utilisateurs ?
3. Confiance / Validité
   - Comment associer, utiliser des connaissances acquises ou apprises sur les utilisateurs et maintenir (renforcer) la confiance (privacy) ?
4. Centralisée vs. Distribuée
   - Comment vont se combiner connaissances distribuées et connaissances centralisées ?
   - Comment apprendre sur des données, de la connaissance, distribuée ?

D'un point de vue plus académique **la communauté KDD a identifié la liste suivante** [Ext-47] :

- Developing a Unifying Theory of Data Mining
- Scaling Up for High Dimensional Data and High Speed Data Streams
- Mining Sequence Data and Time Series Data
- Mining Complex Knowledge from Complex Data
- Data Mining in a Network Setting
- Distributed Data Mining and Mining Multi-agent Data
- Data Mining for Biological and Environmental Problems
- Data-Mining-Process Related Problems
- Security, Privacy and Data Integrity
- Dealing with Non-static, Unbalanced and Cost-sensitive Data

1.2 Le contenu de ce mémoire

Une première étape de la réflexion qui est présentée dans ce mémoire a été consacrée à l'analyse exploratoire de données : le but a été de développer de nouvelles méthodologies d'analyse (étape 1 du processus de data mining voir 1.1.1). Cette réflexion est présenté dans la partie :

- **"Exploratory Analysis using Kohonen Maps"**
  - The Many Faces of a Kohonen Map [MP-2]
  - Looking for a relevant similarity criterion [MP-3]
  - Combining several SOM approaches in data mining [MP-4]

Fort de cette expérience une nouvelle méthode d'analyse d'un modèle, méthode basée sur une analyse de sensibilité a été développée. Cette méthode robuste a été employée dans le cadre de la sélection de variable (préparation des données : étape 2 d'un processus de data mining) puis dans le cadre de l'interprétation d'une modélisation (interprétation : étape 4 d'un processus de data mining). Elle est présentée dans la partie :

- **"Variable Selection and Model Interpretation"**
  - An new Input Variable Importance Definition [MP-5]
  - Contact Personalization using a Score Understanding Method" [MP-6]
  - A naïve understanding of the naive bayes classifier [MP-7]

Enfin la sélection d'instance soit pour préparer les données soit pour réduire le coût d'étiquetage des données est abordé dans la dernière partie du mémoire :

- **"Instance Selection"**
  - Active Learning using Adaptive Curiosity [MP-8]

En 2002 j'ai réalisé une feuille de route scientifique passant par différentes étapes. Ces étapes sont représentées par les 7 articles scientifiques énumérés ci-dessus. Ces 7 articles sont cohérents entre eux et ils sont le résultat d'une démarche scientifique. Au final ce mémoire d'HDR est un condensé de la route suivie.

Les projets, encadrements, cours, logiciels, etc... auxquels j'ai participé en lien avec cette activité " Data Mining : Exploration, Sélection, Compréhension" peuvent être trouvés dans la partie "Curriculum Vitae" de ce document.

La partie "Curriculum Vitae" détaille aussi toutes les parties de mes activités qui ne ressortent pas directement de la démarche décrite ci-dessus.

“Exploratory Analysis using Kohonen Maps”

Chapitre 2

The Many Faces of a Kohonen Map

Contents

The Self-Organizing Map (SOM) is an excellent tool for exploratory data analysis. It projects the input space on prototypes of a low-dimensional regular grid which can be efficiently used to visualize and explore the properties of the data. In this chapter we present a novel methodology using SOM for exploratory analysis, dimensionality reduction and/or variable selection for a classification problem. The methodology is applied to a real case study and the results are compared with other techniques.

2.1 Introduction

The Self-Organizing Map (SOM) [Ext-48] is an excellent tool for data survey because it has prominent visualization properties. It creates a set of prototype vectors representing the data set and carries out a topology preserving projection of the prototypes from the d-dimensional input space onto a low-dimensional grid (two dimensions in this chapter). This ordered grid can be used as a convenient visualization surface for showing different features of the data.

When the number of SOM units is large, similar units have to be grouped together (clustered) so as to ease the quantitative analysis of the map. Different approaches to clustering of a SOM have been proposed [Ext-49; Ext-50] such as hierarchical agglomeration clustering or partitive clustering using k-means. This SOM-based exploratory analysis is therefore a two-stage procedure :

1. a large set of prototypes (much larger than the expected number of clusters) is formed using a large SOM ;
2. these prototypes are combined to form the final clusters.

Such an analysis deals with the cases and constitutes only a first step. In this chapter we follow the pioneering work of Juha Vesanto [Ext-51] on the use of Kohonen maps for data mining and we propose a second step, which involves a very similar techniques, but deals with the analysis of the variables : each input variable can be described by its projection upon the map of the cases. A visual inspection can be performed to see where (i.e. for which prototype(s) of the SOM) each variable is strong (compared to the other prototypes). It is also possible to compare the projections of different variables. This manual examination of the variables becomes impossible when the number of input variables is large and we propose an automatic examination : the projections of each variable on the map of the cases is taken as a representative vector of the variable and we train a second SOM with these vectors ; this second map (map of the variables) can then be clustered, allowing to automatically group together variables which have similar behaviors.

The organization of the chapter is as follows : the next section deals with the real case studies and in section 2.3 we present our methodology for exploratory analysis with SOM. In section 2.4 we present our methodology for dimensionality reduction and variable selection with SOM. Section 2.5 describes experimental conditions and comparative results between our methodology and others machine learning techniques. A short conclusion follows.

2.2 Case Study

The case study is the on-line detection of the fraudulent use of a post-paid phone card. Post-paid cards are characterized by :

– card number (12 digits written on the card)
– card identifier (4 digits only known by the owner of the card)
– used in public phones (only need to enter the identifier)
– used in any fixed phone (enter the 16 digits for identification)

Here the “fraud” term includes all cases which may lead to a fraudulent non-payment by the caller. The purpose is to prevent non-payments by warning the owners of phone card that the current use of their card is unusual.

The original database contains 15330 individuals described with 226 input variables of various kinds :

– sociological variables
– a series of indicators of traffics ;
– variables of descriptive statistics.

Using a large number of these variables in the modeling phase achieves good fraudulent/non-fraudulent classification performances but such models cannot be applied on-line because of computing and data ex-

traction time constraints. It is thus necessary to reduce the number of variables while maintaining good performance.

We split the data into 3 sets : a training set, a validation set and a test set which contain respectively 70%, 15% and 15% of the cases. Whatever the method evaluated below, the test set is never used to build the classifier. 92 % of the examples in the database belong to the class “not fraudulent” and 8 % belong to the class “fraudulent”.

2.3 Methodology

2.3.1 A Two-Step Two-Level Approach

The methodology used in this chapter is depicted in the Figure 2.1. The primary benefits of each two-level approach are the reduction of the computational cost [Ext-49; Ext-50] and an easier interpretation of the map structure. The second benefit of this methodology is the simultaneous visualization of clusters of cases and clusters of variables for exploratory analysis purposes. Finally, dimensionality reduction and/or variable selection can be implemented.

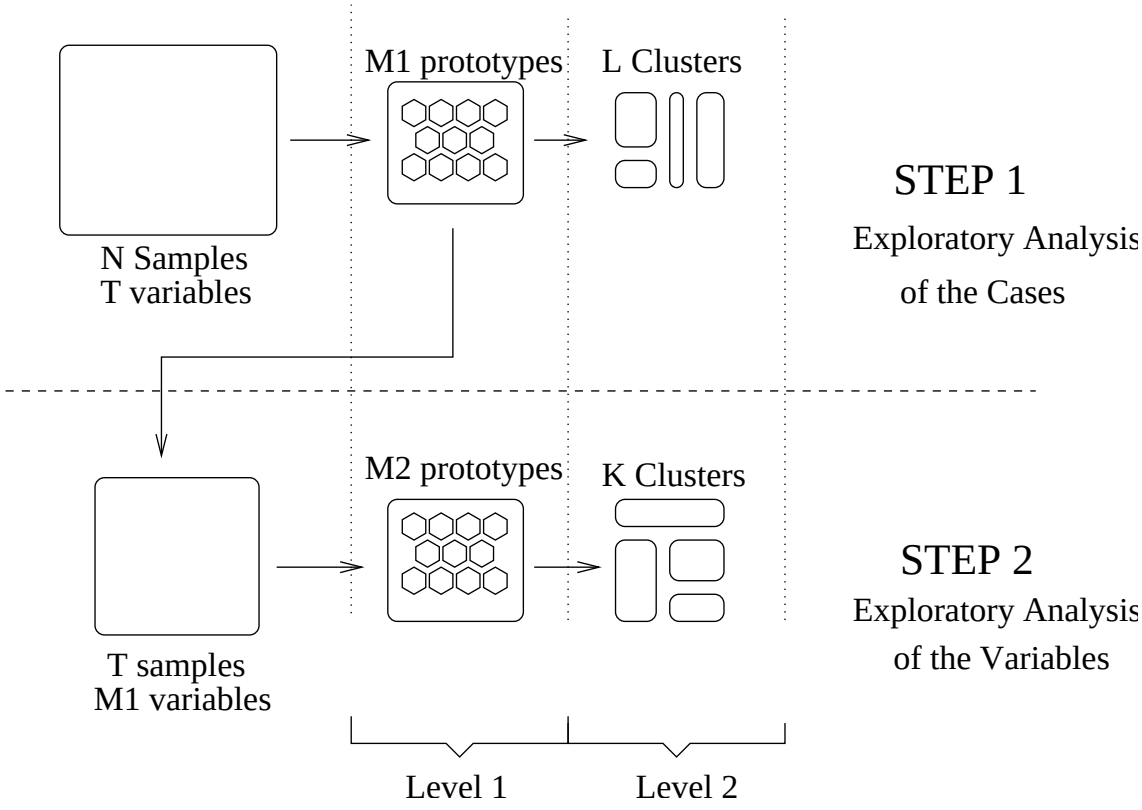

FIG. 2.1 – The Two-Step Two-Level Approach.

All the SOM in this chapter are square maps with hexagonal neighborhoods and are initialized with Principal Component Analysis (PCA). We use a validation set or a cross-validation to measure the error reconstruction and select the map size above which the reconstruction error does not decrease significantly (the reconstruction error for a given size is estimated as an average on 10 attempts).

2.3.2 Top View : Exploratory Analysis of the Cases

The first step of the method is to build a SOM of the cases[1]. The best map size, for the case study was determined to be 12x12. We used the training set and the validation set to achieve a final training of the SOM of the cases.

This map allows to track down the characteristic behaviors of the cases : a standard clustering algorithm can be run on top of the map, revealing groups of cases with similar behaviors (see Figure 2.2). This clustering is done onto the prototypes of the SOM themselves, not on the prototypes weighted by the number of cases belonging to each prototype. Extensive tests have not shown any significant difference between k-means and hierarchical agglomerative clustering with the Ward criterium for this clustering of the map.

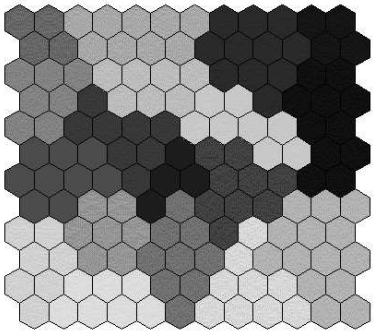

FIG. 2.2 – Groups of cases with similar behaviors found using a hierarchical clustering.

Projecting the class information (fraudulent use or not in our case study ; this information is not used for the construction of the map, see Figure 2.3) on the map allows to investigate the distinctive profiles of the classes in terms of all the input variables.

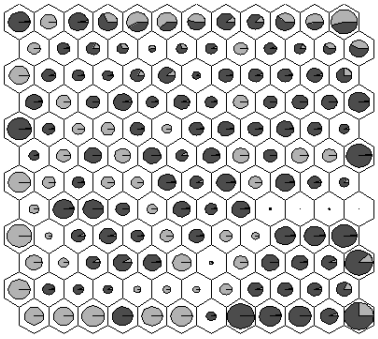

FIG. 2.3 – The two populations for each prototype. The size of the pie indicates the number of cases belonging to each prototype. The lighter the color, the less fraudulent the behavior. For each pie the light grey proportion indicates the proportion of fraudulent behavior. We can project other auxiliary data in a similar manner.

This constitutes the first step. We then proceed to the second step : each input variable is described by its projection upon the map of the cases. Upon visual inspection (see Figure 2.4), one can determine how a variable behaves on the map and relate this behavior to the clusters of cases.

[1] All the experimentations on SOM have been done with the SOM Toolbox package for Matlab © [Ext-52]

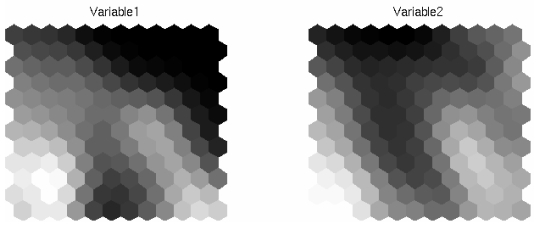

FIG. 2.4 – The projections of the first two variables on the map of the cases : the darker the color, the stronger the value for the corresponding prototype.

It is also possible to visually analyze the relationships between different variables. This visual method consists in the visualization of each projection on the map of the cases and to group together variables with similar projections (see Figure 2.5). However this visual inspection of the variables becomes impossible when the number of input variables grows.

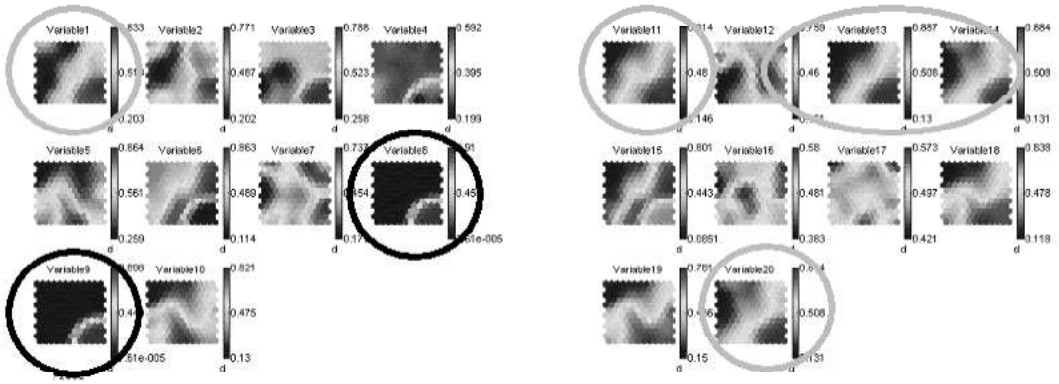

FIG. 2.5 – Each subfigure above shows the projection of 10 variables. Visual inspection allows to find some strongly correlated variables (two “obvious” groups in this example) but is of limited efficiency when the number of variables is large.

Nevertheless, an automatic clustering of the variables according to their projection can be achieved : the projections of each variable on the map of the cases are taken as a representative vector of the variable and we train a second SOM with these vectors ; this second map (map of the variables) can then be clustered, allowing to automatically group together variables which have similar behaviors.

2.3.3 Side View : Exploratory Analysis of the Variables

In the second step, we build a second SOM to explore the variables as follows : each input variable is described by its projection upon the map of the cases, hence by a vector having as many dimensions as the number of neurons of the map of the cases. These variables descriptors are used to train the second map, the map of the variables.

For this SOM we cannot use a validation set since the database is the codebook of the SOM of the cases and is therefore quite small. We use a 5-fold cross-validation [Ext-53] method to find the best size of the SOM of the variables. The selected size is 12 x 12. Knowing the best size of SOM of the variables, we used all the codebooks of the SOM of the cases to perform a final training of the SOM of the variables.

This map allows to explore the relationships between variables and to study the correlation between variables ; we also run a standard clustering algorithm on top on this map to create groups of variables with similar behaviors. Again, this clustering is done onto the prototypes of the SOM themselves, not on the prototypes weighted by the number of variables belonging to each prototypes. The clusters found on the map of the variables can be visualized as for the map of cases (see Figure 2.6).

Figure 2.6 summarizes the results of this analysis of the variables : subfigure (a) shows the map of the variables and its clustering ; subfigures (b) and (c) show the projections of the variables for two clusters of variables. The similarity of the projection for variables belonging to the same cluster is obvious and it can be seen that different clusters indeed correspond to different behaviors of the variables.

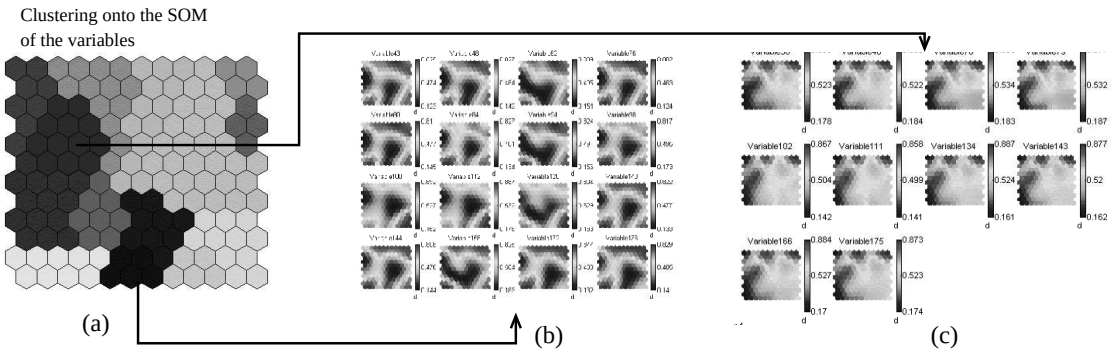

FIG. 2.6 – Groups of variables with similar characteristics using K-means clustering

We used K-means for the clustering onto the SOM of the variables. Here again we cannot use a validation set to determine the optimal K value and we used a 5-fold cross-validation. We chose the value of K^* above which the error reconstruction does not decrease significantly (the result for a given size is an average on 20 attempts). The selected value is $K^* = 11$.

2.3.4 Top View vs. Side View and Exploratory Data Analysis

Figure 2.7 sums up the complete process described above.

At this point, we end up with two clusterings, a clustering of cases and a clustering of variables, which are consistent together : groups of cases have similar behaviors relative to groups of variables and reciprocally, a situation reminiscent of the duality property of PCA. This allows a much easier exploratory analysis and can also be used for dimensionality reduction and/or variable selection since variables of the same group contribute in the same way to the description of the cases.

The map of the variables allows an easier interpretation of the clusters of cases by re-ordering the variables according to their cluster. We see in Figure 2.8(a) the clusters on the map of the cases ; in Figure 2.8(b) the mean value of the variables for the cases belonging to the cluster (A) without re-ordering ; in

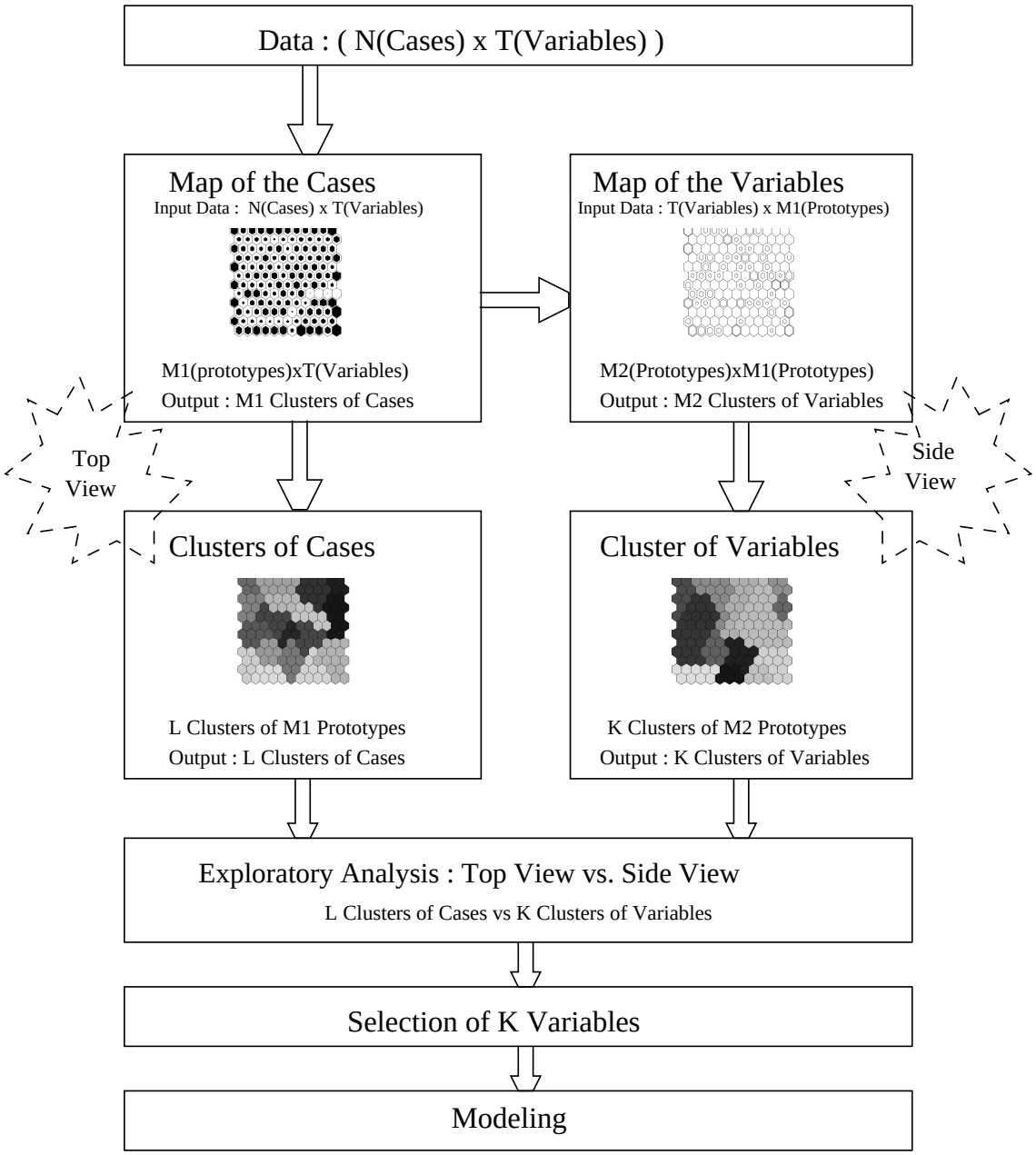

FIG. 2.7 – Top View vs. Side View

Figure 2.8(c) the mean value of the variables for the cases belonging to the cluster (A) re-ordered according to their cluster. Figure 2.8(c) immediately shows how the different clusters of variables contribute to the formation of the cluster of cases A. Such accurate visual analysis is impossible with the raw ordering of the variables (Figure 2.8(b)).

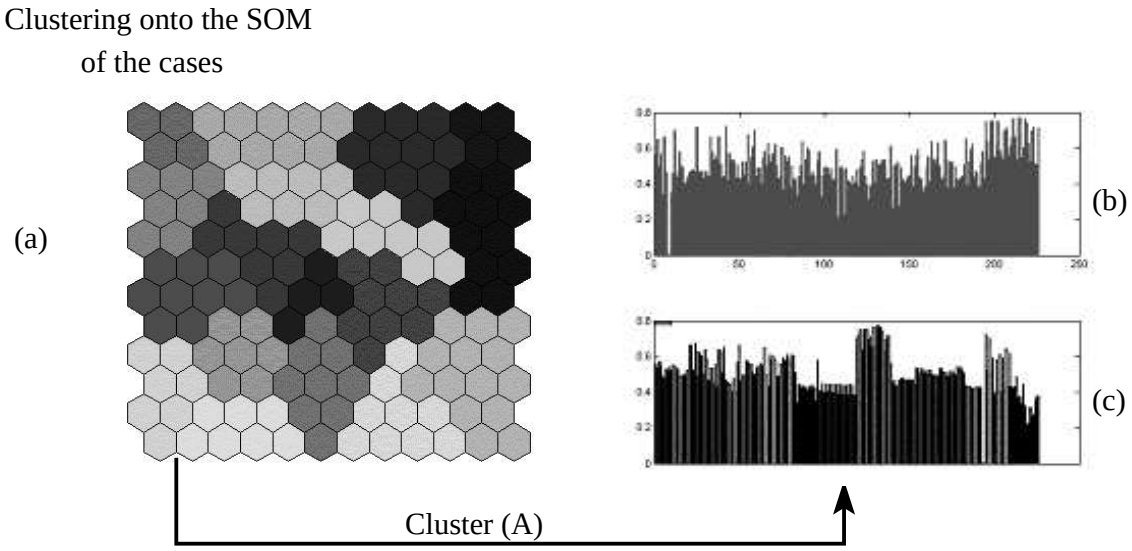

FIG. 2.8 – Re-ordering the variables according to their cluster allow an easier interpretation of the cluster of the cases.

Figure 2.9 illustrates the complete exploratory analysis process which can be done using the method described above. The clustering of the SOM of the cases identifies 12 clusters of cases (a) The projection of the class information allows to visualize the fraudulent behaviors (b). The cluster (A) of the SOM of the cases (in the south-west corner of the map) exhibits fraudulent behavior (b).

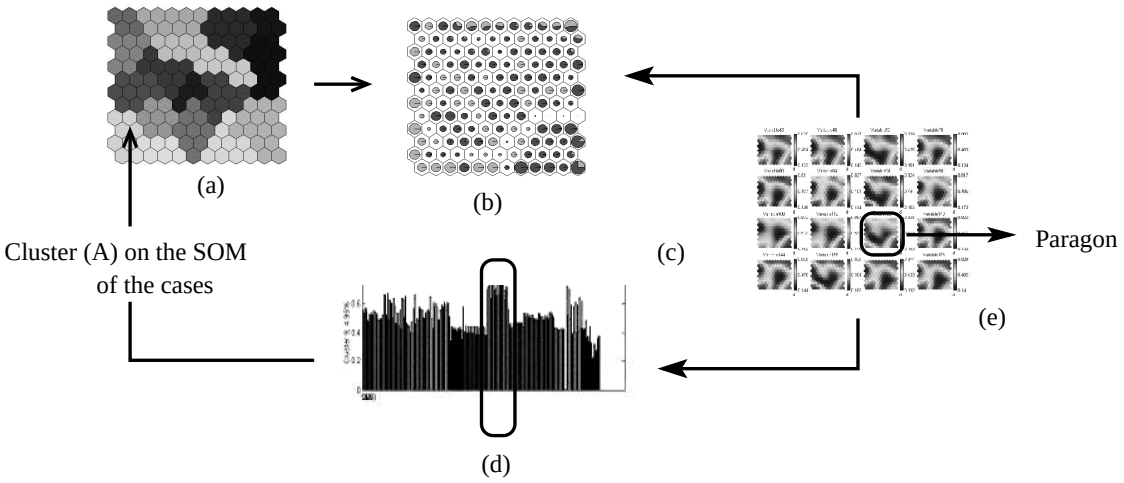

FIG. 2.9 – Example of exploratory analysis.

The clustering obtained on the SOM of the variables allows the re-ordering of the mean value of the variables for the cases belonging to each cluster (c). For the cluster (A) of the SOM of the cases we see a cluster of stronger variables (d). This group of variable is presented in (e) : all the variables are strongly correlated. The grouping of these variables in this cluster is naturally interpreted : these variables represent information about card phone (the amount of communication via a card phone under five temporal observa-

tion windows) and indicate that a specific card phone usage pattern is strongly correlated with a fraudulent behavior.

2.4 Dimensionality Reduction vs. Variable Selection

In this chapter, “dimensionality reduction” refers to techniques which aim at finding a sub-manifold spanned by combinations of the original variables (“features”), while “variable selection” refers to techniques which aim at excluding variables irrelevant to the modeling problem. In both cases, this is a combinatorial optimization problem.

The direct approach (the “wrapper” method) re-trains and re-evaluates a given model for many different feature/variable sets. The “filter” method approximation optimises simple criteria which tend to improve performance. The two simplest optimization methods are forward selection (keep adding the best feature/variable) and backward elimination (keep removing the worst feature/variable) [Ext-54; Ext-55].

As we have seen that each cluster of variables gathers variables with very close profiles, we can exploit this clustering for variable selection in a very natural way : we choose one representative variable per cluster, as the “paragon” of the cluster, i.e. the variable which minimizes the sum of the distances to the other variables of the cluster.

We choose to implement dimensionality reduction by building one feature per cluster as a sum of the variables of the cluster (variables are mean-centered and reduced to unit variance before summing). Both techniques are illustrated in Figure 2.10.

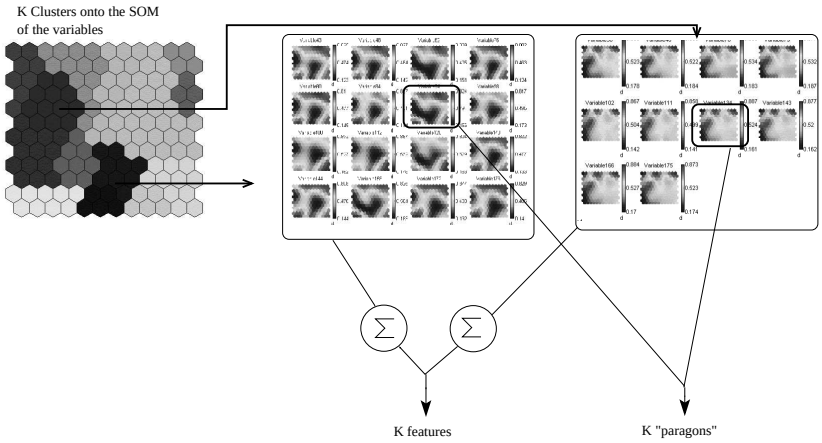

FIG. 2.10 – SOM-based dimensionality reduction and variable selection

Both methods reduce the number of input variables to the number K^* of clusters found on the map of the variables. Modeling after variable selection relies on fewer input variables, therefore relying on less information, while modeling after dimensionality reduction relies on fewer features which may gather all the relevant information in the input variables but are often impossible to interpret in an intelligible way.

2.5 Methodology : Comparison and Results

Other machine learning techniques also allow to realize variable selection such as decision trees, Bayesian networks or dimensionality reduction methods such as PCA. We shall compare the methodology described above to such techniques and this section details the experimental conditions for this comparison.

We shall report a comparison of the results obtained on our case-study :

– on the one hand we shall compare the performance of models which use dimensionality reduction : a neural network trained with all the input variables, a neural network which uses the K^* variables found with dimensionality reduction method described below, and a PCA where we kept the first K^* eigenvectors.
– on the other hand we shall compare the performance of models which use a variable selection : a neural network which uses the K^* variables found with the variable selection method proposed below, a Bayesian network, and a decision tree.

2.5.1 Experimental Conditions

Principal Component Analysis

The principal components are random variables of maximal variance constructed from linear combinations of the input features. Equivalently, they are the projections onto the principal component axes, which are lines that minimize the average squared distance to each point in the data set [Ext-56]. The Principal Component Analysis (PCA) has been constructed on the training set and projected using the first $K^* = 11$ eigenvectors on the validation set and the test set. This may not be the optimal number of eigenvectors but, for comparison purposes, the number of eigenvectors kept has to correspond to the number of clusters of variables found in section 2.3.3.

Multi-layer Perceptrons

Each neural network, in this chapter, is a multilayer perceptron, with standard sigmoidal functions, $K^* = 11$ input neurons, one hidden layer with P neurons and one output. We used the stochastic version on the squared error cost function. The training process is achieved when the cost does not decrease significantly as compared to the previous iteration on the validation set. The learning rate is $\beta = 0.001$.

The optimal number P^* of hidden units was selected for the final cost, between 2 and 50 for each case : the neural network trained with all the input variables, the neural network which uses the $K^* = 11$ variables found with the dimensionality reduction method described above, the neural network where we kept the first $K^* = 11$ eigenvectors found with the PCA and the neural network which uses the $K^* = 11$ variables found with the variable selection method proposed above (the result for a given size of neural network is the average estimated on 20 attempts). The P^* values are respectively 12, 10, 6 and 10.

Decision Tree

We used a commercial version of the algorithm C4.5 [Ext-57]. The principal training parameters and the pruning conditions are :

– the splitting on the predictor variables continues until all terminal nodes in the classification tree are "pure" (i.e., have no misclassification) or have no more than the minimum of cases computed from the specified fraction of cases (here 100) for the predicted class for the node ;
– the Gini measure that reaches a value of zero when only one class is present at a node (with priors estimated from class sizes and equal misclassification costs).

With these parameters the number of variables used by the decision tree is 17, that is more than $K^* = 11$.

Bayesian Network

The Bayesian network (BN) found is similar to the "Naïve Bayes" which assumes that the components of the measurement vector, i.e. the features, are conditionally independent given the class. Like additive regression, this assumption allows each component to be modeled separately. Such an assumption is very

restrictive but on this real problem a naïve Bayes classifier gives very good results (see [Ext-58]). The BN uses 37 variables[2], which is more than three times more than $K^* = 11$.

2.5.2 Results

The various classification performances are given below in the form of lift curves. The methodology described in the chapter gives excellent results (see Figure 2.11)

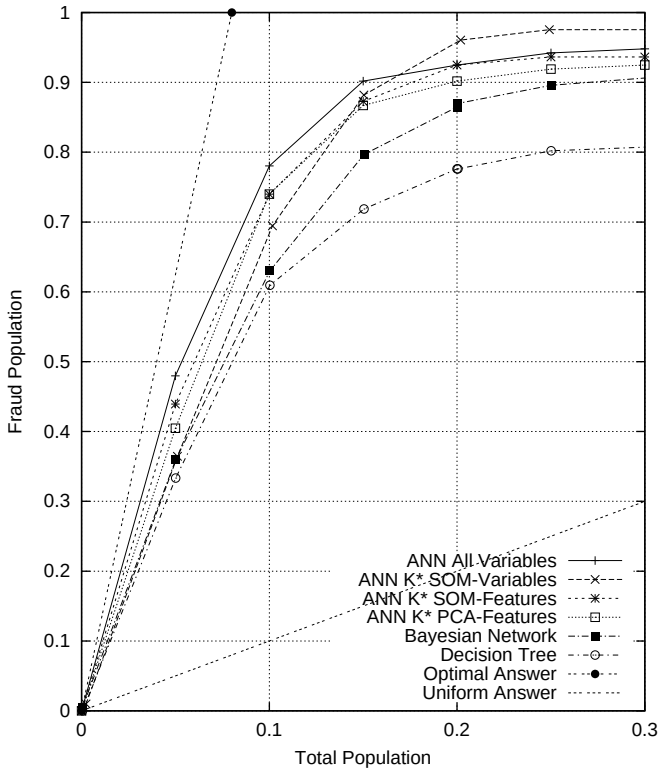

FIG. 2.11 – Detection rate (%) of the fraudulent users obtained with different learning methods (ANN : Artificial Neural Network), given as a lift curve.

Regarding the variable selection method, the performances of the neural network trained with the selected variables are better than the performances of the Decision Tree and the Bayesian Network. As compared to the neural network trained with all the input variables (226 variables), the neural network trained with the selected variables only ($K^* = 11$ variables) shows a marginal degradation of the performance for very small segments of the population and even has a slightly better behavior for segments larger than 20% of the population. The SOM-based dimensionality reduction method has a performance similar to the PCA-based method.

These comparisons show that, on this real application, it is possible to obtain excellent performances with the methodology described above and in particular with the variable selection method, hence allowing a much simpler interpretation of the model as it only relies on a few input variables.

[2]The BN was built by Prof. Munteanu and coworkers, ESIEA, 38 rue D. Calmette Guérin 53000 Laval France, in the framework of the contract “Bayes-Com” with France Telecom. The research report is not available.

2.6 Conclusion

In this chapter we have presented a SOM-based methodology for exploratory analysis, dimensionality reduction and/or variable selection. This methodology has been shown to give excellent results on a real case study when compared with other methods both in terms of visualization/interpretation ability and classification performance.

We have successfully applied the methodology described in this chapter on a variety of problems. It allows :

- to track down characteristic behavior of cases ;
- to visualize synthetically various behaviors and their relationships ;
- to group together variables which contribute in a similar manner to the constitution of the clustering of cases ;
- to analyze the contribution of every group of variables to every group of cases ;
- to realize a selection of variables ;
- to make all interpretations with the initial variables.

Another example of the application of this methodology can be found in [Ext-59]. The authors show how SOM can be used to build successive layers of abstraction starting from low-level traffic data to achieve an interpretable clustering of customers and how the unique visualisation ability of SOM makes the analysis quite natural and easy to interpret.

Chapitre 3

Looking for a relevant similarity criterion for HRTF clustering

Contents

For high-fidelity Virtual Auditory Space (VAS), binaural synthesis requires individualized Head-Related Transfer Functions (HRTF). An alternative to exhaustive measurement of HRTF consists in measuring a set of representative HRTF in a few directions. These selected HRTF are considered as representative because they summarize all the necessary spatial and individual information. The goal is to deduce the HRTF in non-measured directions from the measured ones by appropriate modeling. Clustering is applied in order to identify the representative directions, but the first issue relies on the definition of a relevant distance criterion. The paper presents a comparative study of several criteria taken from literature. A new insight in HRTF (dis)similarity is proposed.

3.1 Introduction

Binaural synthesis is a powerful tool for rendering 3D audio scene. Sound spatialization is based on binaural filters derived from the Head-Related Transfer Function (HRTF), which describes the acoustic path between the sound source and the listener's ears. HRTF highly depends on the individual morphology, but acquiring individualized HRTF is still a key issue of current research in binaural technologies. One solution is HRTF measurement, which is quite long and uncomfortable for subjects [Ext-60] [Ext-61] [Ext-62] [Ext-63]. What's more individual HRTF measurement should be discarded for commercial use of binaural technologies on a massive scale. Another solution is BEM modeling [Ext-64] [Ext-65], but this method does not provide accurate modeling in high frequencies because of computational limitations.

A third approach is investigated in the present paper. The idea is to measure HRTF only in a few directions. It is based on data reduction performed by HRTF clustering [Ext-66] [Ext-67]. The HRTF database is analyzed according to a given criterion of HRTF similarity focused on the magnitude spectrum[1] of HRTF. As a result, the HRTF are grouped into several clusters, which denotes the main features of HRTF. For each cluster, a representative HRTF is identified as the closest HRTF to all the HRTF contained in the cluster. Therefore, it is intended that one given HRTF in any direction can be deduced from the HRTF measured in the representative directions, which suggests a simplified protocol of HRTF measurement. Several methods, such as HRTF interpolation [Ext-67] or neural network modeling, are available for deriving HRTF in any direction from the representative HRTFs. This issue will not be dealt with in the present paper, which will be focused on the first step of HRTF clustering.

However, HRTF clustering relies on a similarity or distance criterion, which should be carefully defined according to the data considered. Several distance criteria designed for HRTF are available from literature, however, they are not specific to HRTF clustering. It is intended to compare them when they are applied to HRTF clustering. First, an overview of HRTF (dis)similarity criteria[2] will be given. Five distance criteria are selected. They will be first examined only from the point of view of HRTF (dis)similarity (*A priori* assessment), disregarding clustering purposes. Then their performances for HRTF clustering will be assessed (*A posteriori* assessment), after a brief recall of clustering methodology. The paper will conclude by summarizing the main results of the two studies (*a priori* and *a posteriori* assessments). By merging the two points of view, it will be investigated whether one particular criterion stands out from the others or not.

3.2 Overview of distance criteria used for HRTF similarity

The goal of a distance criterion for HRTFs is to quantify the (dis)similarity between two HRTFs. In the present paper, the HRTF (dis)similarity will be judged only from the point of view of signal processing. The HRTFs are compared according to their magnitude spectrum. The distance criterion will be used here for clustering purposes. It is intended to identify common features within the HRTFs of a whole database, in order to sort HRTFs by similarity. Apart for clustering, distance criteria are also required for HRTF modeling purposes, in order to compare the modeled HRTF with the original one. For these various problems, several distance criteria have been defined. An exhaustive list of all the criteria available from literature is beyond the scope of our study. Only five "standard" criteria will be considered in the following.

[1] The phase spectrum, which is related to temporal cues such as ITD (Interaural Time Difference), is not considered here.

[2] It should be noticed that some criteria used in the following may be not considered as "pure" distance criteria according to a mathematical sense, insofar as they do not fulfill all the properties of a mathematical distance.

3.2.1 Definition of the distance criteria

MSE Criteria

The first criteria, which is certainly the most obvious, is the well-known MSE (Mean Square Error) distance criterion. It is defined as :

$$C_{MSE} = \frac{1}{N}\sum_{i=1}^{N}[H_1(i) - H_2(i)]^2 \tag{3.1}$$

where $H_1(i)$ is the magnitude spectrum of one HRTF and $H_2(i)$ that of another HRTF. The index *i* refers to the frequency index, and N is the number of FFT points.

The MSE criterion can be modified by taking into account the frequency selectivity of the auditory system [Ext-68]. Since the auditory ability of frequency analysis is poorer for high frequencies than for low frequencies, it is proposed to lower the high frequencies part by frequency weighting. The frequency selectivity is well described by the concept of the critical bands, the bandwidth of which follows the frequency resolution of the auditory system. The critical bandwidth is 100 Hz for low frequencies (frequencies below 500 Hz) and increases up to 3500 Hz for f=13500 Hz. Its value (in Hz) is given for frequency *f* (in kHz) by ("Munich" Formula [Ext-68]) :

$$\Delta(f) = 25 + 75(1 + 1.4F^2)^{0.69}. \tag{3.2}$$

Thus the frequency weights $\alpha(i)$ are computed as the inverse of the critical bandwidth :

$$\alpha(i) = \frac{1}{a_0\Delta(f_i)} \tag{3.3}$$

where a_0 is a normalization value :

$$a_0 = \sum_{i=1}^{N}\frac{1}{\Delta(f_i)}$$

ensuring that :

$$\sum_{i=1}^{N}\alpha(i) = 1.$$

Fig. 3.1 illustrates the frequency weights. The MSE criterion including frequency weighting according to the critical bands (which will be referred to as the *CB criterion*) is thus defined by :

$$C_{CB} = \frac{1}{N}\sum_{i=1}^{N}\{\alpha(i)[H_1(i) - H_2(i)]\}^2 \tag{3.4}$$

Fahn Criterion

HRTF clustering has been already investigated by Fahn & al for HRTF interpolation purposes [Ext-67], a problem very close to our study. The memory cost of binaural synthesis is high if the HRTF measured for all the directions are stored. One solution is to interpolate HRTF in any direction from a limited number of HRTF stored in a few directions. But there are many ways to chose these "useful" HRTF. One of this method is clustering and Fahn & al showed that this latter gave better interpolation than uniform sampling. The performance evaluation was based on a "reconstruction error" defined as :

$$C_F = \frac{\sum_{i=1}^{N}[H_1(i) - H_2(i)]^2}{\sum_{i=1}^{N}[H_1(i)]^2} \tag{3.5}$$

This criterion is the third distance criteria used in our study and will be called the *Fahn criterion*. It differs from the MSE criterion mainly by the fact that the MSE distance is weighted by the energy of one HRTF.

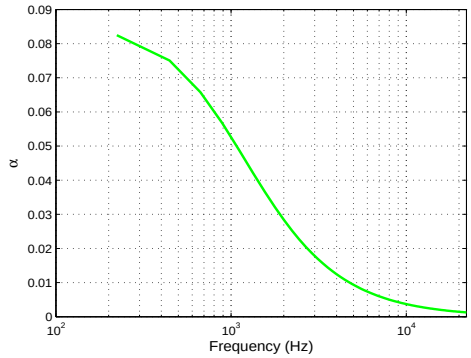

FIG. 3.1 – Frequency weighting according to critical band

Avendano Criterion

The fourth criterion is due to Avendano & al, who have introduced a new “error measure" in a paper about the modeling of the contralateral HRTF [Ext-69]. This error is based on the MSE distance expressed on a dB scale :

$$C_A = 10log10\left\{\frac{\sum_{i=1}^{N}[H_1(i) - H_2(i)]^2}{\sum_{i=1}^{N}[H_1(i)]^2} + 1\right\}. \tag{3.6}$$

Another advantage of this error criterion is that zero error (i.e. perfect modeling) does not lead to infinity, but to 0 dB, which is more relevant.

Durant Criterion

The last criterion is given by Durant & al in a study about filter design based on Genetic Algorithm for HRTF approximation. The authors have proposed a modified error measure computed as :

$$C_D = \frac{1}{N}\sum_{i=1}^{N}\left\{20log10\left[\frac{H_2(i)}{H_1(i)}\right] - \bar{d}\right\}^2 \tag{3.7}$$

with :

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{N}20log10\left[\frac{H_2(i)}{H_1(i)}\right].$$

In this criterion, the distance between the two HRTFs is evaluated by magnitude ratio instead of magnitude difference. As the Avendano criterion it is expressed on a dB scale. With the parameter $\bar{d}$, the authors intended to discard the effect of overall gain mismatch. This idea is clever for HRTF modeling, since the reproduction of the main features of the spectral magnitude (i.e. the peaks and notches) is the first goal. Often it is considered that the absolute level is secondary. From Equ. 3.7, it can also be noticed that the *Durant* criterion is similar to a variance.

3.2.2 *A priori* assessment

The previous criteria are first examined in order to assess how they account for HRTF (dis)similarity. The HRTF database of the CIPIC is considered [Ext-63]. Under the assumption that increasing angular

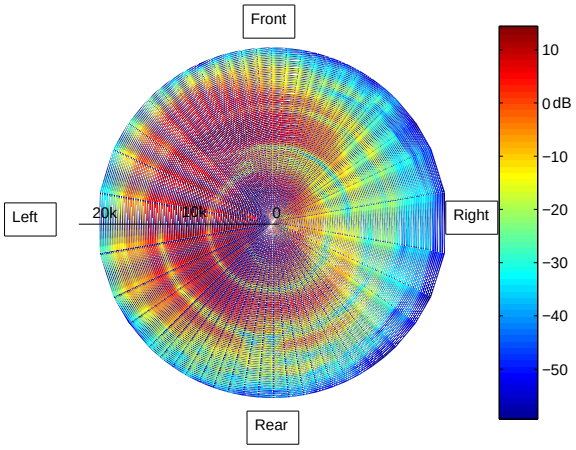

FIG. 3.2 – HRTF magnitude spectrum (dB) in function of the azimuth angle in the horizontal plane (Left ear of subject 003 taken from the CIPIC database) : “Polar" representation of HRTF magnitude spectrum, where the radius corresponds to the frequency axis and the angle to the azimuth angle.

difference between HRTF direction leads to increasing HRTF dissimilarity[3], the behavior of each criterion is checked for various pairs of HRTF corresponding to low and high angular mismatch.

Methodology

The goal is to collect a set of HRTF pairs with controlled dissimilarity, in order to assess the distance criteria. The HRTF database of the CIPIC [Ext-63], which provides a huge amount of HRTF data with 45 individuals (including one dummy head) and 1250 directions measured in the 3D sphere for each subject, will be used. Fig. 3.2 depicts the variation of the magnitude spectrum of HRTF in function of azimuth angle in the horizontal plane for one subject of the CIPIC database. Since the HRTFs illustrated in Fig. 3.2 are measured from the left ear, the HRTF magnitude is the highest on the left and decreases for right locations because of the acoustic diffraction induced by the head. Peaks and notches, which are mainly due to pinnae resonance, are also observed. Going from the left to the right, the magnitude spectrum varies quite continuously with the azimuth angle. Therefore it may be expected that when comparing two HRTFs located at azimuth angle θ_1 and θ_2 for instance in the front horizontal plane, their dissimilarity increases with their angular difference, which defines their *angle mismatch* :

$$d\theta = |\theta_2 - \theta_1|. \tag{3.8}$$

On the other hand, from a psychoacoustic point of view, we know that increasing angular mismatch leads to increasing error of localization. Thus it can be reasonably assumed that the HRTFs in the horizontal plane provides a wide range of HRTF dissimilarity. The HRTF dissimilarity relies both on signal processing (magnitude spectrum) and perception (localization error). However, before constituting the HRTF pair, it should be noticed that low dissimilarity may occur for strong angle mismatch because of the symmetry between front and rear HRTFs. For instance if the two HRTFs considered are taken from two locations which are symmetric with respect to the interaural axis, the HRTF are very similar despite a strong angular mismatch (cf. Fig. 3.2). Therefore it is preferred to consider separately the front and rear HRTFs in order to

[3]In this case, the HRTF dissimilarity is related to a localization error.

keep a confident link between the HRTF dissimilarity and the angle mismatch. On this condition, the HRTF dissimilarity varies in a monotone way with the angular mismatch.

In the CIPIC database, 25 directions are measured in the front horizontal plane, corresponding to azimuth angle varying from $-80°$ (on the left) to $80°$ (on the right), in the interaural polar coordinates. From these 25 HRTFs, 300 pairs are obtained with angular mismatch $d\theta$ varying from $5°$ to $160°$. Several pairs are associated to the same value of angular mismatch. In the same way, 300 pairs are also obtained from the rear horizontal plane. Thus a total of 1200 pairs, including 600 pairs both from the left and the right ear, is collected for each individual. The five dissimilarity criteria presented in Section 3.2 are then evaluated[4] for all the individuals available in the CIPIC database [Ext-63]. The data obtained from all the individuals, the left/right and the front/rear sets are merged for each value of angle mismatch. The statistical analysis of the criteria values in function the angle mismatch is presented in the following Section.

Results

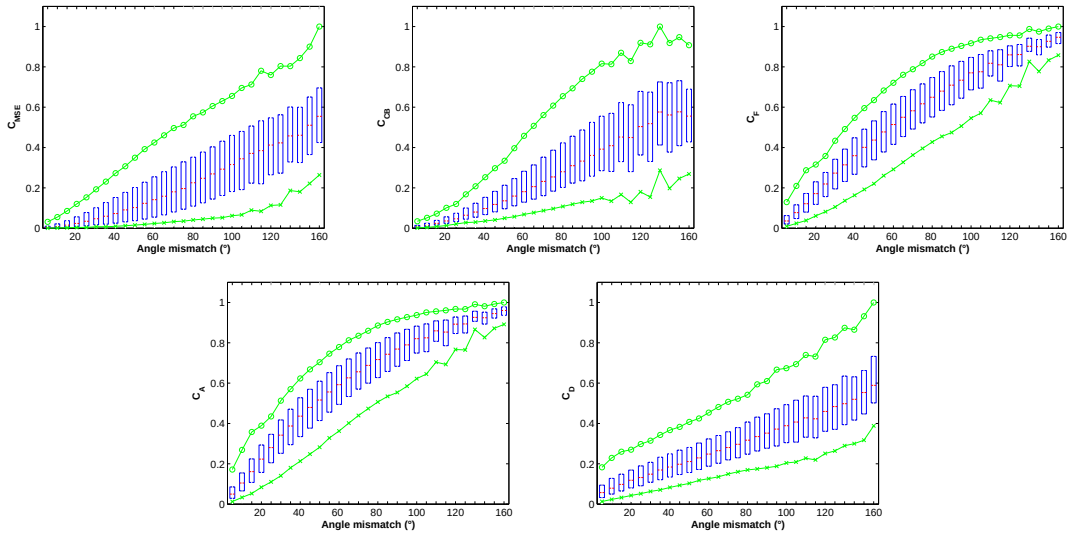

FIG. 3.3 – A priori assessment of the dissimilarity criteria : criterion values in function of the angular mismatch. From left to right and top to bottom : *MSE* criterion, *CB* criterion, *Fahn* criterion, *Avendano* criterion, and *Durant* criterion. The blue boxes describe the lower and upper quartile. The median is depicted by a red line. The green curves show the 5th (lower curve depicted by crosses) and 95th (upper curve depicted by circles) centiles.

The behavior of the various criteria in function of the angular mismatch are displayed in Fig. 3.3. The results are plotted as blue boxes delimited by the lower (25%) and upper (75%) quartile. The median is depicted by a red line. In addition, green curves show the 5th (lower curve depicted by crosses) and 95th (upper curve depicted by circles) centiles, in order to illustrate the extent of the rest of the data. Since the range of values strongly varies from one criterion to another, the values displayed are all normalized by the maximum value of the 95th centile, in order to compare the criteria with the same scale. The values obtained before normalization for 5, 10 and $15°$ angle mismatch are given in Tab. 3.1.

[4]When computing the criterion value, the HRTF considered as H_1 is always the HRTF with maximum energy, i.e. corresponding to the azimuth the more on the left for the left ear set and the more on the right for the right ear set. This choice has no influence for most of the criteria except for the *Fahn* and *Avendano* criteria, which include a normalization by the energy of the HRTF H_1.

Criterion	5°	10°	15°
C_{MSE}	0.0319	0.0677	0.116
C_{CB}	3.8310^{-6}	8.10^{-6}	1.3510^{-6}
C_F	0.0321	0.0691	0.108
C_A	0.137	0.29	0.446
C_D	11.8	16.3	20.3

TAB. 3.1 – Median value of the 5 criteria for 5, 10, 15° angle mismatch.

Criterion	5°	10°	15°
C_{MSE}	0.0624	0.139	0.241
C_{CB}	6.10^{-6}	8.8110^{-6}	1.30^{-6}
C_F	0.0376	0.0609	0.0827
C_A	0.158	0.246	0.323
C_D	12.7	16.0	17.1

TAB. 3.2 – Interquartile range of the 5 criteria for 5, 10, 15° angle mismatch.

A confident criterion should have monotone variation with increasing HRTF dissimilarity and low deviation for equivalent levels of dissimilarity, because it is intended to link criteria values with angular mismatches. From Section 3.2, it should be kept in mind that all the criteria are null or positive. The criterion value is null for perfect similarity (i.e. $H_1 = H_2$) and increases for increasing dissimilarity. In Fig. 3.3 the five criteria all exhibit almost linear increase with the angular mismatch. However, the values of the *Fahn* and *Avendano* criteria reach a ceiling for the highest angle mismatch (i.e. for mismatch greater than 100°). Except for *Durant* criterion, the deviation also increases with the angular mismatch. This phenomenon is particularly strong for the *MSE* and the *CB* criteria. Low deviation for small difference of azimuth is not surprising, since it can be observed from Fig. 3.2 that for small variation of azimuth, the HRTF variations are very close, whereas for greater variation of azimuth the HRTF variation are less consistent. The low deviation for small angular mismatch provides fine discrimination for low dissimilarity. From this point of view, the constant deviation of the *Durant* criterion is a drawback.

It is also worth examining the extent of criteria values (i.e. the range delimited by the lower and upper green curves in Fig. 3.3) in function of the angular mismatch. It is striking that the range of criteria values for a given angular mismatch is wider for the *MSE* and the *CB* criteria than for the other criteria. Particularly, for the *MSE* criterion, the 5th-centile curve keeps very close to zero whatever the angular mismatch is, which means that this criterion may give low value although the HRTF dissimilarity is quite strong, which is not confident. The same defect is observed for the *CB* criterion. On the contrary, the *Fahn* and *Avendano* criteria show narrow extent of values, which suggests that these criteria provide a fine discrimination of HRTF dissimilarity. From these results, these two criteria can be considered as the most suited as distance criteria for HRTF dissimilarity. The influence of magnitude smoothing [Ext-70] of HRTF spectrum has been also studied, but no difference with the previous results has been pointed out.

3.2.3 Criterion calibration

When using dissimilarity criteria, one difficulty is to link the criterion values with dissimilarity level in terms of the compared data. Considering two different pairs of HRTFs, if we suppose for instance that the *MSE* criterion gives a value of 0.03 for one pair and a value of 0.12 for the other, it is not obvious to know if these values denote low or high dissimilarity. From the previous analysis (cf. Section 3.2.2), we have now some knowledge about the physical and psychoacoutic meaning of the dissimilarity criteria. First, Tab. 3.1

shows that a *MSE* criteria value of 0.03 corresponds to an angular mismatch of 5°. The dissimilarity can be interpreted in two ways, by considering : either the difference between the HRTF magnitude spectrum or the localization mismatch between the two HRTFs. In Fig. 3.2, it can be observed that an angular mismatch of 5° leads to small variation of magnitude spectrum. In terms of auditory perception, a localization mismatch of 5° is very close to the lowest Minimum Audible Angle (MAA) [Ext-71] and so can be considered as hardly noticeable[5]. As a result, an angular mismatch of 5° is a low level of dissimilarity, whereas a mismatch greater than 10° corresponds to a noticeable level of dissimilarity, which allows us to calibrate each criterion. Tab. 3.1 gives the calibration values for the 5 criteria. Moreover, it is also interesting to know for a given step of decrease or increase of a criterion value whether this step is significant or not. The curves plotted in Fig. 3.3 can be used to interpret a given increase or decrease in terms of angular mismatch in order to assess its significance.

3.3 A posteriori comparison of distance criteria via HRTF clustering

After the previous *a priori* study, the present section will present an *a posteriori* study, where the five (dis)similarity criteria described in Section 3.2 will be assessed for clustering purpose.

Among clustering methods [Ext-72] the Self-Organizing Map (SOM) [Ext-48] is an excellent tool for data survey because it has prominent visualization properties. It creates a set of prototype vectors representing the data set and carries out a topology preserving projection of the prototypes from the N-dimensional input space onto a low-dimensional grid (two dimensions in the present paper). This ordered grid can be used as a convenient visualization surface for showing different features of the data[6] [MP-9]. The SOM method is used in the following sections to compare the five criteria, by judging their ability to produce an homogeneous clustering and low quantification errors.

3.3.1 Methodology - Organization of the experiment

The HRTF data used for the clustering are first presented. Then it is described how the criteria are included in the training of a SOM. Thirdly, the three axis of the experiment are explained.

2.3.1.1 The data

Clustering of one or two HRTFs

In the CIPIC database (see Section 3.2.2), each individual is represented by his(her) HRTFs for various azimuths and elevations (θ,ϕ) described in the interaural polar coordinates. A total number of 1250 directions is available for each individual (see figure 3.4).

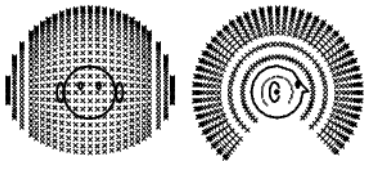

FIG. 3.4 – Graphical description of the 1250 directions (CIPIC database of HRTF).

[5]However it should be noticed that the auditory perception of HRTF mismatch is not so simple and can not be considered only as a pure localization mismatch in a thorough analysis. Perception of spectrum difference should also be taken into account. The present paper provides only a first analysis.

[6]These visualization surfaces are not shown in this paper because the main interest here is to rank the criteria.

For each position (θ,ϕ), the HRTF is therefore represented by a vector of 100 components, one component per frequency. In the following study, the input vectors considered for HRTF clustering consists either of 100 components (if only one ear is considered) or of 200 components (if the ipsilateral and contralateral HRTF are considered) (see Fig. 3.5).

To cluster HRTFs one can considers input vectors which are represented with 100 components if only one ear is considered and represented as a vector with 200 components if the two ears are considered (see Fig. 3.5).

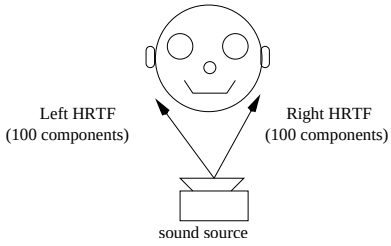

FIG. 3.5 – Using one or two HRTFs.

HRTF preprocessing

The amplitude scale of the raw HRTFs is linear. It is transformed into a logarithmic scale closer to our auditory perception than linear scale (see [Ext-73] for instance).

In terms of the amplitude range, we consider that a lowest threshold of $-80dB$ (10^{-4} in the linear scale of amplitude) is sufficient from a psychoacoustic point of view. The input vectors (HRTF) are transformed as follows :

$$Hl_{(\lambda,\theta,\phi)}(i) = 20\log_{10}\left(max(H_{\lambda,\theta,\phi}(i), 10^{-4})\right) \tag{3.9}$$

where H denotes a HRTF in the linear scale, Hl a HRTF in the logarithmic scale and λ refers to the individual.

An example of Hls is given in Fig. 3.6 for the individual 003 of the database and for three positions : (θ=-80, ϕ=-45), (θ=0, ϕ=90) and (θ=80, ϕ =230). Even on a log-scale, the spectra exhibit strongly localized features (i.e. peaks and notches) which are critical for the sound localization. The accurate modeling of such features from a few measurements only is therefore a real challenge.

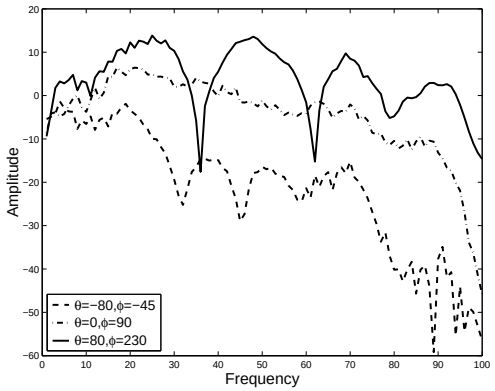

FIG. 3.6 – HRTFs measured for three directions : (θ=-80, ϕ=-45), (θ=0, ϕ=90) and (θ=80, ϕ =230) - Individual 003 of the CIPIC database.

Statistical Learning Set

When clustering data, it is well known that it is necessary to split the data into several sets : a **training set** used to adjust the parameters of the model and a **test set** to estimate the generalization error of the modeling (in order to prevent from over-training).

In our case, the CIPIC database is composed of 45 individuals, each described by 1250 or 2500 HRTFs. The data have been split into two sets : 23 individuals for the training set and 22 individuals for the test set. Therefore the training set consists of 1250 x 23 = 28750 vectors and the test set of 1250 x 22 = 27500 vectors.

2.3.1.2 Applying the five distance criteria into a SOM algorithm

The basic SOM algorithm

The basic SOM algorithm[7] is described below[8]. All the SOM in this chapter are square maps with hexagonal neighborhoods and are initialized with Principal Component Analysis (PCA). First the size of the Self-Organizing Map (SOM) [Ext-48] i.e k, the number of clusters and the topology of the SOM have to be fixed.

The basic SOM algorithm comprises five steps :

Given :

- (A) choose the number, k, of clusters (H_1 prototypes) ;
- (B) choose a topology of the map ;
- (C) initialization : choose random values for the k prototypes ;
- **Pour** (D) for all iteration t **faire**

  (D-1) random selection of an example H_2 from the training set,
  (D-2) election of the nearest prototype (the "winner") using a distance criterion,
  for example if the mean squared error is used

$$\arg\min_k \|H_1^k - H_2\| \tag{3.10}$$

$$\arg\min_k \left[\frac{1}{N}\sum_{i=1}^{N}[H_1^k(i) - H_2(i)]^2\right] \tag{3.11}$$

  where k denote the kth prototype of the map and i refers to the frequency index.
  (D-3) bring the winner near the example H_2 with a learning rate α ;
  (D-4) bring the neighbours of the winner, at this iteration t, near the example H_2
  (D-5) go to step D-1 until convergence is not reached[a]

  Fin Pour

[a] The convergence is obtained when winners do not move significantly.

Algorithme 1: SOM Algorithm

Projecting the position information (this information is not used for the construction of the map) on the map allows to investigate the distinctive profiles of the clusters in terms of position and dispersion (see Fig. 3.7).

[7] All the experimentation on SOM have been done with the SOM Toolbox package for Matlab © [Ext-52]

[8] Here the algorithm is presented in a very simple way just to introduce the "winner" notion, for more details see [Ext-48]

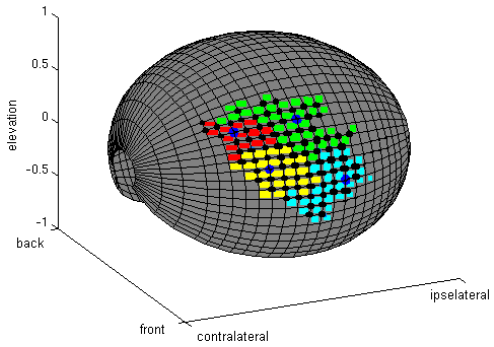

FIG. 3.7 – Illustration of clusters of HRTFs.

At the end of the training the whole test set is presented to the map. For each example of this test set a winner is selected using the criterion used to build the map.

The modified SOM algorithm

The training procedure includes the notion of "winner" using a distance criteria. It is straightforward that the equation 3.11 can be changed by any of the criteria described in section 2 to elect the winner.

It should be remarked that there is only a scale difference between the *Fahn* and *Avendano* criteria (Section 3.2) : the former is based on a linear scale, whereas the latter is logarithmic. In both cases, the election of the winner prototype gives the same result. The clustering results obtained with a SOM trained with either the *Fahn* or the *Avendano* criteria would be the same. Therefore only the *Avendano* results will be presented below.

2.3.1.3 Three axis of investigation

The study focuses on three issues : (1) the choice of the distance criterion, described in section 3.2 ; (2) the number of clusters (k) ; (3) the input data.

For (1) : the four distance criteria have been described above.

For (2) : The number of clusters corresponds to the SOM size. For instance a SOM with a topology 4x4 contains 16 clusters. The final aim of this study is to find few representative HRTFs so only small value of k are considered. SOM larger than 8x8 ($k = 64$) have not been examined[9].

For (3) : These experiment includes two ways of considering the input data. Either the left and right HRTF are considered independently, i.e. the input vector is : H = H_{1L} or H_{1R} (vector of length N). Or the left and right HRTF are pulled together (see Fig. 3.5), in order to take advantage of shared information between the left and right HRTF about the overall diffraction by the listener's head. The input vector is then : H = [H_{1L} H_{1R}] (vector of length 2*N). It is examined whether it is useful to consider both the ipsilateral and contralateral HRTF for describing a direction and if taking into account this solution provides any advantage.

[9]The following information may be useful for the readers who want to carry out the same experiments using Matlab Tool box [Ext-52]. The number of iterations for the rough tuning phase is 1500 for 2x2 and 4x4 SOMs, and 4000 iterations for 6x6 and 8x8 SOMs ; the number of iterations for fine tuning phase is 500 for all SOM size. The time need to train all the SOM used in this chapter has been 30 days on a Pentium IV 3.8 GHz.

3.3.2 Clustering results

2.3.2.1 Introduction

Three errors are defined to compare the clustering performances of the distance criteria.

• The **global average quantification error** is defined as :

$$Eq = \sum_{t=1}^{T} \sum_{p=1}^{P} \sum_{i=1}^{N} |H_1^k(i) - H_2(i)|, \tag{3.12}$$

where H_2 denotes an input HRTF presented to the map, H_1 represents one of the k prototypes (the winner according to a given criterion), T is the number of individuals in the test set, P is the number of considered positions and i refers to the frequency index. This global error versus the SOM size and the criteria is shown on Fig. 3.8(a). This error is thus a very global error which merges all individuals of the test set (22), all positions (1250) and all frequencies (100).

• The **average quantification error per position** whatever individual λ is defined as :

$$Eq(\lambda, \theta, \phi) = \sum_{i=1}^{N} |H_1^k(i) - H_2(i)| \tag{3.13}$$

The dispersion of this average quantification error per position is illustrated by Fig. 3.8(b) for each criterion, given a SOM size. This statistical analysis uses all the positions (1250) and all the individuals (22) of the test set. Therefore 27500 errors are aggregated inside each box plot.

• The **quantification error per frequency** is defined as :

$$Eq(i) = |H_1(i) - H_2(i)|, \tag{3.14}$$

The dispersion of the quantification error per frequency is depicted for each criterion in Fig. 3.8(c), 3.8(d), 3.8(e) and 3.8(f). As previously the statistics include all the positions (1250) and all the individuals (22) of the test set, which leads to 27500 errors for each frequency.

These three errors are used below to compare the four criteria considering one or two HRTFs as input data.

2.3.2.2 Clustering the right ear HRTFs

In this first experiment, the input data consists only of the right ear HRTFs. Fig. 3.8(a) shows the influence of the number of clusters on the average error (Equ. 3.12) for each criterion. The common trend is the decrease of the average error when the size of the SOM increases. Of course the error will be null if the number of clusters is equal to the number of vectors constituting the training set[10]. It is intended to reach a good compromise between the number of clusters and the error. From Fig. 3.8(a) it can be seen that a SOM of size 6x6 (36 clusters) gives a reasonable error for each criterion. What's more a SOM size of 8x8 does not provide a great improvement. Based on this error, the ranking order of the four criteria is (beginning from the best) : *MSE* (1), *Avendano* (1), *CB* (2) and *Durant* (3). The average quantification error obtained by the *MSE* and *Avendano* criteria for a SOM size of 6x6 is 3.8 dB. In terms of angular mismatch (cf. Section 3.2.2), this error value can be considered as equivalent to the level of dissimilarity observed in average between two HRTFs taken in the horizontal plane with an azimuth difference of $75°$. This is a strong dissimilarity, but the level of data reduction is also considerable, since clustering by a 6x6 SOM means that 27500 vectors are described by only 36 representatives.

[10] For instance the asymptotic result for the *MSE* criterion is close to 3 dB for 144 clusters [Ext-74].

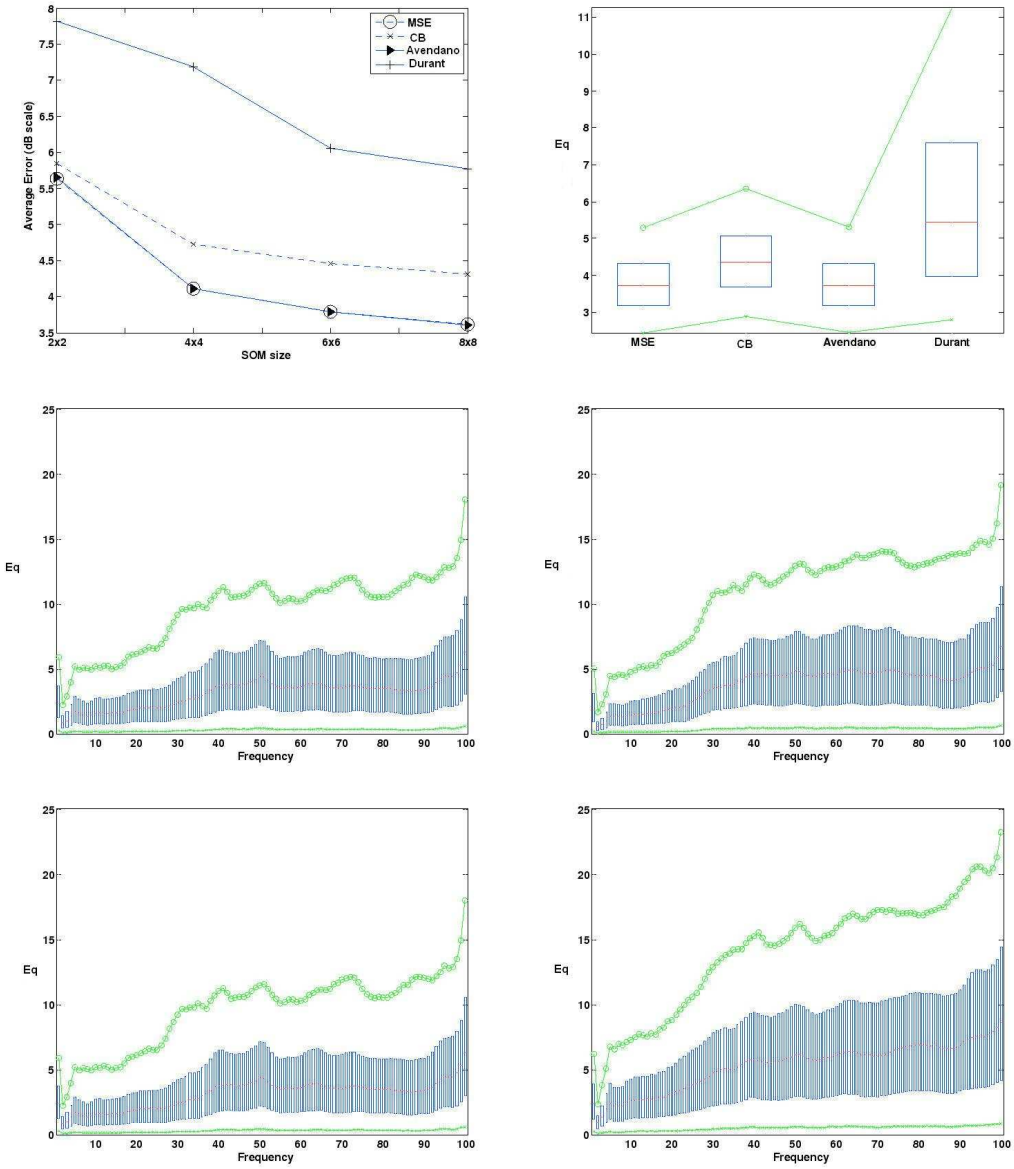

FIG. 3.8 – From left to right and up to down : (a) Average quantification error (dB) (see Equ. 3.12) versus SOM size for each criterion (*MSE*, *CB*, *Avendano* and *Durant* criterion) ; (b) Average quantification error per position (see Equ. 3.13) for each criterion for a SOM which contains 36 clusters (6x6) ; (c)(d)(e)(f) Average quantification error per frequency (see Equ. 3.14) respectively for the *MSE*, *CB*, *Avendano* and *Durant* criterion, for a SOM which contains 36 clusters (6x6). (c)(d)(e)(f) : The blue boxes describe the lower and upper quartile. The median is depicted by a red line. The green curves show the 5th (lower curve depicted by crosses) and 95th (upper curve depicted by circles) centiles.

Now the SOM size is fixed to 6x6 (36 clusters) :

– Fig. 3.8(b) describes the average error per position (Equ. 3.13) versus the criterion. The “best" criterion is the one which provides the smallest quantification error (i.e. the smallest median value) with low dispersion (i.e. narrow box plot). The same ranking order as in Fig. 3.8(a) is derived from Fig. 3.8(b), considering either the median value of the error or its dispersion. However it is still impossible to decide between the *MSE* and *Avendano* criteria.

– Fig. 3.8(c), 3.8(d), 3.8(e), 3.8(f) show the distribution of average errors (Equ. 3.14) in function of frequency for the four criteria. The criteria are judged according to the median value of the quantification error and the size of the box plot. The conclusions are the same as for the previous results (Fig. 3.8(a) and 3.8(b)) : i.e. *MSE* (1), *Avendano* (1), *CB* (2) and *Durant* (3).

As a result, the *MSE* and *Avendano* criteria stands out as the best criteria from this experiment. They should be considered as equivalent without further information.

2.3.2.3 Clustering both the right and left ear HRTFs

The results obtained when using both the ipsilateral and contralateral HRTF (H_{1L} and H_{1R}) are presented[11] in Fig. 3.10(a) to 3.10(f). The quantification error is slightly greater than when considering only the right ear HRTFs, but the difference is poorly significant[12]. A detailed analysis of the figures leads to the same ranking order of the distance criteria as previously.

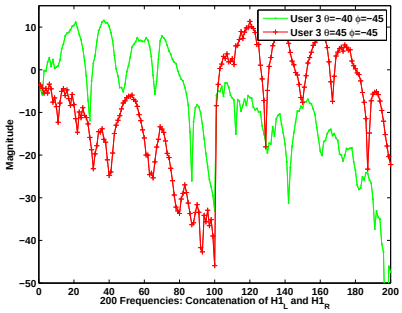

FIG. 3.9 – Combining the left and right ear. Each plot depicts the concatenated HRTFs $H_1 = [H_{1L}H_{1R}]$ represented in log-scale (see equation 3.9).

At first sight it may be surprising that considering both the ipsilateral and contralateral HRTF for describing a direction does not provide any advantage. This result suggests that the information contained in $H = [H_{1L}H_{1R}]$ is not greater, in the sense of clustering, than the information provided only by H_{1L} or H_{1R}. More precisely the additional information conveyed by the HRTFs of the second ear is greater than the information of the first ear for certain positions (i.e. when the second ear is the ipsilateral one), but noisier for other positions (i.e. when the second ear is the contralateral one). This phenomenon is illustrated in Fig. 3.9.

In Fig. 3.9, the 100 first components on both curve represent the signal perceived by the left ear and the 100 following components are the signal perceived by the right ear. The right ear is illuminated by the sound source for location ($\theta = -40$,$\phi = -45$), but is shadowed for location ($\theta = 45$,$\phi = -45$). This is the opposite for the left ear. It is obvious that including the second ear HRTFs in the clustering algorithm adds information for the first location, whereas it adds only noise for the second location. Therefore considering all the database using two HRTFs for each position does not give any improvement.

[11] The 200 frequencies are used to elect the winner (see section 3.3.1) but only the 100th frequencies are used to compute the errors presented here since one wants to compare to the results presented in section 3.3.2

[12] Except for Durant which is realy improved using the two HRTFs on a position

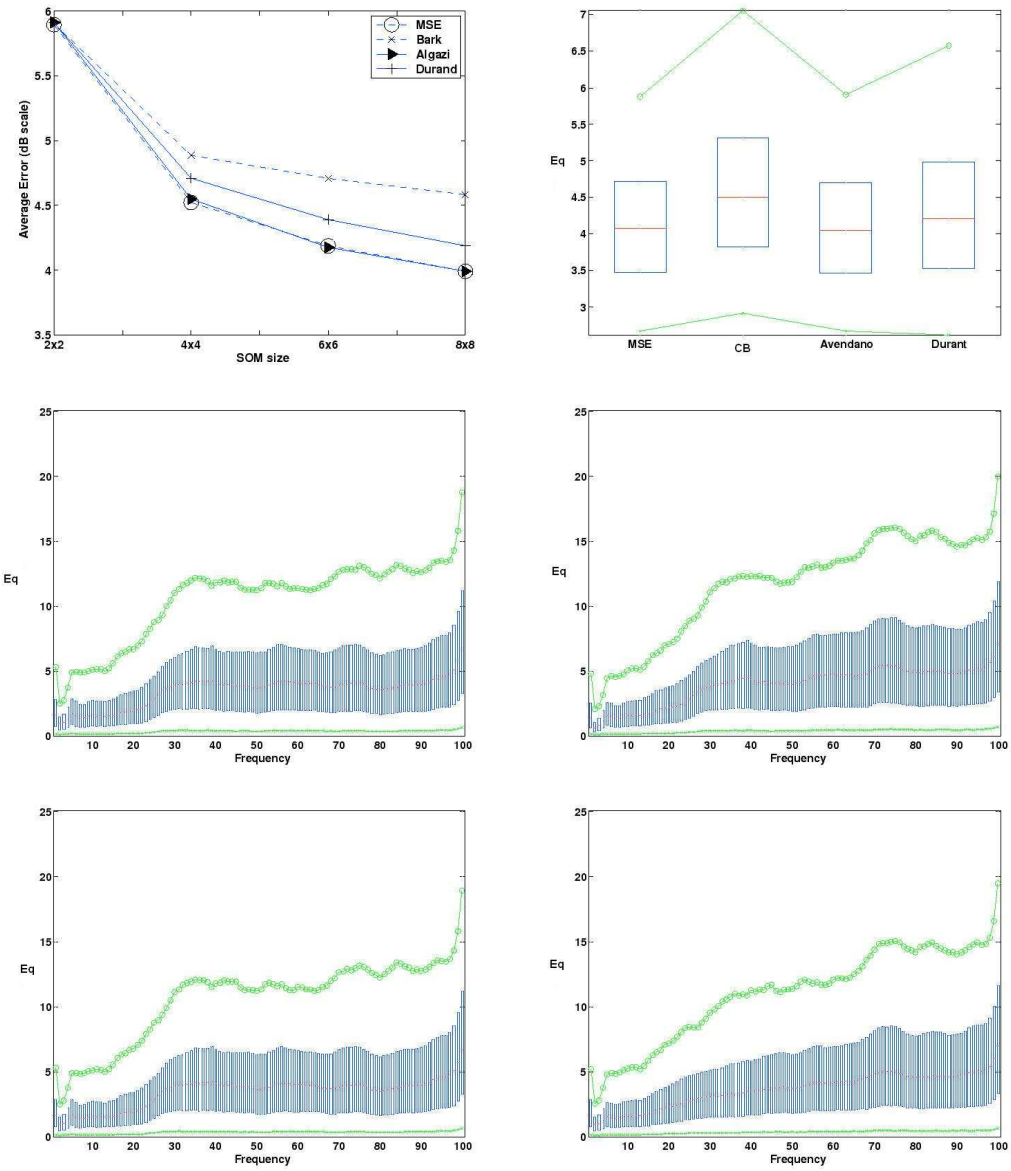

FIG. 3.10 – From left to right and up to down : (a) Average quantification error (Equ. 3.12) versus the SOM size for each criterion (*MSE*, *CB*, *Avendano* and *Durant*) ; (b) Average quantification error per position (Equ. 3.13) for each criterion for a SOM which contains 36 clusters (6x6) ; (c)(d)(e)(f) Average quantification error per frequency (Equ. 3.14) respectively for the *MSE*, *CB*, *Avendano* and *Durant* criterion, for a SOM which contains 36 clusters (6x6). (c)(d)(e)(f) : The blue boxes describe the lower and upper quartile. The median is depicted by a red line. The green curves show the 5th (lower curve depicted by crosses) and 95th (upper curve depicted by circles) centiles.

3.4 Conclusion

HRTF (dis)similarity has been investigated through five distance criteria taken from literature. The criteria were assessed and compared in two ways : first by examining their behavior towards a sample of HRTFs with “controlled" dissimilarities, which are linked to various levels of localization mismatch, second by evaluating their performances for HRTF clustering. It is striking that the two studies point out the same criterion, namely the *Avendano* criterion. In addition, it has been shown how to link the value of a distance criterion to a physical scale of HRTF dissimilarity, in order to know whether a given value means either a low or a high dissimilarity, which is of prime interest when using distance criteria.

HRTF clustering has been used successfully for reducing the size of a HRTF database. Input data, which consists of 27500 HRTFs, can be described by only 36 representatives. The study first considered only one HRTF by direction. It was also examined whether it is useful to consider both the ipsilateral and contralateral HRTF for describing a direction, but the results suggest that this solution provides no advantage.

From the HRTF representatives it is expected to derive clever modeling of individualized HRTF, which will be the next step. Beyond data reduction, HRTF clustering is also a powerful tool for investigating the spatial and individual dependence of HRTF, which could be analyzed in the light of auditory perception.

Chapitre 4

Combining several SOM approaches in data mining

Contents

The very rapid adoption of new applications by some segments of the ADSL customers may have a strong impact on the quality of service delivered to all customers. This makes the segmentation of ADSL customers according to their network usage a critical step both for a better understanding of the market and for the prediction and dimensioning of the network. Relying on a "bandwidth only" perspective to characterize network customer behaviour does not allow the discovery of usage patterns in terms of applications. In this paper, we shall describe how data mining techniques applied to network measurement data can help to extract some qualitative and quantitative knowledge.

4.1 Introduction

Broadband access for home users and small or medium business and especially ADSL (Asymmetric Digital Subscriber Line) access is of vital importance for telecommunication companies, since it allows them to leverage their copper infrastructure so as to offer new value-added broadband services to their customers. The market for broadband access has several strong characteristics :

- there is a strong competition between the various actors,
- although the market is now very rapidly increasing, customer retention is important because of high acquisition costs,
- new applications or services may be picked up very fast by some segments of the customers and the behaviour of these applications or services may have a very strong impact on the quality of service delivered to all customers (and not only those using these new applications or services).

Two well-known examples of new applications or services with possibly very demanding requirements in term of bandwidth are peer-to-peer file exchange systems and audio or video streaming.

The above characteristics explain the importance of an accurate understanding of the customer behaviour and a better knowledge of the usage of broadband access. The notion of “usage" is slowly shifting from a “bandwidth only" perspective to a much broader perspective which involves the discovery of usage patterns in terms of applications or services. The knowledge of such patterns is expected to give a much better understanding of the market and to help anticipate the adoption of new services or applications by some segments and allow the deployment of new resources before the new usage effects hit all the customers.

Usage patterns are most often inferred from polls and interviews which allow an in-depth understanding but are difficult to perform routinely, suffer from the small size of the sampled population and cannot easily be extended to the whole population or correlated with measurements [Ext-75]. “Bandwidth only" measurements are performed routinely on a very large scale by telecommunication companies [Ext-76] but do not allow much insight into the usage patterns since the volumes generated by different applications can span many orders of magnitude.

In this paper, we report another approach to the discovery of broadband customers’ usage patterns by directly mining network measurement data. After a description of the data used in the study and their acquisition process, we explain the main steps of the data mining process and we illustrate the ability of our approach to give an accurate insight in terms of usages patterns of applications or services while being highly scalable and deployable. We focus on two aspects of customers’ usages : usage of types of applications and customers’ daily traffic ; these analyses suppose to observe the data at several levels of detail.

4.2 Network measurements and data description

4.2.1 Probes measurements

The network measurements are performed on ADSL customer traffic by means of a proprietary network probe working at the SDH (Synchronous Digital Hierarchy) level between the Broadband Access Server (BAS) and the Digital Subscriber Line Access Multiplexer (DSLAM). This on-line probe allows to read and store all the relevant fields of the ATM (Asynchronous Transfer Mode) cells and of the IP/TCP headers. From now, 9 probes equip the network ; they observe about 18000 customers non-stop (a probe can observe about 2000 customers on a physical link). Once the probe is in place, data collection is performed automatically. A detailed description of the probe architecture can be found in [Ext-77].

4.2.2 Data description

For the study reported here, we gathered one month of data, on one site, for about two thousand customers. The data give the volumes of data exchanged in the upstream and downstream directions of twelve types of applications (web, peer-to-peer, ftp, news, mail, db, control, games, streaming, chat, others and unknown) sampled for each 6 minutes window for each customer. Most of the types of applications correspond to a group of well-known TCP ports, except the last two which relate to some well known but “obscure" ports (others) or dynamic ones (unknown). Since much of peer-to-peer traffic uses dynamic ports, peer-to-peer applications are recognized from a list of application names by scanning the payloads at the application level and not by relying on the well-known ports only. This is done transparently for the customers ; no other use is made of such data than statistical analysis.

Figure 4.1 plots the distribution of the total monthly traffic on the applications (all days and customers included) for one site in September 2003 (the volumes are given in bytes). About 90 percent of the traffic is due to peer-to-peer, web and unknown applications and all the monitored sites show a similar distribution.

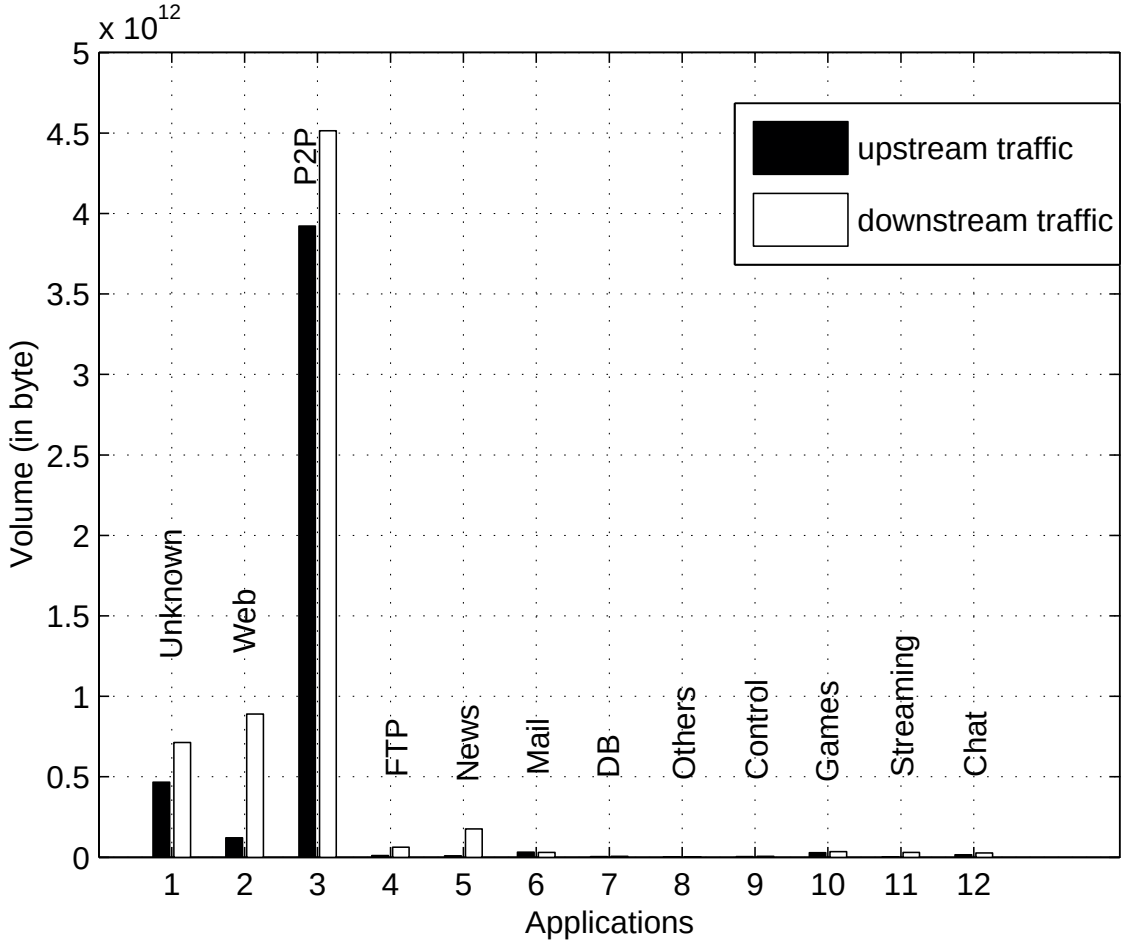

FIG. 4.1 – Volume of the traffic on the applications

Figure 4.2 plots the average hourly volume for the same month and the same site, irrespective of the applications. We can observe that the night traffic remains significant.

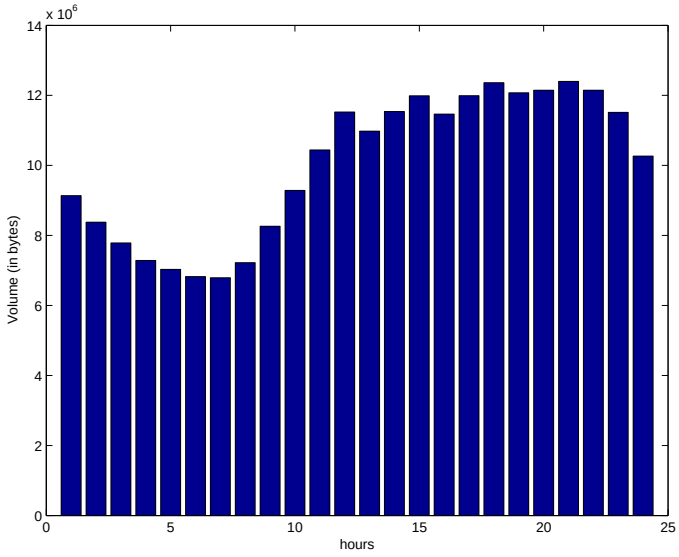

FIG. 4.2 – Average hourly volume

4.3 Customer segmentation

4.3.1 Motivation

The motivation of this study is a better understanding of the customers' daily traffic on the applications. We try to answer the question : **who is doing what and when ?**

To achieve this task we have developed a specific data mining process based on Kohonen maps. They are used to build successive layers of abstraction starting from low level traffic data to achieve an interpretable clustering of the customers.

For one month, we aggregate the data into a set of daily activity profiles given by the total hourly volume, for each day and each customer, on each application (we confined ourselves to the three most important applications in volume : peer-to-peer, web and unknown ; an extract of the log file is presented Figure 4.3). In the following, "usage" means "daily activity" described by hourly volumes. The daily activity profiles are recoded in a log scale to be able to compare volumes with various orders of magnitude.

4.3.2 Data segmentation using self-organizing maps

We choose to cluster our data with a Self Organizing Map (SOM) which is an excellent tool for data survey because it has prominent visualization properties. A SOM is a set of nodes organized into a 2-dimensional[1] grid (the map). Each node has fixed coordinates in the map and adaptive coordinates (the weights) in the input space. The input space is spanned by the variables used to describe the observations. Two Euclidian distances are defined, one in the original input space and one in the 2-dimensional space.

[1] All the SOMs in this chapter are square maps with hexagonal neighborhoods.

The self-organizing process slightly moves the location of the nodes in the data definition space -i.e. adjusts weights according to the data distribution. This weight adjustment is performed while taking into account the neighbouring relation between nodes in the map.

The SOM has the well-known ability that the projection on the map preserves the proximities : observations that are close to each other in the original multidimensional input space are associated with nodes that are close to each other on the map.

After learning has been completed, the map is segmented into clusters, each cluster being formed of nodes with similar behaviour, with a hierarchical agglomerative clustering algorithm. This segmentation simplifies the quantitative analysis of the map [Ext-78], [MP-9]. For a complete description of the SOM properties and some applications, see [Ext-48] and [Ext-79].

4.3.3 An approach in several steps for the segmentation of customers

We have developed a multi-level exploratory data analysis approach based on SOM. Our approach is organized in five steps (see Figure 4.6) :

• In a first step, we analyze each application separately. We cluster the set of all the daily activity profiles (irrespective of the customers) by application. For example, if we are interested in a classification of web down daily traffic, we only select the relevant lines in the log file (Figure 4.3) and we cluster the set of all the daily activity profiles for the application. We obtained a map with a limited number of clusters (Figure 4.4) : the typical days for the application. We proceed in the same way for all the other applications.

As a result we end up, for each application, with a set of “typical application days" profiles which allow us to understand how the customers are globally using their broadband access along the day, for this application. Such “typical application days" form the basis of all subsequent analysis and interpretations.

client	day	application	volume
client 1	day 1	unknown-up	volume-day-unknown-up-11
client 1	day 1	P2P-up	volume-day-P2P-up-11
client 1	day 2	unknown-up	volume-day-unknown-up-12
...	...	...	...
client 2	day 1	web-down	volume-day-web-down-21
client 2	day 3	unknown-up	volume-day-unknown-up-23
client 2	day 3	web-up	volume-day-web-up-23
client 2	day 3	web-down	volume-day-web-down-23
client 2	day 5	P2P-down	volume-day-P2P-down-25
...	...	...	...

FIG. 4.3 – log file : each application volume (last column) is a curve similar to the one plotted Figure 4.2

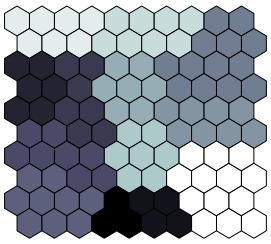

FIG. 4.4 – Typical Web-down days

• In a second step we gather the results of previous segmentations to form a global daily activity profile : for one given day, the initial traffic profile for an application is replaced by a vector with as many dimensions as segments of typical days obtained previously for this application.

The profile is attributed to its cluster ; all the components are set to zero except the one associated with the represented segment (Figure 4.5). This component is set to one. We do the same for the other applications. The binary profiles are then concatenated to form the global daily activity profile (the applications are correlated at this level for the day).

• In a third step, we cluster the set of all these daily activity profiles (irrespective of the customers). As a result we end up with a limited number of “typical day" profiles which summarize the daily activity profiles. They show how the three applications are simultaneously used in a day.

• In a fourth step, we turn to individual customers described by their own set of daily profiles. Each daily profile of a customer is attributed to its “typical day" cluster and we characterize this customer by a

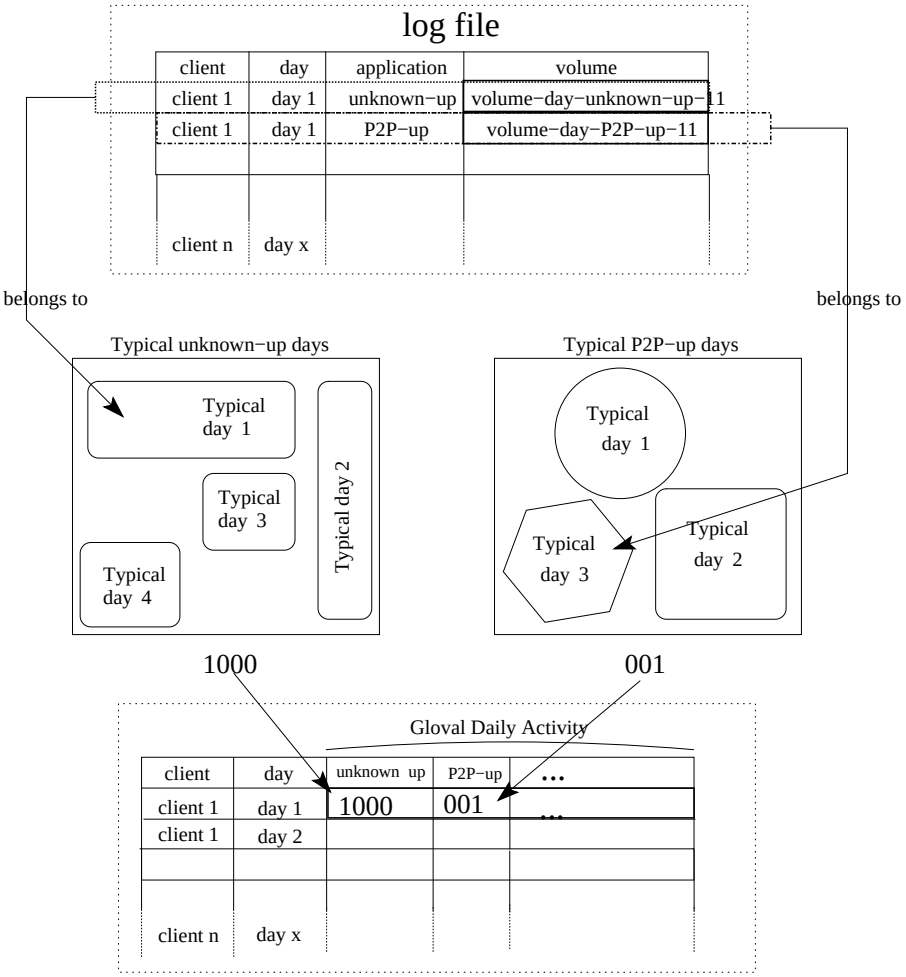

FIG. 4.5 – Binary profile constitution

profile which gives the proportion of days spent in each “typical day" for the month.

• In a fifth step, we cluster the customers as described by the above activity profiles and end up with “typical customers". This last clustering allows to link customers to daily activity on applications.

The process (Figure 4.6) exploits the hierarchical structure of the data : a customer is defined by his days and a day is defined by its hourly traffic volume on the applications. At the end of each stage, an interpretation step allows to incrementally extract knowledge from the analysis results. The unique visualization ability of the self organizing map model makes the analysis quite natural and easy to interpret. More details about such kind of approach on another application can be found in [Ext-80].

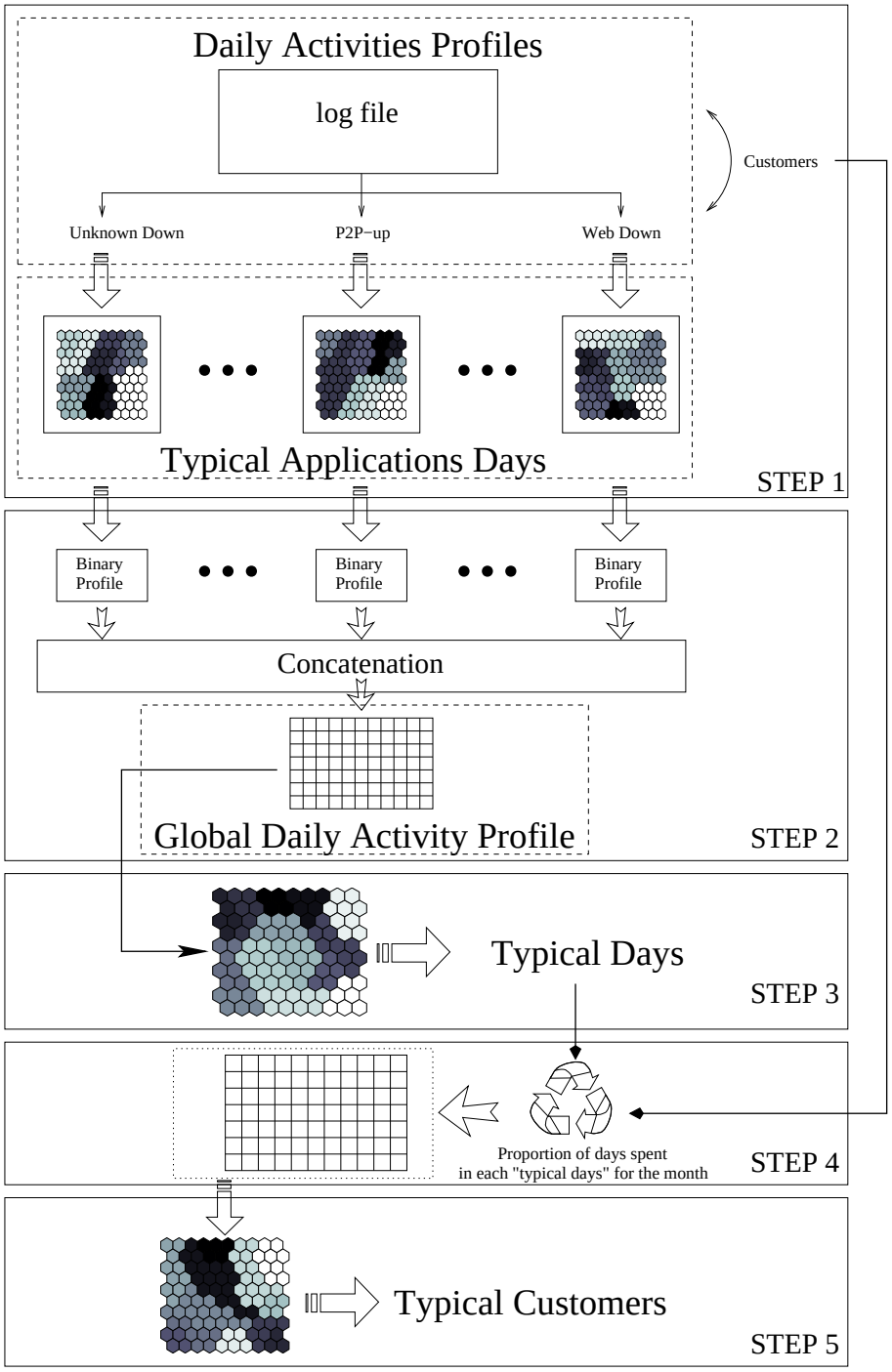

FIG. 4.6 – The multi-level exploratory data analysis approach.

4.3.4 Clustering results

We experiment with the site of Fontenay in September 2003. All the segmentations are performed with dedicated SOMs (experiments have been done with the SOM Toolbox package for matlab [Ext-52].

The first step leads to the formation of 9 to 13 clusters of “typical application days" profiles, depending on the application. Their behaviours can be summarized into inactive days, days with a mean or high activity on some limited time periods (early or late evening, noon for instance), and days with a very high activity on a long time segment (working hours, afternoon or night).

Figure 4.7 illustrates the result of the first step for one application : it shows the mean hourly volume profiles of the 13 clusters revealed after the clustering for the web down application (the mean profiles are computed by the mean of all the observations that have been classified in the cluster ; the hourly volumes are plotted in natural statistics). The other applications can be described similarly.

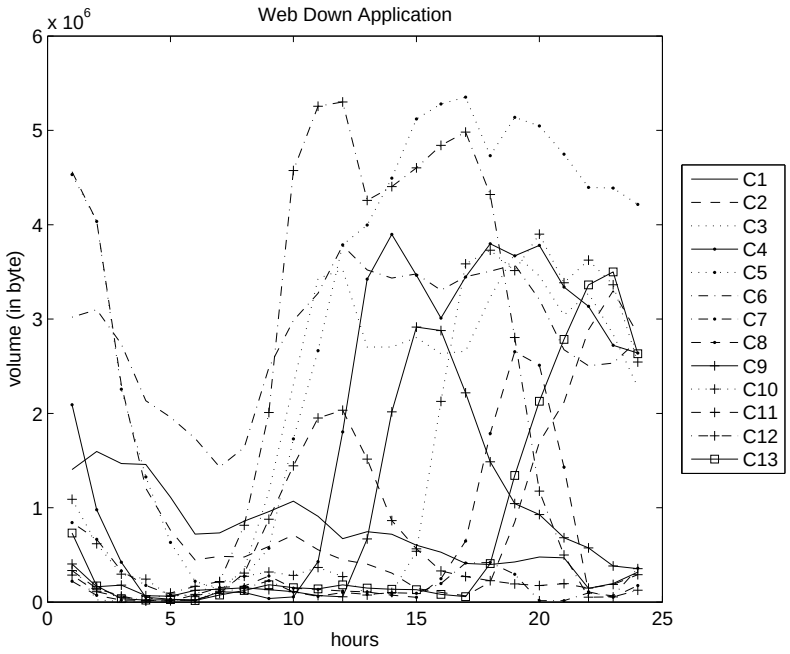

FIG. 4.7 – Mean daily volumes of clusters for web down application

The second clustering leads to the formation of 14 clusters of “typical days". Their behaviours are different in terms of traffic time periods and intensity. The main characteristics are a similar activity in up and down traffic directions and a similar usage of the peer-to-peer and unknown applications in clusters. The usage of the web application can be quite different in intensity. Globally, the time periods of traffic are very similar for the three applications in a cluster. 10 percent of the days show a high daily activity on the three applications, 25 percent of the days are inactive days. If we project the other applications on the map days, we can observe some correlations between applications : days with a high web daily traffic are also days with high mail, ftp and streaming activities and the traffic time periods are similar. The chat and games applications can be correlated to peer-to-peer in the same way.

The last clustering leads to the formation of 12 clusters of customers which can be characterized by the preponderance of a limited number of typical days.

Figure 4.8 illustrates the characteristic behaviour of one “typical customer" (cluster 6) which groups 5 percent of the very active customers on all the applications (with a high activity all along the day, 7 days out

of 10 and very little days with no activity). We plot the mean profile of the cluster (computed by the mean of all the customers classified in the cluster (up left, in black). We also give the mean profile computed on all the observations (bottom left, in grey), for comparison.

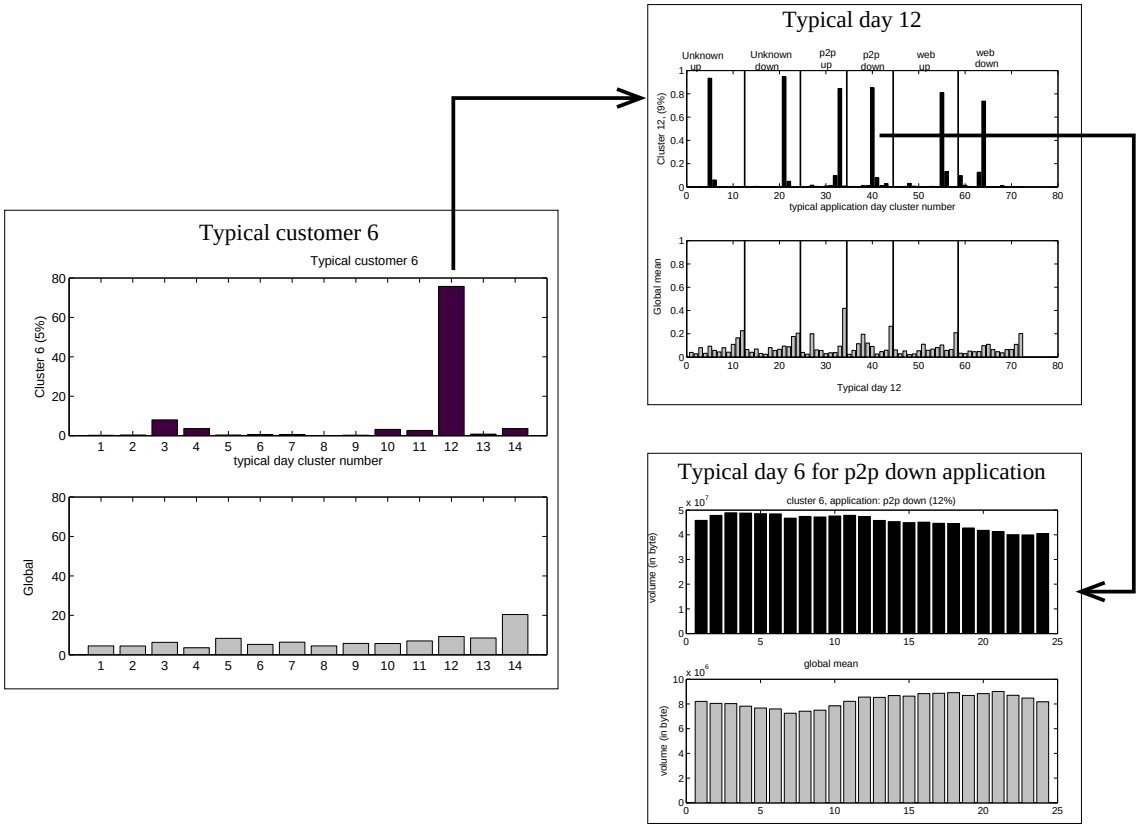

FIG. 4.8 – Profile of one cluster of customers (up left) and mean profile (bottom) and profiles of associated typical days and typical application days

The profile can be discussed according to its variations against the mean profile in order to reveal its specific characteristics. The visual inspection of the left part of Figure 4.8 shows that the mean customer associated with the cluster is mainly active on “typical day 12" for 78 percent of the month. The contributions of the other “typical days" are low and are lower than the global mean. Typical day 12 corresponds to very active days. The mean profile of “typical day 12" is shown in the right top part of the figure in black. The day profile is formed by the aggregation of the individual application clustering results (a line delimits the set of descriptors for each application). We also give the mean profile computed on all the observations (bottom, in grey).

Typical day 12 is characterized by a preponderant typical application day on each application (from 70 percent to 90 percent for each). These typical application days correspond to high daily activities.

For example, we plot the mean profile of “typical day 6" for the peer-to-peer down application in the same figure (right bottom ; in black the hourly profile of the typical day for the application and in grey the global average hourly profile ; the volumes are given in bytes). These days show a very high activity all along the day and even at night for the application (12 percent of the days). Figure 4.8 schematizes and synthesizes the complete customer segmentation process.

Our step-by-step approach aims at striking a practical balance between the faithful representation of the

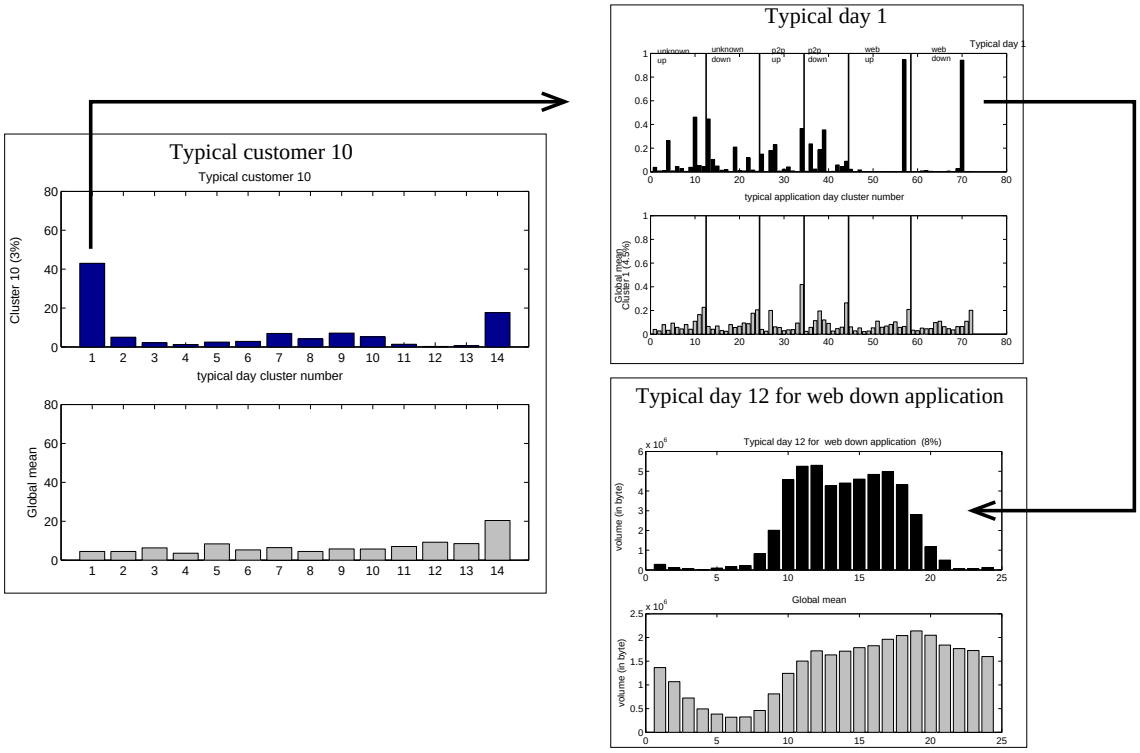

FIG. 4.9 – Profile of another cluster of customers (top left) and mean profile (bottom) and profiles of associated typical days and typical application days

data and the interpretative power of the resulting clustering. The segmentation results can be exploited at several levels according to the level of details expected. The customer level gives an overall view on the customer behaviours. The analysis also allows a detailed insight into the daily cycles of the customers in the segments. The approach is highly scalable and deployable and clustering technique used allows easy interpretations. All the other segments of customers can be discussed similarly in terms of daily profiles and hourly profiles on the applications.

We have identified segments of customers with a high or very high activity all along the day on the three applications (24 percent of the customers), others segments of customers with very little activity (27 percent of the customers) and segments of customers with activity on some limited time periods on one or two applications, for example, a segment of customers with overall a low activity mainly restricted to working hours on web applications. This segment is detailed in Figure 4.9.

The mean customer associated with cluster 10 (3 percent of the customers) is mainly active on “typical day 1" for 42 percent of the month. The contributions on the other “typical days" are close to the global mean. Typical day 1 (4.5 percent of the days) is characterized by a preponderant typical application day on web application only (both in up and down directions) ; no specific typical day appears for the two other applications. The characteristic web days are working days with a high daily web activity on the segment 10h-19h.

Figure 4.10 depicts the organization of the 12 clusters on the map (each of the clusters is identified by a number and a colour). The topological ordering inherent to the SOM algorithm is such that clusters with close behaviours lie close on the map and it is possible to visualize how the behaviour evolves in a smooth manner from one place of the map to another. The map is globally organized along an axis going from the

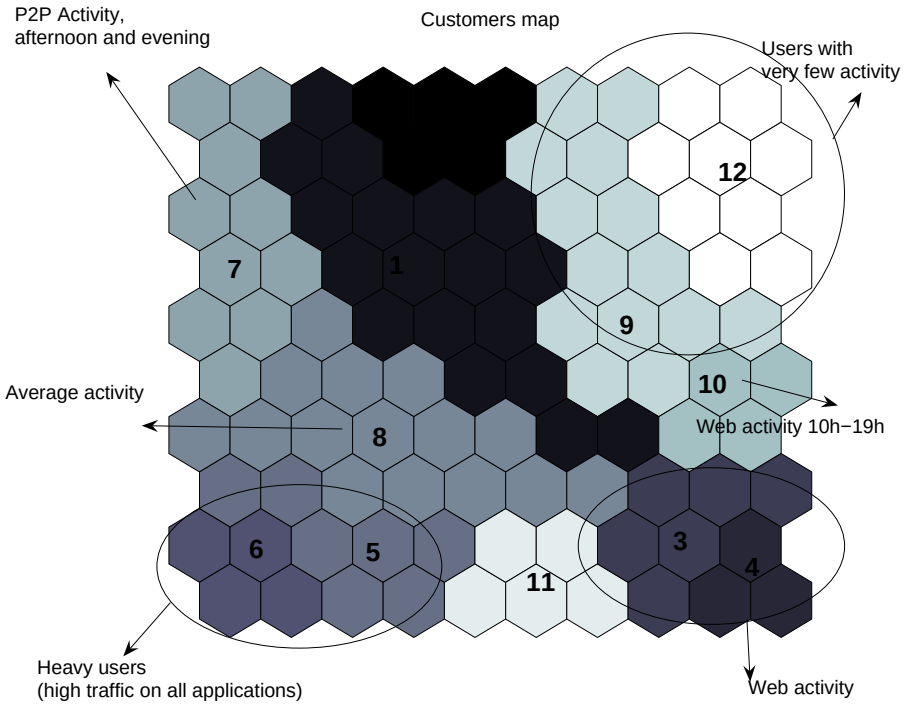

FIG. 4.10 – Interpretation of the learned SOM and its 12 clusters of customers

north east (cluster 12) to the south west (cluster 6), from low activity to high activity on all the applications, non-stop all over the day.

4.4 Conclusion

In this paper, we have shown how the mining of network measurement data can reveal the usage patterns of ADSL customers. A specific scheme of exploratory data analysis has been presented to give lightings on the usages of applications and daily traffic profiles. Our data-mining approach, based on the analysis and the interpretation of Kohonen self-organizing maps, allows us to define accurate and easily interpretable profiles of the customers. These profiles exhibit very heterogeneous behaviours ranging from a large majority of customers with a low usage of the applications to a small minority with a very high usage.

The knowledge gathered about the customers is not only qualitative ; we are also able to quantify the population associated to each profile, the volumes consumed on the applications or the daily cycle.

Our methodologies are continuously in development in order to improve our knowledge of customer's behaviours.

“Variable Selection and Model Interpretation”

Chapitre 5

A Input Variable Importance Definition based on Empirical Data Probability Distribution

Contents

Variable and feature selection have become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. We propose a new method to score subsets of variables according to their usefulness for the performance of a given model. This method is applicable on every kind of model and on classification or regression task. We assess the efficiency of the method with our results on the NIPS 2003 feature selection challenge and with an example of a real application.

5.1 Introduction

Up to 1997, when a special issue on relevance including several papers on variable and feature selection was published [Ext-54], few domains explored more than 40 features. The situation has changed considerably in the past few years, notably in the field of data-mining with the availability of ever more powerful data warehousing environments.

A recent special issue of JMLR [Ext-81] gives a large overview of techniques devoted to variable selection and an introduction to variable and feature selection can be found in this special issue [Ext-82]. A challenge on feature selection has been organized during the NIPS 2003 conference to share techniques and methods on databases with up to 100000 features.

The objective of variable selection is three-fold : improve the prediction performance of the predictors, provide faster and more cost-effective predictors, and allow a better understanding of the underlying process that generated the data.

Among techniques devoted to variable selection we find filter methods, which select variables by ranking them with correlation coefficients, and subset selection methods, which assess subsets of variables according to their usefulness to a given model.

Wrapper methods [Ext-83] use the elaborated model as a black box to score subsets of variables according to their usefulness for the modeling task. In practice, one needs to define : (i) how to search the space of all possible variable subsets ; (ii) how to assess the prediction performance of a model to guide the search and halt it ; (iii) which predictor to use.

We propose a new method to perform the second point above and to score subsets of variables according to their predictive power for the modeling task. It relies on a definition of the variable importance as measured from the variation of the predictive performance of the model (classification or regression). The method is motivated and described in section 5.2. Having presented the NIPS feature selection challenge in 5.3, we compare in section 5.4 the performance of the proposed method with other techniques on this challenge. Section 5.5 shows an example of application in a practical context and we conclude in 5.6.

5.2 Analysis of an Input Variable Influence

5.2.1 Motivation and previous works

Our motivation is to measure variable importance given a predictive model. The model is considered a perfect black box and the method has to be usable on a very large variety of models for classification (whatever the number of classes) or regression problems.

When a predictive model has been built, a question often raised in practice is '*What* would happen to this individual *if* this variable was set to a different value *?*'. A simple way to answer this question is to plot the variation of the output of this predictive model for this individual versus the variation of the variable [Ext-84; Ext-85; Ext-86].

For non-linear models the variation of the output can be non-monotonous. Hence, the influence of an input variable cannot be evaluated by a local measurement. The measurement of the difference of the output of a model with respect to the variation of an input variable provides a more global information and can be applied to discrete variables. However, the choice of the variation range should depend on the variable : too small a value has the same drawback as the partial derivatives (local information and not well suited for discrete variables), too large a value can be misleading if the function (the model) with respect to an input V is non-monotonous, or periodic.

A characteristic of the 'what if ?' simulation is that it relies on the generalization capabilities of the model since the output of the model is calculated with values of the variables which can be away from the training set ; for instance, a discrete variable can be treated as a continuous one. The 'what if ?' simulation is extended to define causal importance and saliency measurement by Féraud et al. in [Ext-87]. Their definition

however does not take into account the true interval of variation of the input variables. They propose to use a prior on the possible values of the input variables. The knowledge needed to define this prior depends on the specificities of the input variable (discrete, positive, bounded, etc). Such individual knowledge is clearly difficult and costly to obtain for databases with a large number of variables. A more automatic way than this 'prior' approach is needed.

A first step in this direction is given by Breiman in [Ext-88] (paper updated for the version 3.0 of the random forest) where he proposes a method which relies on the distribution of probability of the variable studied. Each example is randomly perturbed by randomly drawing another value of the studied variable among the values spanned by this variable across all examples. The performance of the perturbed set are then compared to the 'intact' set. Ranking variable performance differences allows to rank variable importance. This method allows to automatically determine the possible values of a variable from its probability distribution, even if perturbing every example only once does not explore the influence of the full probability distribution of the variable. Moreover, although [Ext-88] seems to restrict the method to random forests, it can obviously be extended to other models.

The method described in this chapter combines the definition of the 'variable importance' as given in Féraud et al. [Ext-87] with an extension of Breiman's idea [Ext-88]. This new definition of variable importance both takes into account the probability distribution of the studied variable and the probability distribution of the examples.

5.2.2 Definition of the variable importance

The importance of an input variable is a function of examples I (see Figure 5.1) probability distribution and of the probability distribution of the considered variable (V_j).

Let us define :

- V_j the variable for which we look for the importance ;
- V_{ij} the realization of the variable V_j for the example i ;
- I_m the example m a vector with n components ;
- f the predictive model ;
- $P_{V_j}(u)$ the probability distribution of the variable V_j ;
- $P_I(\nu)$ the probability distribution of examples I.

and

$$f_j(a;b) = f_j(a_1,...,a_n;b) = f(a_1,...,a_{j-1},b,a_{j+1},...,a_n) \tag{5.1}$$

where a_p is the p^{th} component of the vector a.

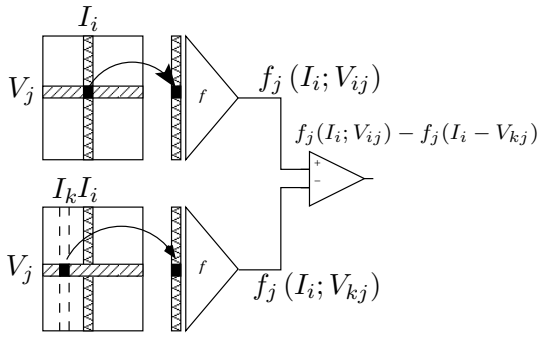

FIG. 5.1 – Graphical representation of the random draw

The importance of the variable V_j (see Figure 5.1) is the average of the measured variation of the predictive model output when examples are perturbed according to the probability distribution of the va-

riable V_j. The perturbed output of the model f, for an example I_i is the model output for this example but having exchanged the j$^{\text{th}}$ component of this example with the j$^{\text{th}}$ component of another example, k. The measured variation, for the example I_i is then the difference between the 'true output' $f_j(I_i; V_{ij})$ and the 'perturbed output' $f_j(I_i; V_{kj})$ of the model. The importance of the variable V_j is then the average of $|f_j(I_i; V_{ij}) - f_j(I_i; V_{kj})|$ on both the examples probability distribution and the probability distribution of the variable V_j. The importance of the variable V_j for the model f is then :

$$S(V_j|f) = \iint P_{V_j}(u)du P_I(v)dv \, |f_j\,(I_i; V_{ij}) - f_j\,(I_i; V_{kj})| \tag{5.2}$$

5.2.3 Computation

Approximating the distributions by the empirical distributions, the computation of the average of $S(V_j|f)$ would require to use all the possible values of the variable V_j for all examples available. For N examples and therefore N possible values of V_j the computation time scales as N^2 and become very long for large databases.

There are, at least, two faster heuristics to compute $S(V_j|f)$:

1) We draw simultaneously I_i and V_{kj} and compute one realization of $|f_j(I_i, V_{ij}) - f_j(I_i, V_{kj})|$. The measure of the average of $S(V_j|f)$ is then realized by means of a Kalman filter until convergence (see [Ext-89] to initialize and set the Kalman filter parameters).

2) $S(V_j|f)$ can be written :

$$S(V_j|f) = \int P_I(v)dv \int P_{V_j}(u)du \, |f_j\,(I_i; V_{ij}) - f_j\,(I_i; V_{kj})| \tag{5.3}$$

Approximating the probability distribution of the data by the empirical distribution of the examples :

$$S(V_j|f) = \frac{1}{N} \sum_{i \in N} E\left\{|f_j\,(I_i; V_{ij}) - f_j\,(I_i; V_{kj})|\right\} \tag{5.4}$$

As the variable probability distribution can be approximated using representative examples (P) of an ordered statistic :

$$S(V_j|f) = \frac{1}{N} \sum_{i \in N} \sum_{p \in P} |f_j\,(I_i; V_{ij}) - f_j\,(I_i; v_p)| \,\text{Prob}(v_p) \tag{5.5}$$

The computation can also be stopped with a Kalman filter. This method is especially useful when V_j takes only discrete values since the inner sum is exact and not an approximation.

5.2.4 Application to feature subset selection

The wrapper methodology [Ext-83] offers a simple and powerful way to address the problem of variable selection, regardless the chosen learning machine. The learning machine is considered a perfect black box and the method lends itself to off-the-shelf machine learning software packages. Exhaustive search can only be performed if the number of variables is small and heuristics are otherwise necessary. Among these, backward elimination and 'driven' forward selection which can both rely on the variable importance described above.

In backward elimination one starts with the set of all variables and progressively eliminates the least important variable. The model is re-trained after every selection step. In forward selection, as in [Ext-88], at a first step we train a model with all variables then we rank the variables using the method described in

this paper and in a second step we train models where variables are progressively incorporated into larger and larger subsets according with their ranks.

Comparison between both methods will be discussed elsewhere. Hereafter we restrict the discussion to backward elimination. We note here that both methods have the appealing property of depending on one parameter only, the degradation of the performance of the model trained with the subset relatively to the best possible performance reached.

To speed up the backward elimination another parameter is added. At each step of the backward elimination we remove all variables with an importance smaller than a very low threshold (10^{-6}). With this implementation the backward elimination method has only two simple parameters, a performance threshold to define the selected subset and an importance threshold to discard variables with 'no' importance.

5.3 Feature Selection Challenge

5.3.1 Introduction

Asserting the performance of data-mining methods is always a difficult task. Standard 'benchmark' problems such as the databases of the UCI repository are not well-suited to investigate the properties of variable selection techniques since most of the databases include only a small number of variables.

The purpose of the NIPS 2003 workshop on feature extraction was to bring together researchers of various application domains to share techniques and methods. Organizers of the challenge[1] formatted a number of datasets for the purpose of benchmarking feature selection algorithms in a controlled manner. The data sets were chosen to span a wide variety of domains. They chose data sets that had sufficiently many examples to create a large enough test set to obtain statistically significant results. The input variables are continuous or binary, sparse or dense. All problems however are two-class classification problems. The similarity of the tasks will allow participants to enter results on all data sets to test the genericity of the algorithms.

Each dataset was split in 3 sets : training, validation and test set. Only the training labels were provided. During the development period, challengers could send classification results (on the five datasets or on only one) and received in return validation set error rate. At any time the participants could submit their final classification results. A submission was considered final if the author(s) made a simultaneous submission on the five data sets before the deadline. A very large number of submissions were made on each dataset (840 for the most tried) but there were only 136 final submissions and 56 final valid submissions (organizers kept the five better results of every challenger).

5.3.2 Datasets

We describe here very briefly the five datasets. The number of examples for each train, valid and test set are given in Table 5.1. Manipulations of the datasets described below were performed by the organizers before the challenge.

- The task of ARCENE is to distinguish cancer versus normal patterns from mass-spectrometric data (continuous input variables). For data compression reasons organizers of the challenge thresholded the values. Before the benchmark linear SVM trained on all features had 15 % test error rate.
- The task of GISETTE is to discriminate between confusable handwritten digits : the four and the nine (sparse continuous input variables, many methods have been tried on this dataset, see `yann.lecun.com/exdb/mnist/`). The dataset was normalized so that the pixel values would be in the range [0,1] then values below 0.5 have been thresholded by the organizer to increase data sparsity. Before the benchmark linear SVM trained on all features had 3.5 % test error rate.

[1] All the informations about the challenge, the datasets, the results can be found on : `www.nipsfsc.ecs.soton.ac.uk`

– The task of DEXTER is to filter texts about 'corporate acquisitions' (sparse continuous input variables, see `kdd.ics.uci.edu/databases/reuters21578/`). The order of the features and the order pattern were randomized. Before the benchmark linear SVM trained on all features had 5.8 % test error rate.
– The task of DOROTHEA is to predict which compounds bind to Thrombin (sparse binary input variables). Before the benchmark 'lambda method' trained on all features had 21 % test error rate (no linear SVM tried).
– The task of MADELON is to classify artificial data (continuous input variables) with only 5 useful features. Before the benchmark organizers of the challenge used a K-nearest method, with $K = 3$, with the 5 useful features only which gives a 10 % error rate.

TAB. 5.1 – Data statistics

Dataset	Fraction of probes	Number of Features	Training set	Validation set	Test set
Arcene	30 %	10000	100	100	700
Gisette	50 %	5000	6000	1000	6500
Dexter	50 %	20000	300	300	2000
Dorothea	50 %	100000	800	350	800
Madelon	96 %	500	2000	600	1800

Probes refer to 'random features' distributed similarly to the real features and added to every dataset. This allows organizers to rank algorithms according to their ability to filter out irrelevant features.

5.4 Results and Comparison of the NIPS 2003 challenge

5.4.1 Test conditions on the proposed method

As we wished to investigated the performance of our variable importance measurement, we chose to use a single learning machine for all datasets (no bagging, no Ada-boost, no other bootstrap method) : a MLP neural network with 1 hidden layer, tangent hyperbolic activation function and stochastic back-propagation of the squared error as training algorithm. We added a regularization term active only on directions in weight space which are orthogonal to the training update [Ext-90].

For each dataset we split the training set in two sets : a training (70 %) and a validation set (30 %) ; the validation set of the challenge is then used as a test set. We made a single final submission before December first and we decided to keep this submission after December first (the valid submissions made before December first received the labels of the validation set, allowing a new attempt which was to be sent before December 8th). Therefore we compare below the results obtained with the proposed method with the valid results sent before December first.

The preprocessing used is only a zero-mean, unit-variance standardization. The strategy used to constitute the selected variable subset is the standard backward elimination. The subset of variables was chosen as the smallest subset allowing a performance greater than 95 % of the best performance reached during the selection process.

5.4.2 Comparison with others results

Comparison with baseline results

Our results compared to the baseline results, linear SVM and 'lambda method' (features selection by correlation with the target followed by Golub's classifier; see `clopinet.com/isabelle/Projects/NIPS2003/Slides/NIPS2003-Datasets.pdf`) are presented in Table 5.2.

TAB. 5.2 – Test Balanced Error Rate

Dataset	1	2	3	4	5	6
Arcene	30	1.5	-	15	30	29.65
Dexter	50	0.61	-	5.8	20	9.70
Dorothea	50	0.07	-	-	21	22.24
Madelon	96	1.6	10	-	41	16.38
Gisette	50	1.8	-	3.5	30	3.48

1 : Fraction of probes of the dataset (%)
2 : Fraction of features used (%)
3 : K-nearest BER (%)
4 : Linear SVM's BER (%)
5 : Lambda method's BER (%)
6 : Neural network BER (%),with the best subset of variables

The results presented in Table 5.2 show that all the results obtained are included between the results of the lambda method and the linear SVM. The Fraction Of Features (FoF) is defined as the ratio of the number of used variables by the classifier to the total number of variables in the dataset and the Balanced Error Rate (BER) as the average of the error rate on positive class examples and the error rate on negative class examples.

Clearly restricting ourselves to a simple model with no bootstrap techniques cannot allow us to reach very good BER, particularly on databases as ARCENE where the number of example is quite small.

Comparison with same model using all the variables

Our results compared to the results of a neural network trained with all the variables are presented in Table 5.3.

TAB. 5.3 – Test Balanced Error Rate

Dataset	(1)	(2)	(3)
Arcene	1.5	20.2	29.65
Dexter	0.61	15.1	9.70
Dorothea	0.07	30	22.24
Madelon	1.6	31.5	16.38
Gisette	1.8	4.3	3.48

(1) : Fraction of features used (%); (2) : Neural network BER (%), with all the variables; (3) : Neural network BER (%), with the best subset of variables.

The performance of the model trained with the subset relatively to the performance of the model trained with all variables are improved on every dataset excepted ARCENE. For this dataset it seems that the use of only 70 examples for training with no bootstrap does not allow to have good variable selection considering generalization performances.

An example of the BER obtained during the backward elimination phase is presented on the Figure 5.2 for GISETTE.

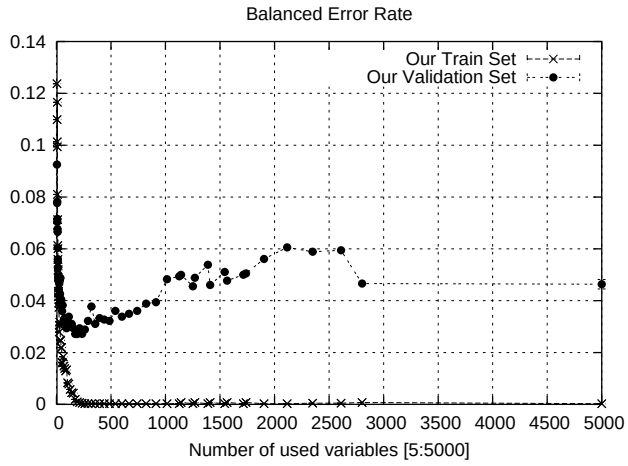

FIG. 5.2 – Balanced Error Rate during the backward elimination for GISETTE

The BER does not increase as we remove (backward elimination) variables until a very small number of variables is reached where the BER starts to be affected and increases sharply. Note that during the first step of the backward elimination we remove all variables with 'no' importance. For GISETTE this first step removes 44 % of the variables (there are 50 % of probes in GISETTE). At the end of the backward elimination we keep 90 variables, that is 1.8 % of the features, and we only have 5.56 % of probes, that is 5 'dummy' variables, among these 90 variables. This shows that for this database, even with a single MLP, the variable selection task has been well performed.

Comparison with all other valid submissions

The organizers of the challenge rated the classification results only with the BER. For methods having performance differences that are not statistically significant, the method using the smallest number of features win.

'Variable selection' is always somewhat ambiguous when the result is judged from the BER only, specially when different learning machines are used, since it is more a matter of balance between BER and FoF rather than a matter of BER only : to prefer a BER=0.1 using 50 % of features to a BER=0.12 using 10 % of the features is mostly a matter of application requirements. In some applications, one would trade some accuracy for less features, as in the real-time application we describe below for instance.

We first note that our results with a single MLP compare quite favorably with results by Amir Reza (results named 'SimpleNN' on the web site challenge) also using a single MLP and all features even if, as stated above, the use of a single learning machine without bootstrap does not lead to excellent BERs. However as can be seen below our BER results are close to the average results. The point here is just to stress that our model, although admittedly not the most adapted for accuracy on some datasets, indeed reaches a 'reasonable' BER (see also point 2) below).

What we expect from a variable selection technique is to adapt itself in such situation by removing as many features as possible. Therefore, what we can expect from the combination of our simple model and our selection technique is to keep a BER reasonably close to the average while using significantly less features on all datasets.

Below we use a representation of the results which allows a comparison of the proposed method to other methods on the five datasets. This representation has two axis (see Figure 5.3) : the first axis of comparison is the ratio between the BER of a submitted method and our BER (BER*) and the second axis is the ratio between our FoF (FoF*) and the FoF of a submitted method. For each dataset our results are placed in the center of the figure. Each author(s) is represented with a marker symbol. A marker is placed for each method of this author(s) and for each dataset. This allows to compare the results for each dataset. This simple representation allows to define 4 classes of methods on every dataset : 1) a better BER and a better FoF ; 2) a better BER but a worse FoF ; 3) a better FoF but a worse BER ; 4) a worse FoF and a worse BER. As authors were able to send more than one submission, authors may have more than five identical marker symbol.

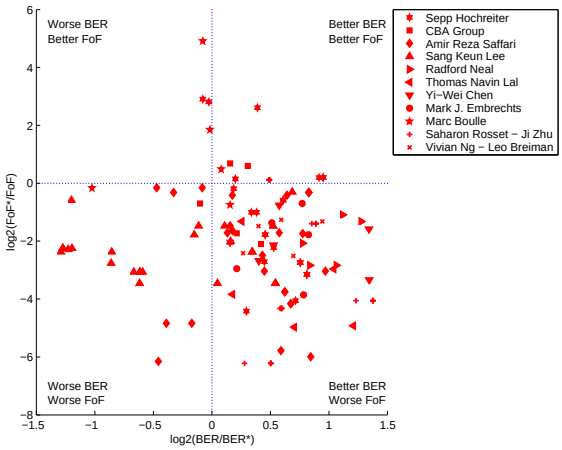

FIG. 5.3 – Results on the test set for all the valid final submissions (56) labelled by author(s).

The figure 5.3 shows results of methods which never used 100 % of the variables compared to the results obtained with our method. This figure shows that compared to the proposed method :

1. No method gave better results (better BER and better FoF) on the five datasets.
2. No method obtained a significantly better BER (less than 0.8 BER* regardless of the FoF) on the five datasets.
3. Half of the authors have tried a method which gives a worse BER and a worse FoF.
4. Only 4 authors proposed methods allowing to have a better BER and less features on some datasets.
5. The proposed method, combined with backward elimination using only one neural network, selects very few variables compared with the other methods.

The points above show that the proposed variable selection technique exhibits the expected behavior by both keeping the BER to a reasonable level (better than the BER with all features, except for ARCENE as already discussed, close to the average result of the challenge) and dramatically reducing the number of features on all datasets.

5.5 Application to fraud detection

The case study is the on-line detection of the fraudulent use of a post-paid phone card. Here the ‘fraud’ term includes all cases which may lead to a fraudulent non-payment by the caller. The purpose is to prevent non-payments by warning the owners of phone card that the current use of their card is unusual. The original database contains 15330 individuals described with 368 inputs variables of various natures. The database contains 97 % examples which belong to the class ‘not fraudulent’ and 3 % which belong to class ‘fraudulent’.

Using all variables in the modeling phase allows to obtain good fraudulent/non fraudulent classification performances but this model cannot be applied on-line because of computing and data extraction time constraints. It is thus necessary to reduce significantly the number of variables while keeping good performances.

The BER on the test set versus the number of variables is given in Figure 5.4. This figure shows that with the proposed method one can obtain the same BER with 100 variables than with 368 variables and a small degradation using 90 variables. Accepting a degradation of the performance by 10 % allows to retain only 40 variables.

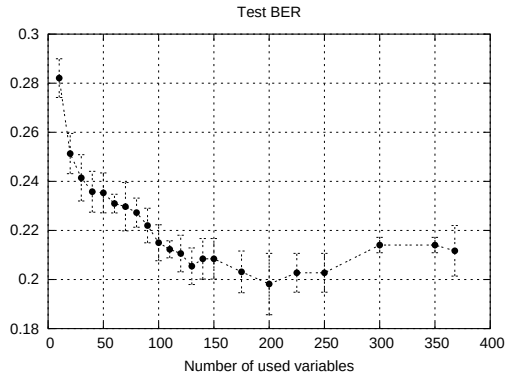

FIG. 5.4 – BER on the test set versus the number of variables.

The classification performances are given below in the form of lift curves in Figure 5.5 using 40, 90 and all the variables. Regarding the variable selection method, the performances of the neural network trained with 90 variables shows a marginal degradation of the performance as compared to the neural network trained with all the 368 variables. The neural network trained with 40 variables shows no degradation of the performance for small segments of the population : the selectivity is the same up to a lift ratio of 0.6 which is a key issue for such systems where only small segments of the population can be processed in real time.

These results show that, on this real application, it is possible to obtain excellent performances with the methodology described in this paper. Moreover, it allows a much simpler interpretation of the model as it only relies on much fewer input variables but such business-oriented discussion is out of the scope of this paper.

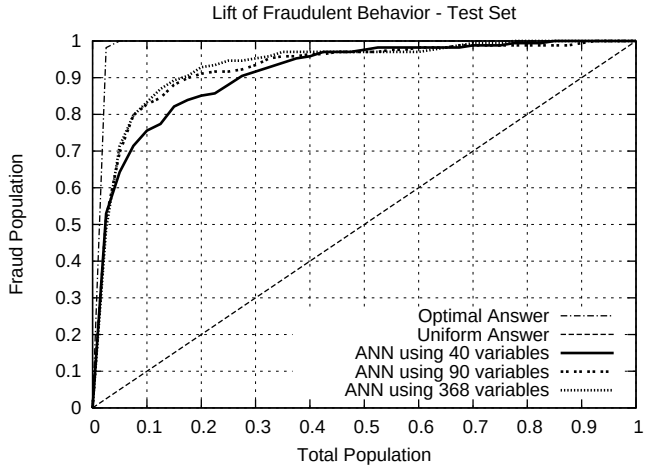

FIG. 5.5 – Detection rate (%) of the fraudulent users obtained with different number of variables (ANN : Artificial Neural Network), given as a lift curve.

5.6 Conclusions

We presented a new measure which allows to estimate the importance of each input variable of a model. This measure has no adjustable parameter, is applicable on every kind of model and for classification or regression task.

Experimental results on the NIPS 2003 feature selection challenge show that using this measure coupled with backward elimination allows to reduce considerably the number of input variables with no degradation of the modeling accuracy. Experimental results on a real application show the effectiveness of this approach.

Chapitre 6

Contact Personalization using a Score Understanding Method

Contents

This chapter presents a method to interpret the output of a classification (or regression) model. The interpretation is based on two concepts : the variable importance and the value importance of the variable. Unlike most of the state of art interpretation methods, our approach allows the interpretation of the model output for every instance. Understanding the score given by a model for one instance can for example lead to an immediate decision in a Customer Relational Management (CRM) system. Moreover the proposed method does not depend on a particular model and is therefore usable for any model or software used to produce the scores.

6.1 Introduction

The most elaborate way, in a CRM system, to build knowledge on customer is to produce scores. Tools which produce scores allow to project, on a given population, quantifiable information. The score is an evaluation for all instances of a target variable to explain. The score (the output of a model) is computed using input variables which describe instances. Scores are then "injected" in the information system (IS), for example, to personalize the customer relationship.

Nevertheless, sometimes the scores are not directly usable. For example if a scoring model identifies a customer interested in churning, the score does not say anything on the action needed to avoid his cancellation. To prevent this intention to churn, the fragility of the customer and its causes have to be identified.

We propose to solve this problem by interpreting the classification produced by the model for every instance. To make possible the industrial implementation of this solution we propose a completely automatic method. The interpretation of the score is delivered for every instance to feed the information system. This knowledge could then be exploited to provide information personalized in the customer relationship management.

The proposed method is independent of the model used to build the scores. The most powerful model can be used without changing the difficulty of its interpretation. This interpretation method could thus remove one of the principal difficulty of the use of models like Support Vector Machines (SVM), Random Forest (RF) or artificial neural networks (ANN)in the marketing services.

6.2 Positioning and previous works

6.2.1 Variable importance

V_j	:	an input variable j ;
X	:	a vector of J dimension ;
K	:	the number of training examples ;
X_n	:	a example n ;
X_{nj}	:	the component j of the vector X_n ;
F	:	the predictive model ;
p	:	the component p of the output vector ;
$F^p(X)$	:	the output value of the component p of the output vector of the model ;
and	:	$F_j^p(a;b) = F_j^P(a_1,\ldots,a_{j-1},b,a_{j+1},\ldots,a_J)$;

TAB. 6.1 – Notations

The field of machine learning abounds in techniques able to effectively solve problems of regression and/or classification. These techniques build a model from a training data base made up of a finite number of examples. The built model is used to associate an input vector to an output vector on a class label.

The large number of the models (linear regression, ANN, naive bayes, Random Forest (RF), Parzen window...) existing in the literature lead to a number of interpretation methods, generally specific to each model. The interpretation of the model is often based on : the parameters and the structure of the model [Ext-91], statistical tests on the coefficient's model [Ext-92], geometrical interpretations [Ext-93], rules [Ext-94] or fuzzy rules [Ext-95]. Resulting interpretations are often complex based on averages (for several individuals), for a given model (ANN, Decision Tree), or for a given task (regression OR classification).

Another approach consists in analysing the model as a black box with a sensibility analysis method. In these "What if ?" analyses, the structure and the parameters of the model are only needed to compute the output of the model. This independence gives valid interpretation methods whatever the model.

To analyze in detail the state of the art approaches, notations which will be used below in this chapter are introduced in table 6.1. In this table $F_j^p(a;b)$ denotes the output p of the model when the component j, value a, is replaced by the value b. The proposed method analyses the outputs of the model one by one. Therefore the simplified notation F_j will be used (instead of F_j^p). All calculations presented in this chapter are identical whatever the output p of the model.

Framling [Ext-96] introduces a variable importance measure, I, based on sensitivity analysis :

$I(V_j|F, X_n, p = [F_j(X_n, \max(V_j)) - F_j(X_n, \min(V_j))]/[\max[F(X_n), \forall n] - \min[F(X_n), \forall n]]$;

where $\max(V_j)$ and $\max(V_j)$ denotes respectively the maximum and the minimum value of V_j.

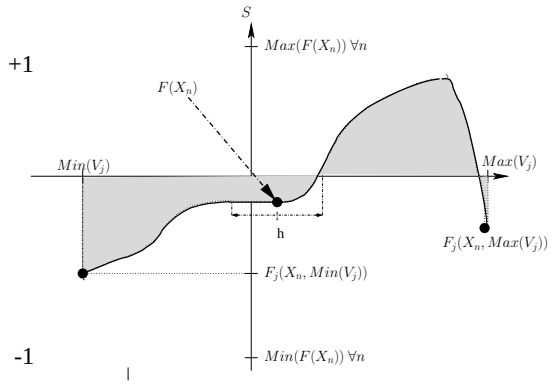

FIG. 6.1 – What if simulation : Output values of the model vs. values of V_j.

This measurement is interesting but can be misleading when F is not monotonous (see Figure 6.1). In this illustrative example, the variable V_j is important for the model F : according to the values of this variable an example can be classified in class $+1$ or -1. However $F(X_n, \max(V_j))$ and $F(X_n, \min(V_j))$ are close, which leads to underestimate the importance of the variable V_j. Moreover, this method is based on extremums variable and thus very sensitive to noise.

Another approach is based on the variation of the model output for a variation h of the variable V_j and an example X_n (see Fig. 6.1). When h tends towards zero, this measurement corresponds to the partial derivative of the model compared to the variable V_j. In this case, measurement is local and can give an erroneous importance measurement : the partial derivative at the point $F(X_n)$ is null for this example whereas the variable V_j is important. When h is larger, as in the previous case, this measurement can be misleading when F is not monotonous. The problem is the same when these measurements are averaged on all examples.

Feraud et al. [Ext-87] proposes a method based on the integral of the variations of the outputs model. This measurement is well adapted to non monotonous functions. On the illustrative example (see Fig. 6.1), this measurement is related to the surface under the curve. As this surface is important, the variable V_j is important. The principal drawback of this method is that it does not take into account the distribution of the examples to define the interval of integration.

We propose a method of variable importance measurement based on the integral of the output variations of the model using the probability distributions of the examples. This measurement was tested successfully for classification problems in [MP-10]. This method will be used in this chapter as the “variable importance” definition.

6.2.2 Variable influence

For a given problem, a subset of relevant variables can be chosen using the variable importance measurement. This variable selection increases the model robustness and facilitates the model interpretation. However, the notion of variable importance, for an instance X_n, is not sufficient to interpret its classification.

One way to complete the interpretation is to analyse the importance of the value of the considered variable V_j on the output value of the model. In Figure 6.1 the example X_n belongs to the class -1. What indicates the value of the variable V_j for this example ? Is it possible to change its class by modifying the V_j value ? We propose to answer questions such as these ones using a measurement of the value of a given variable V_j for an example X_n. The importance of the value of a variable will be called its "influence".

To produce an interpretation of the model Féraud et al. [Ext-87] propose to segment examples and then characterize each cluster using the variables importance and influences inside every cluster. In this chapter the objective is to propose a method which produces, automatically (without human assistance), an interpretation of the score for each example (instead for each cluster).

Therefore an "influence measurement" relative to every example will be proposed in the next section. Among existing methods the method proposed in [Ext-96] by Framling is the closest. But Framling uses extremums and an assumption of monotonous variations of the output model versus the variations of the input variable. The proposed "influence" measure is based on the distribution of the examples and is therefore more robust to outliers.

6.3 Method description

6.3.1 Importance of an input variable for an example

Considering[1] the model F, the example X_n, the input variable V_j and the variable to be explained p, the sensitivity of the model $S(V_j/F, X_n, p)$ is defined as the sum of the variations observed on the output p when perturbing the example X_n using the probability distribution of the input variable V_j.

The perturbed output of the model F, for an example X_n is the model output for this example but having replaced the value of the variable V_j with the value for an example k. The measured variation, for the example X_n, is then the difference between the "true output" $F_j(X_n)$ and the "perturbed output" $F_j(X_n, X_k)$ of the model.

The sensitivity of the model is then the mean value of $||F_j(X_n) - F_j(X_n, X_k)||^2$ for the probability distribution of the variable V_j. Approximating the variable probability distribution by the empirical distribution of the examples :

$$S(V_j|F, X_n, p) = \sum_{k=1}^{K} ||F_j(X_n) - F_j(X_n; X_k)||^2 \tag{6.1}$$

A sensitivity distribution is available by carrying out this sensitivity measurement on the output p and whatever is the input variable[2] V_j. The importance of the variable V_j to the example X_n, $I(V_j|F, X_n, p)$, is then defined as the rank o of the model sensitivity, $S(V_j|F, X_n, p)$, in the sensitivity distribution $S(V_j|F, X_i, p)$ $\forall i, j$:

$$I(V_j|F, X_n, p) = \tag{6.2}$$
$$P[(S(V_j|F, X_i, p) \forall i, \forall j) \leq S(V_j|F, X_n, p)] \geq o$$

[1] Definitions I and I_v are presented here for one variable V_j, of the input vector of the model, and one output p, of the output vector. These definitions are the same whatever the considered variables j and p.

[2] The importance is not intrinsic to one input variable but to all variables. The distribution is established for all the input variables and using all the examples

This measurement provides the variable importance of an input variable to an example relatively to all others examples and all others input variables. This relative measurement gives relevant information to every instance.

6.3.2 Influence on an example of an input variable value

An input variable can "pull up" (high value) or "pull down" (low value) the model output. For the example X_n the "natural" value of the output model p is by definition $F(X_n)$ (which can also be denoted by $F_j(X_n, X_n)$). The perturbed value considering the input variable V_j is $F_j(X_n, X_k)$.

The distribution of $F_j(X_n, X_k)$ represents the "potential" values for the example X_n if its variable V_j was different. The position of the natural value of X_n ($F(X_n)$) within this distribution gives information on the value of V_j (X_{nj}). The influence of the variable V_j on an example X_n is then defined, $I_v(V_j|F, X_n, p)$, as the rank r of the "natural" output model within the "potential values" :

$$I_v(V_j|F, X_n, p) = P[(F_j(X_n, X_k)\forall k) \leq F(X_n)] \geq r. \tag{6.3}$$

For example, for a two classes classification problem (output -1 or $+1$), a high value of the rank of I_v shows a positive influence on the class $+1$ and a negative one on the class -1. Reciprocally a low value of the rank of I_v shows a positive influence on the class -1 and negative one on the class $+1$.

6.3.3 Automation of the interpretation : discussion

In business applications of CRM, scores identify customers most interested to react positively to a marketing campaign. For example, rather than to send a mail to all its customers to offer a product, a company will prefer to target the subset of its customers having the most "appetency" for the product. The marketing campaign will be less expensive, and the customers who are not interested by the product will have a lower probability to receive the publicity's product in their post-box (or mailbox).

The score interpretation brings additional information to improve the effectiveness of marketing campaigns. The score understanding provides means to support and personalize commercial action. For example if a customer is identified as fragile because he wishes to renew his mobile phone, the telecommunication company will be able to react by proposing a subscription with a reduction on the purchase price of a mobile phone. If the fragility of another customer corresponds to an under use of its "pay monthly plan", the company will be able to propose a better adapted plan.

In our system (see Figure 6.2), scores and score interpretations are evaluated in the deployment phase. Customer identifiers having the highest scores and the corresponding interpretation are send to the CRM system. This system uses the score understanding to personalize customer relationships.

The proposed method in this chapter analyses the sensitivity of the model output p considering each input variable independently.

The different steps needed to obtain the score understanding can require a long computation time. To speed up this computation two solutions are possible. The first solution extracts "an abstract" of each input variable using for example the method presented in [Ext-97] or centile information for continuous value and the method presented in [Ext-98] for categorical variables. The second one consists in memorising the $S(.)$ distribution.

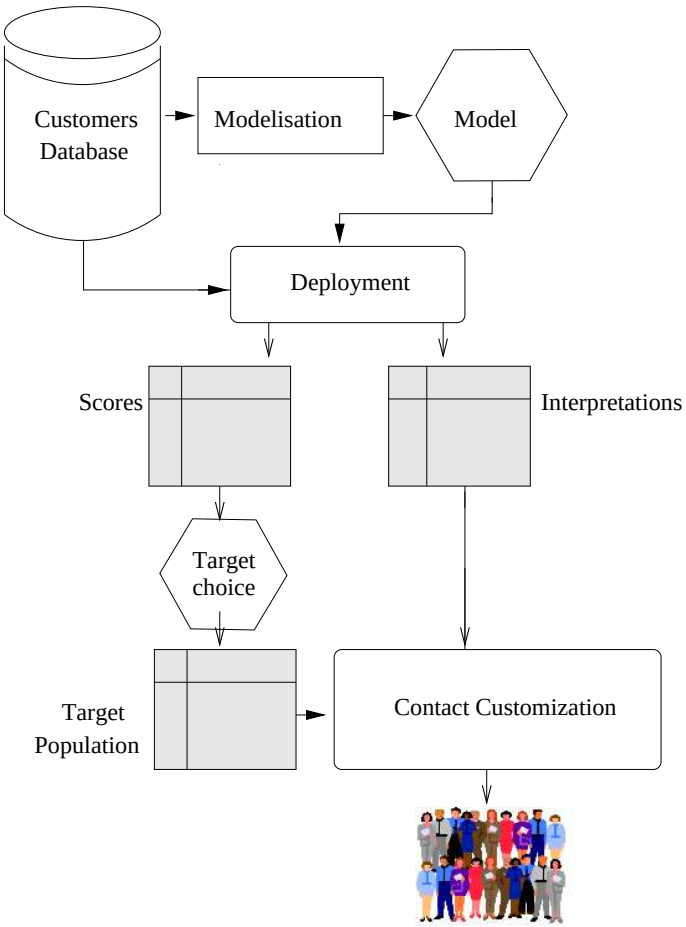

FIG. 6.2 – Application architecture

6.4 Illustration on a toy example

6.4.1 Toy example

A toy example has been constructed to test and observe the model interpretation method proposed in this chapter. This toy example is presented in Figure 6.3. In this figure the class -1 is in black and the class $+1$ is in gray. The Figure 6.4 illustrates "a priori" influence zones of the two dimensions : (1) areas of points A and C : examples where both V_1 and V_2 influence the class, (2) area of point B : examples where only V_1 influences the class, (3) areas of point D and F : examples where only V_2 influences the class and (4) area of point E : examples where any dimension influences the class.

Data : 1000 examples for the training set and 1000 for the test set, were randomly drawn ($V_1 \in [0:2], V_2 \in [0:2]$).

Models - Two types of model were tested on this toy example : (1) a Neural Network [Ext-75] (NN) using one hidden layer, a sigmoid activation function, the standard back propagation algorithm (stochastic version) and the squared error for cost function. Using a cross validation procedure the number of hidden units has been fixed to 4 ; (2) a Parzen Window [Ext-99] (PW) using an Gaussian Kernel and the L2 norm ($P(y_i|X_n) = \left(\sum_{n\ y=y_i} K(X_n, X_k) / \sum_n K(X_n, X_k)\right)$ where $K(X_n, X_k) = exp(||X_n - X_k||^2)/(2\sigma^2)$).

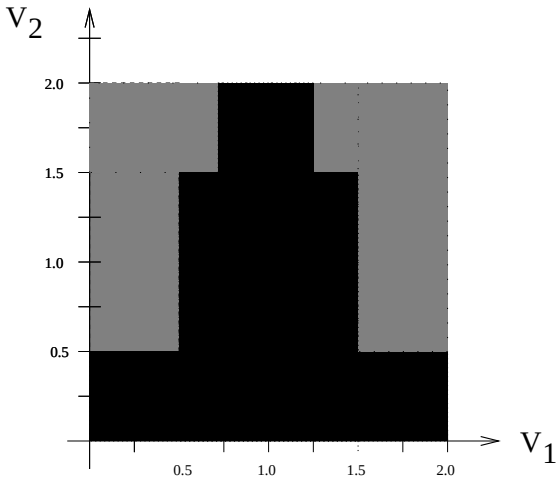

FIG. 6.3 – Toy example : two classes

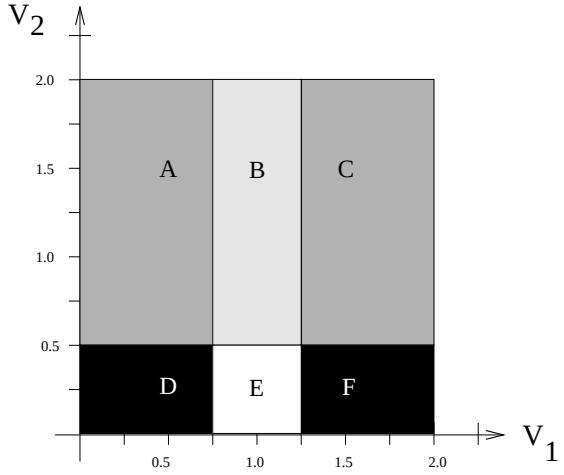

FIG. 6.4 – Influence zones

The parameter σ was fixed to 0.1 using a cross validation procedure. Whatever the model the data were standardized before training.

6.4.2 Construction of the elements of the interpretation

Among the 1000 test examples, 6 representative examples of influence zones of variables V_1 and V_2 were selected to illustrate the method. Their location is indicated in the figure 6.4 and they are named from A to F : A(0.25,1.50), B(1.00,1.50), C(1.75,1.50), D(0.25,0.25), E(1.00,0.25), F(1.75,0.25).

The interpretation as of the these 6 examples requires the following steps (for $n \in \{A, B, C, D, E, F\}$) :

- for $I(V_j/F, X_n, p)$:
  (1.1) calculation of $S(V_j/F, X_i, p)\ \forall j, \forall i$
  (1.2) sorting $S(.)$
  (1.3) determination of the rank of $S(V_j/F, X_n, p)$;

- for $I_v(V_j/F, X_n, p)$:
  (2.1) calculation of the $F(X_n, X_k)\ \forall k$;
  (2.2) sorting $F(.)$
  (2.3) determination of the rank of $F(X_n)$;

6.4.3 Results and discussion

The Figure 6.6 shows the sensibility distribution ($S(.)$, equation 6.1) obtained for V_1 using the NN and the PW on the training set. The x-coordinate represents a sensitivity value and the y-coordinate its corresponding rank in the distribution. The sensibility ranks progresses by stages (the result, not presented here, is the same for $V2$). Sensibility distributions are constituted of some important modalities relatively to the considered classification problem and the models used. These distributions concatenate the effect of individual sensibilities and influence zones : zones where the input variables have no interest, zones where they have high interest and transitory zones.

Figure 6.5 presents the distributions of "potential" output for the test point F and both the input variables V_1, V_2 using the NN. The obtained distribution using the input variable V_1 has an only one modality : $F(X_n, X_k) = -1.0\ \forall k$. This result is consistent since this variable has no influence for this example F. The

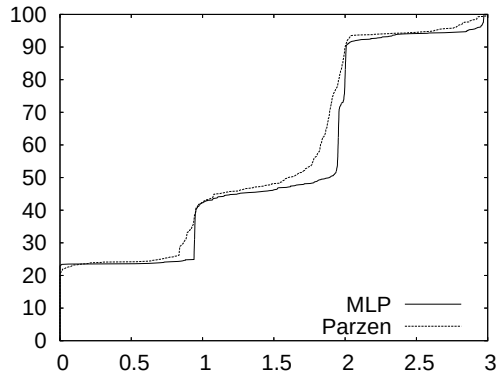

FIG. 6.5 – Ordered sensibility distribution for V_1.

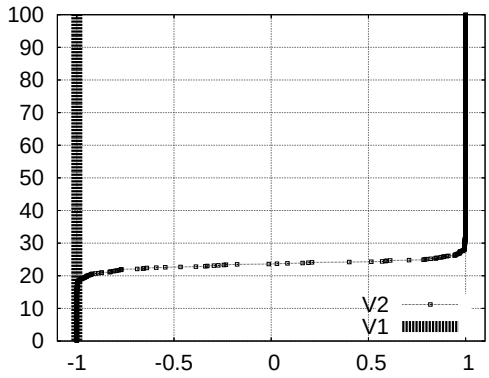

FIG. 6.6 – Ordered "potential" output for the test point 'F' and V_1,V_2 using the MLP.

obtained distribution using the input variable V_2 has 3 modes : $F(X_n, X_k) = -1$, $-1 \leq F(X_n, X_k) \leq +1$, $F(X_n, X_k) = +1$.

Figures 6.6 and 6.5 show that it could be interesting to use a rank range instead of a single rank. Quintiles, Q_1, Q_2, Q_3, Q_4 and Q_5, will be now used with the respective labels : "Very weak", "Weak", "Average", "Strong", "Very Strong". Each rank belongs to one of these quintiles (value of Q in the Table 6.2) and has therefore the corresponding label. The joint observation of Table 6.2), Figure 6.2 and Figure 6.5 shows a total coherence in the obtained results.

The influence of an input variable (I_v) has to be evaluated also in conjunction with the variable importance (I). If $I = 0$ the corresponding I_v is unimportant. Variables with a small I should not be used in the interpretation. In this case the interpretation has to be based only on the important variables (in these cases the value I_v is not presented in the Table 6.2).

6.4.4 Two examples of obtained interpretations

Two interpretations using Table 6.2 are presented here. The first interpretation is for the test point A using the Parzen Window. The interpretation contains 3 elements : (1) the point belongs to the class $+1$ with a probability (the CRM score) of 0.99 (the value of $F(X_A)$) because :

* (2) : V_1 which is very important indicates that it belongs strongly to the class $+1$

* (3) : V_2 which is moderately important indicates that it belongs strongly to the class $+1$

The second interpretation is for the test point D [3] using the MLP. The interpretation contains 2 elements : (1) the point belongs to the class -1 with a probability (the CRM score) of 1.00 (the value of $F(X_D)$) because :

* (2) : V_2 which is very important indicates that it belongs strongly to the class -1

The inspection of obtained interpretations, Table 6.2, on all points of the figure 6.3 shows that interpretations are consistent whatever the tested model ; thus is an important advantage of the proposed method. The interpretation method is also usable for other applications : the importance (I) and the influence (I_v) (of an input variable) being known, the class of an example (a customer in our application of this method) could be changed or reinforced.

[3] For the point D which belongs to the class -1, and reciprocally for the point A of the class $+1$, a low rank of I_v indicates a positive influence on the class -1 and negative one on the class $+1$, see section 6.3.2

Interpretation using the MLP				
V_j, X_n	S	I	$F(X_n)$	I_v
V_1, X_A	1.24	Q_4 (o=63)	+1.00	Q_5 (r=99)
V_2, X_A	0.96	Q_3 (o=49)	+1.00	Q_5 (r=99)
V_1, X_B	2.70	Q_5 (o=89)	-1.00	Q_1 (r=14)
V_2, X_B	0.00	-	-	-
V_1, X_C	1.24	Q_4 (o=63)	+1.00	Q_5 (r=99)
V_2, X_C	0.93	Q_2 (o=31)	+1.00	Q_5 (r=99)
V_1, X_D	0.00	-		-
V_2, X_D	3.03	Q_5 (o=95)	-1.00	Q_2 (r=22)
V_1, X_E	0.00	-	-	-
V_2, X_E	0.00	-	-	-
V_1, X_F	0.00	-	-	-
V_2, X_F	3.05	Q_5 (o=98)	-1.00	Q_2 (r=21)
Interpretation using the Parzen window				
V_j, X_n	S	I	$F(X_n)$	I_v
V_1, X_A	1.16	Q_4 (o=63)	+0.99	Q_4 (r=74)
V_2, X_A	0.97	Q_3 (o=53)	+0.99	Q_4 (r=74)
V_1, X_B	2.28	Q_5 (o=89)	-0.99	Q_2 (r=25)
V_2, X_B	0.00	-	-	-
V_1, X_C	1.16	Q_4 (o=63)	+0.99	Q_4 (r=75)
V_2, X_C	0.90	Q_2 (o=35)	+0.99	Q_4 (r=67)
V_1, X_D	0.00	-	-	-
V_2, X_D	2.96	Q_5 (o=96)	-0.99	Q_1 (r=12)
V_1, X_E	0.00	-	-	-
V_2, X_E	0.00	-	-	-
V_1, X_F	0.00	-	-	-
V_2, X_F	3.02	Q_5 (o=90)	-0.99	Q_1 (r=12)

TAB. 6.2 – Interpretation of the 6 test points

6.5 Transposition to a real application

6.5.1 Introduction to the “Why” and “How” notions

The aim of the transposition detailed in this section is a proof of concept, intended for a Orange™ Business Unit, of the interpretation method presented in this chapter. The purpose is to show that the interpretation method can be used in the context of CRM.

The way to improve customer’s relationship is described in the following example. A campaign is designed to reduce customers’ churn. The score (probability that a customer, X_n, churns) interpretation has to explain (i) “Why” the trained model indicates that the customer has this score and (ii) “How” it is possible to decrease this score.

The “Why” and “How” information are not useful for all customers. Marketers need this information only for customers on which the campaign will be applied. These customers are selected using their churn probability (high scores). These customers are named “the target”.

Using the “Why” and “How” information, marketers will write a more personalized script to retain customers. The commercial script can be personalized for each customer relationship. In the discussion between the teleoperator and the customer is rarely possible to influence more than one aspect of this customer (one input variable of the classification model which produces scores). Therefore an only one variable will be kept in the Why and How interpretations as described in the next section.

6.5.2 Implementation

The Why notion uses the definition of I presented in section 6.3.1. This definition is used, here, only for the most important variable. This variable describes a "profile" on the customer X_n and we define a Why notion by :

$$Why(X_n|F,p) = \underset{V_j}{\operatorname{argmax}} \left[I(V_j|F,X_n,p)\right] \tag{6.4}$$

The computation time of $Why(X_n)$ is in $O(KJ)$. This computation can be simplified only if the V_{dj}, the number d of different values of the variable V_j are considered. In this case the computation time of $Why(X_n)$ is in $O\left(\left(\sum_{j=1}^{j} V_{dj}\right) J\right)$. Computation time can exceed a day (since more than one million of customers are concerned) and become useless in the CRM-Analytics loop (see Figure 6.2). To reduce this computation time, variables which have more than 100 different values are discretized using centiles. Therefore a variable has now a maximum of T modalities ($T \leq 100, \forall j$). The why notion uses then for $S(.)$ the computation :

$$S(V_j|F,X_n,p) = \sum_{t=1}^{T} ||F_j\left(X_n\right)) - F_j\left(X_n; V_{tj}\right)||^2 \, P(V_{tj}) \tag{6.5}$$

where $P(V_{tj})$ is the probability of V_{tj}.

The "How" interpretation looks for values of variables that positively change the score of a customer ("pull down" value for churn or vice versa "pull up" value for "appetency"). This interpretation is tied to I_v (see equation 6.3). Here for the Orange Business Unit application, the "How" is limited to the more positive variable, such as ($F_j(.,.) \in [0:1]$) :

$$How(X_n|F,p) = \underset{V_j}{\operatorname{argmin}} \left[\underset{t}{\operatorname{argmin}} \left[F_j(X_n, V_{tj})\right]\right] \tag{6.6}$$

Here the problem is to prevent churn and to find the "worst" variable. Furthermore, variables that cannot be changed, such as sex, birthday or address, are not tested.

6.5.3 Experiments on Orange scores

Orange scores are calculated with the SAS™, Kxen™ or Khiops™ software (depending on the Business Unit and the country). Results presented here have been obtained using the Kxen software using a model close to a ridge regression. However the structure of the model is not used as detailed above in this chapter.

For confidentiality reasons results of the "why" and "how" approaches on recent Orange scores are not presented. Only the "Why" information is illustrated on an older model of churn. This model is computed on a table of 100000 customers. The target is composed of 10 % of customers.

Why	% of the Target	Usage	Product 1	Product 2	Service 1	Customer Indication	Customer Environment	Customer Behavior	...
Usage	58%	0.19	0.00	1.04	0.68	0.99	0.99	0.07	...
Product 1	17%	2.10	6.77	1.20	1.03	1.23	0.95	3.05	...
Product 2	15%	1.85	0.00	0.49	1.15	0.79	1.01	1.06	...
Product 3	6%	1.97	0.08	1.16	3.74	0.66	0.99	1.40	...
...	...	...	...	...	...	...	...	...	...

TAB. 6.3 – "Why" Results

Input variables are defined as follow : indicators of telephone use ; flags on the possession of service or product ; indicators on customer (sex, senior (yes/no), ...) ; indicators of customer environment ; indicators of customer purchasing behaviour ; ...

Table 6.3 shows on the first column the name of the most important variable using the definition equation of 6.4. The second column indicates the percentage of customers for which this variable is the most important. From the third to the last, columns gives ratios. For example the cell at the intersection of the "Usage" column and the "Product 1" line gives the ratio between the mean value of the input variable "Usage" and the mean variable of customer for which the "Product 1" input variable is the most important (in the "Why" sense). This cell indicates customers who have a mean greater than the mean population.

Table 6.3 shows a main profile, which is pointed by the "Usage" variable, that contains 58 % of the "target population". The analysis of the first line of this table indicates (1) for the first column : customers with weak usage of some services (5 times smaller than the mean population) ; (2) for the second column : customers with no services or product of type "Product 1" ; and so on. Therefore a possible marketing campaign can be build to push service usage or to suggest adequate services for their consumption. Others lines and cell of the table 6.3 can be analysed using the same process.

15 models have been tested (for this churn problem) with different numbers of input variables. All tests demonstrate that the approach is useful. The "Why" approach allows to detect profiles in high scores and to provide relevant interpretation. The "How" approach seeks the best value that will allow to reinforce (or change) a score.

6.5.4 Discussions

The Orange case shows the usefulness of the approach to detect high scores profiles. The profiles interpretation is easy since it contains only the most important variable which characterizes the profile itself.

However profile built using only the most important variable is not always the best choice. If all high scores have the same most important variable the second most sensitive variable has to be considered and so on. When the model has a lot of input variables the profile could be difficult to analyse. This is another obstacle for marketing use of the interpretation method.

6.6 Conclusion

A method to interpret results of a predictive model has been presented. Experimental results on a toy problem using two different models and experimental results using another model (from a commercial software) were performed. Results show a very nice behavior of the method. At the moment this method is being industrialized in Orange CRM applications.

Even if the method was elaborated for black box models there are still ways to improve the approaches to speed up computing of sensitivity. The sensitivity analysis of specific model (i.e. logistic regression) could be accelerated by finding an analytic sensitivity function for the model. For example the method is exact for naive bayes model which is used in the Khiops software[4]. The proposed method will be added to the Khiops software next year. Future work concerns the extension of the method to obtain an instance selection method.

Acknowledgments

Authors would like to thank Claude Riwan and the Score Team of Orange France for their contribution to the experimentation of the method presented in this chapter.

[4] http://www.francetelecom.com/en/group/rd/offer/software/applications/providers/khiops.html

Chapitre 7

A naive understanding of the naive Bayes classifier

Contents

This demonstration presents a method to interpret the output of a naive Bayes classifier. The interpretation is based on two concepts : the variable importance ("Why") and the value importance of the variable ("How"). Using the "Why" and "How" information, marketers will write a more personalized script to retain customers. The paper describes all the computation and implementation detail. The demonstration will propose to test the method in a free version of the Khiops^{TM} software.

7.1 Introduction - Context

An industrial customer analysis platform able to build prediction models with a very large number of explicative variables has been developped by Orange Labs [Ext-100]. This platform implements several processing methods for instances and variables selection, prediction and indexation based on a selective naive Bayes model combined with variable selection regularization and model averaging method. The main characteristic of this platform is its ability to scale on very large datasets with hundreds of thousands of instances and thousands of variables. The rapid and robust detection of the variables that have most contributed to the output prediction can be a key factor in a marketing application. This is the subject of this demonstration. The method we propose is derived explicitly for prediction model used in this platform : a naive Bayes classifier.

This demonstration extends the method proposed by Lemaire et al. in [MP-6] which is valid for any models and has been tested successfully on real Customer Relationship Management (CRM) problems.

This method based on sensibility analysis has a very high computational time cost for the "Why" and "how" information. Computation time can exceed a day (since more than one million of customers are concerned) and become useless in the CRM-Analytics loop. To reduce this computation time and industrialize this approach, in the platform [Ext-100], a specialization of this method for the particular case of the naive Bayes classifier is here proposed.

7.1.1 The naive Bayes classifier

The conditional probability of a class is :

$$P(C_z|X_k) = \frac{P(C_z)\prod_{j=1}^{J} P(V_j = V_{jk}|C_z)}{\sum_{t=1}^{T}\left[P(C_t)\prod_{j=1}^{J} P(V_j = V_{jk}|C_t)\right]} \tag{7.1}$$

then the predicted class is given using :

$$\underset{z}{\operatorname{argmax}}\,[P(C_z|X_k)] \tag{7.2}$$

where notations are :

X	: an input vector of J dimension ;	V_{jk}	: value of the variable j for the example X_k ;
K	: number of individuals ;	N_j	: number of different values of the variable V_j ;
X_k	: The k^{th} individual ;	C_z	: z^{th} class of the classification problem ;
V_j	: an input variable j ;	T	: number of classes ;
J	: size of the input vector	$P(C_z)$	: prior on the class z.

7.1.2 Implementation details of the naive Bayes Classifier

Only the information which is present in the model delivered by the software Khiops[1] is used. This model (the .kwc file) contains partitioning (contingency tables for all variables and all classes as described Table 7.1) and model information.

The implementation of the naive Bayes classifier in Khiops uses a m-estimate such as : $P(C_b|I_a) = \frac{N_{ab}+\epsilon}{N_{a.}+J\epsilon}$ and where N_{ab} are as describe in the Table 7.1. Therefore in the Bayes classifier $P(I_a|C_b) = \frac{N_{ab}+\epsilon}{N_{a.}+J\epsilon}\frac{N_{a.}}{N_{b.}}$ is used (where $\epsilon = \frac{1}{K+1}$ and K is the number of examples in the training set when the classifier is trained).

		C_1	C_2
input variable	I_1	N_{11}	N_{21}
defined on	I_2	N_{12}	N_{22}
three modalities	I_3	N_{13}	N_{23}

FIG. 7.1 – Contingency Table

To avoid numerical problem $P(C_x|X_k)$ is computed as : $P(C_x|X_k) = \frac{1}{\sum_{t=1}^{T} e^{L_t - L_x}}$ where $L_t = log(P(C_t)) + \sum_{j=1}^{J} log(P(V_j = V_{jk}|C_t))$ Then the perturbed output of the naive Bayes classifier could be computed using only several additions or subtractions since the difference between parts e^{L_x} in equation (7.4) and $e^{L'_x}$ in equation (7.5) ($\forall x \in T$) is : $L'_t = L_t - log(P(V_q = V_{qk}|C_t)) + log(P(V_q = V_{qn}|C_t))$. Using this equation and a pre-computation of all $P(I_a|C_b), \forall a, b$ (stored in the memory of the computer) an efficient and usable computation of a naive Bayes classifier is available even on a real time desktop of a teleoperator.

[1] `www.francetelecom.com/en/group/rd/offer/software/applications/providers/khiops.html`

7.2 Description of the Understanding Method

The way to improve customer's relationship is described in the following example. A campaign is designed to increase the appetency of a customer to a product. The score (probability that a customer, X_n, buys the product) interpretation has to explain (i) "Why" the trained model indicates that the customer has this score and (ii) "How" it is possible to increase this score. Using the "Why" and "How" information, marketers will write a more personalized script. The commercial script can be personalized for each customer relationship. In the discussion between the teleoperator and the customer, the teleoperator will try to influence aspects of this customer (input variables of the classification model which produces scores) to pull up his appetency to the product.

7.2.1 Why - Variable importance

The why concept has to explain why an instance belongs to a given class using equations 7.1 and 7.2. An example, X_k, belongs for instance to the class 1 since $P(C_1|X_k) > P(C_x|X_k), \forall x \neq 1$. Since the model is naive we propose to use a naive interpretation of the "level" of $P(C_1|X_k)$. Each input variables carries information, a contribution, to build this score : $P(V_j = V_{jk}|C_1)$. The predicted class is the one which maximize the equation 7.2.

Knowing this class, the score of the individual, X_k is given by this maximal probability using equation 7.1. We propose that the importance of a variable, j , for the probability of membership of a reference class, R, be measured by the indicator of importance :

$$P(V_j = V_{jk}|C_R) - \underset{z \neq R}{\operatorname{argmax}} [P(V_j = V_{jk}|C_z)] \tag{7.3}$$

This value, ranging between -1 and +1, measures the positive, neutral or negative contribution of the variable. Others methods [Ext-101] will be added in this software in future works. Variables are ordered according to their importance.

7.2.2 How - Value Influence

The "How" interpretation looks for values of variables that positively change the score of an example ("pull down" value for churn or vice versa "pull up" value for "appetency" in CRM applications). In the algorithm 7.2.2 the "How" interpretation looks for values that "pull up" the value of the score. For the example X_k the "natural" value of the output model is by definition $P(C_z|X_k)$. The perturbed value of the output model is $P_{new}(C_z|X_k)$. This value indicates what will be the output model "if" this example had the V_{jn} value for its V_j input variable.

7.3 Advantages : low complexity and intelligible results

The computational time cost is $O(d)$ for the "Why" information, and $O(\sum_{j=1}^{d} N_j)$ for the "How" information (where J is the size of the input vector and N_j the number of different values of a variable j).

The "score interpretation" was tested on several data files of the UCI : Adult, Iris, Mushroom and Votes. For example for the base Iris, the class to be predicted is the type of Iris among three varieties : Setosa, Virginica and Versicolor. The explanatory variables are four : the length and the width of the petals and the sepals. The interpretation of the scores for the class of reference "Setosa" indicates the classification of importance following for the first individual of the base, predicted of class "Setosa" : (1^{st}) length of the petals with an importance of 0.99 for the interval]-inf ;2.45[- (2^{nd}) width of the petals with an importance of 0.99 for the interval]-inf ;0.8[- (3^{rd}) length of the sepals with an importance of 0,78 for the interval]-inf ;5.45[- (4^{th}) width of the sepals with an importance of 0,5 for the interval [3.35 ; + inf[. For this

Pour all the different input variables V_j **de** $j = 1$ **à** $j = J$ **faire**

$$P_{old}(C_z|X_k) = \frac{\overbrace{P(C_z)\prod_{j=1}^{J} P(V_j = V_{jk}|C_z)}^{e^{L_j}}}{\sum_{t=1}^{T}\left[P(C_t)\prod_{j=1}^{J} P(V_j = V_{jk}|C_t)\right]} \tag{7.4}$$

Pour all the n different values (V_{qn}) of the variable V_q **de** $n = 1$ **à** $n = N_q$ **faire**

$P(n) = 0$; $QV(n) = 0$; $QP(n) = 0$;

$$P_{new}(C_z|X_k) = \frac{\overbrace{\left(P(C_z)\prod_{j=1,j\neq q}^{J} P(V_j = V_{jk}|C_z)\right)P(V_q = V_{qn}|C_z)}^{e^{L'_x}}}{\sum_{t=1}^{T}\left[P(C_t)\left(\prod_{j=1,j\neq q}^{J} P(V_j = V_{jk}|C_t)\right)P(V_q = V_{qn}|C_t)\right]} \tag{7.5}$$

Si $P_{new} > P_{old}$ **Alors**

$P(n) = P_{new}$; $QV(n) = V_{qn}$; $QP(n) = P(V_q = V_{qn}|C_x)$;

Fin Si

Fin Pour

Fin Pour

individual, the probability of membership of the Setosa class is estimated at 1 by the naive Bayes predictor and no improvement of the score is thus obtained.

7.4 On-Line Demonstration

Wider results will be presented at the ECML Conference. A real time demonstration will be proposed, on a credit data base. After some questions put to one visitor of the demonstration one will indicate to him if the credit he would be granted and the reason of this decision and how to increase his chance to obtain it. Visitors can come also with their data.

7.5 Who is it for ?

This method can seem straightforward but it allows extremely effective actions. It is integrated at the moment into the software Khiops and can be applied to any data base including possibly thousands of variables and hundreds of thousand lines. The tool is useful for companies and/or organizations and/or machine learning researchers who want to understand the results of a classification, to increase the knowledge they have on their problem and/or to change the result of a classification (the medical profession for example).

“Instance Selection”

Chapitre 8

Active Learning using Adaptive Curiosity

Contents

Exploratory activities seems to be crucial for our cognitive development. According to spychologists, exploration is an intrinsically rewarding behaviour. That explains the autonomous and active development of children. The developmental robotics aim to design computational systems that are endowed with such an intrinsic motivation mechanism. There are possible links between developmental robotics and classical machine learning. Active learning strategies aim to the most informative examples and adaptive curiosity allows a robot to explore its environement in an intelligente way. In this chapter, the adaptive curiosity framework is reformulated in terms of active learning terminology, and compared directly to existing algorithms in this field. The main contribution of this chapter is a new criterion evaluating the potential interestingness of zones of the sensorimotor space.

8.1 Introduction and notation

Human beings develop in an autonomous way, carrying out exploratory acivities. This phenomenon is an intrinsically motivated behaviors. Psychologists [Ext-102] have propose theory which explain exploratory behaviors as a source of self rewarding. Building such a robot is a great challenge of developmental robotics. The ambition of this field is to build a computational system that try to capture curious situations. Adaptive curiosity [Ext-103] is one possility to aim this objective. This approach push a robot towards situations in which it maximizes its learning progress. The robot first spends time in situations that are easy to learn, then shifts progressively its attention to more difficult situations, avoiding situations in which nothing can be learnt.

This chapter does a bridge between developmental robotic and classical machine learning. Active learning strategies allow a predictive model to construct its training set in interaction with an expert. The learning starts with few labelled examples. Then the model selects examples (with no label) which considers the most informative and asks their associated output to the expert. The model learns faster thanks to active learning strategies, reaching the best performances using less data. These approaches minimize the labeling cost inducted by the training of a model.

On the one hand, active learning brings into play a predictive model that explore the space of unlabelled examples, in order to find the most informative ones. On the other hand, adaptive curiosity allow a robot to explore its environment in an intelligente way, and tries to deal with the dilemma exploration / exploitation. This paper proposes to fit adaptive curiosity to supervised active learning. The organization of this paper is as follow : in section 8.2, adaptive curiosity is presented in a generic way and original implementation choices are described. The next section shows a possible implementation of adaptive curiosity for classification. The behavior of this strategy is examined on a toy example. Considering the obtained results, a new strategy of adaptive curiosity is defined in section 8.4. This new strategy is then compared with two other active learning strategies. Finally, possible improvements of adaptive curiosity are discussed.

Notations : $\mathcal{M} \in \mathbb{M}$ is the predictive model that is trained with an algorithm $\mathcal{L}$. $\mathbb{X} \subseteq \mathbb{R}^n$ represents all possible input examples of the model and $x \in \mathbb{X}$ is a particular examples. $\mathbb{Y}$ is the set of possible outputs (answers) of the model ; $y \in \mathbb{Y}$ refers to a class label which is associated to $x \in \mathbb{X}$.

The point of view of selective sampling[1] is adopted [Ext-106] in this paper. The model observes only one restricted part of the universe $\Phi \subseteq \mathbb{X}$ which is materialized by training examples with no label. (see Figure 8.1). The image of a *"bag"* containing examples for which the model can ask for associated labels is usually used to describe this approach. The set of examples for which the labels are known (at one step of the training algorithm) is called L and the set of examples for which the labels are unknown is called U with $\Phi = U \cup L$ and $U \cap L = \emptyset$.

The concept which is learnt can be seen as a function, $f : \mathbb{X} \rightarrow \mathbb{Y}$, with $f(x_1)$ the desired answer of the model for the example x_1 and $\widehat{f} : \mathbb{X} \rightarrow \mathbb{Y}$ the obtained answer of the model ; an estimation of the concept. The elements of L and the associated labels constitute a training set T. The training examples are pairs of input vectors and desired labels such that $(x, f(x))$.

[1] In practice, the choice of selective [Ext-104] or adaptive [Ext-105] sampling depends primarily on the applicability where the model is authorized, or not, "to generate" new examples.

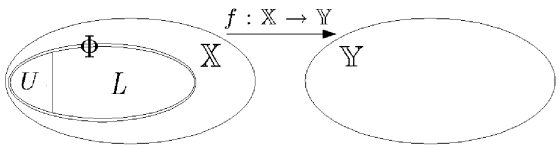

FIG. 8.1 – Handled sets

8.2 Adaptive Curiosity

8.2.1 General remarks

Adaptive curiosity [Ext-103] is the ability for a robot to choose appropriate situations[2] according to its learning[3]. Indeed, the robot can be in a trivial state (or on the contrary, in a too difficult state) in which it can not learn anything. The objective of the robot is to maximize its progress carrying out the good actions in its environment.

Y. Nagai [Ext-107] shows that a robot can learn faster considering situations where the difficulty progressively increases. The aim of adaptive curiosity is to make the robot autonomous in the choice of learnt situations. In the best case, the robot is interested by more and more difficult situations, and leaves situations for which there is nothing to learn.

The first intuition for robot's progress assessment is to compare successive performances. If the robot carries out a task in a better way than previously, one considers it makes progress. With such training rules, the robot can adopt aberrant behaviors.

To illustrate that point, Y. Nagai [Ext-107] uses the example of a robot that learns to estimate its own position after a move. The robot believes to make big progress alternating a collision with an obstacle and immobility. Indeed, "immobility" is the action that allows the robot to predict its next position with the more important precision. Comparing this performance with the previous state (the collision), the progress is maximum.

Adaptive curiosity compares similar situations (and not successive situations) [Ext-103] to measure robot's progress. Several sub-models which are specialized in certain types of situations are trained at the same time. The aim of adaptive curiosity is to make the robot autonomous in the discovery of the environment.

8.2.2 Generic Algorithm

Adaptive curiosity [Ext-103] involves a double strategy. The first strategy makes a recursive partitioning of $\mathbb{X}$, the input space of the model. The second strategy selects zones to be fed with labelled examples (and to be split by recursive partitioning). It is an active learning as far as the selection of a zone defines the subset of examples which can be labelled (those which belong to the zone). Adaptive curiosity is described below in a generic way and illustrated by an algorithm.

The input space $\mathbb{X}$ is recursively partitioned in zones (some of them are included in others). Each zone corresponds to a type of situations the robot must learn. Adaptive curiosity uses a criterion to select zones and preferentially splits area of input space $\mathbb{X}$ in which learning improves. The main idea is to schedule learnt situations in order to accelerate the robot's training.

Each zone is associated with a sub-model which is trained with examples belonging to the zone only. Sub-models are trained at the same time on disjointed examples sets. The partitioning of the input space

[2] A situation is defined as the state of the whole of sensors.

[3] The robot is learning to carry out a task in its environment.

is progressively realized, at the same time new examples are labelled. Just before the partitioning of a zone, the sub-model of the "parent" zone is duplicated in "children" zones. Duplicated sub-models continue independently their learning thanks to the examples which appear in their own zones.

Algorithm (2) shows the general steps of adaptive curiosity. It is an iterative process during which examples are selected and labelled by an expert. A first criterion chooses a zone to be fed with examples (stage A). The following stage consists in drawing an example in the selected zone (stage B). The expert gives the associated label (stage C) and the sub-model is trained with an additional example (stage D). A second criterion determines if the current zone must (or must not) be partitioned. In this case, one seeks (in the "parent" zone) adequate separations to create "children" zones (stage i). Lastly, the sub-model is duplicated into "children" zones (stage ii).

Given :

- a learning algorithm $\mathcal{L}$
- a set $\mathcal{M} = \{m_1, m_2, ..., m_n\}$ of n predictive sub-models
- $U = \{u_1, u_2, ..., u_n\}$, n subsets of unlabelled examples
- $L = \{l_1, l_2, ..., l_n\}$, n subsets of labelled examples
- $T = \{t_1, t_2, ..., t_n\}$ the training subsets corresponding to sub-models, with $t_i = \{(x, f(x))\}\ \forall x \in l_i$

$n \leftarrow 1$

Répéter

(A) Choose a sub-model m_i to be fed with examples
(B) Draw a new example x^* in u_i
(C) Label the instance x^*, $t_i \leftarrow t_i \cup (x^*, f(x^*))$
(D) Train the sub-model m_i thanks to $\mathcal{L}$, U and t_i
Si the split criterion is satisfied **Alors**
(i) Separate l_i in two sub-sets l_j and l_k the most homogeneous as possible
(ii) Duplicate m_i into two sub-models m_j and m_k
(iii) $n \leftarrow n + 1$
Fin Si

Tant $U \neq \emptyset$

Algorithme 2: Adaptive Curiosity

The main purpose of this algorithm is to seek interesting zones in the input space, at the same time the machine discovers data to learn. The algorithm chooses (as soon as possible) the examples belonging to the zones where there is possible progress. Five questions appear :

- How to decide if a zone must be partitioned ?
- How to carry out the partitioning ?
- How many "Children" zones ?
- How to choose zones to be fed in examples ?
- What kind of sub-models must be used ?

8.2.3 Original choices (Oudeyer and al, 2004)

Partitioning

A zone must be partitioned when the number of labelled examples exceeds a certain threshold. Partitioned zones are those which were preferentially chosen during previous iterations. These zones are interesting to be partitioned when more populated. Associated sub-models have done important progress.

To cut a "parent" zone into two "children" zones, all dimensions of the input space $\mathbb{X}$ are considered. For each dimension, all possible cut values are tested using the sub-model to calculate the variance of example's predictions (on both sides of the separation). During this stage, observable data Φ is used. This criterion[4] consists in finding a dimension to cut and a cut value minimizing the variance. This criterion elaborates preferentially pure zones to facilitate the learning of associated sub-models. Another constraint is added by the authors, the cut has to separate labelled examples into two subsets whose cardinalities are about balanced.

Zones selection

At every iteration, the sub-model which most improves is considered as having the strongest potential of improvement. Consequently, adaptive curiosity needs an estimation of sub-model's progress. Firstly, performances of sub-models are measured on labelled data. The choice of a performance measure is required. Secondly, sub-model's performances are evaluated on a temporal window. The sub-model which realizes the most important progress is chosen to be fed with new examples uniformly drawn.

8.3 Implementation for classification

In this section the relevance of the adaptive curiosity approach is evaluated. A toy example is used to examine the behavior of this approach within the active learning framework.

8.3.1 Transposition of original choices

Used model

A logistic regression implemented by a neural network is used [Ext-109], its architecture is represented in figure 8.2. This perceptron has two output neurons (O_1 and O_2) which are dedicated to both classes. This model consists in a single hidden neuron (H). The weights vector $[w_1..w_5]$ gathers parameters which are adjusted during the training stage. The first two network's input (x_1 and x_2) correspond to co-ordinates of the instance $x \in l$. Network's skew is an additional input whose value is 1. It makes possible to vary ordinate at the origin of the linear separating which is learnt by the model. The outputs of this model are normalized by a soft max function in the interval $[0, 1]$. Outputs correspond to probabilities of observing classes, conditionally to the instance which is placed as input of the model. Neural network's training is stopped when the training error does not decrease more than 10^{-8}, and the training step is fixed to 10^{-2}.

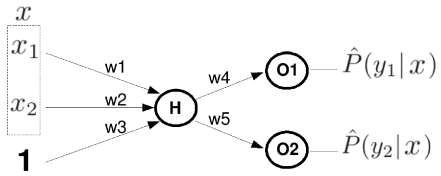

FIG. 8.2 – Neural network for logistic regression (when input vector has 2 dimensions).

Logistic regression is used as a global model (m_* on figure 8.3) which is trained independently of the input space partitioning, using examples which are selected by sub-models ($m_1, m_2...m_5$. on figure 8.3 represent sub-models which are associated with each zone). Sub-models play a role in the selection of zones

[4]This recursive partitioning playing a discretization method. For a state of the art on discretization methods, interested readers, can refer to [Ext-108].

and in the selection of instances to be labelled only. m_* is trained after using these examples. m_* allows to make a coherent comparison between adaptive curiosity and stochastic strategies. Performances of the global model report only the quality of selected examples.

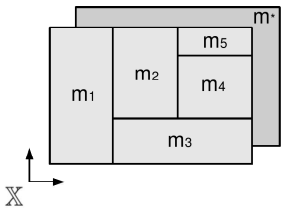

FIG. 8.3 – Local and global models

Partitioning

Zones containing at least 30 labelled examples are split. A cut separates labelled examples into two $\pm 25\%$ balanced subsets (according to the criterion of section 8.2.3). These arbitrary choices are preserved for all experiments in this paper.

Zones selection

The original criterion (section 8.2.3) which selects interesting zones in $\mathbb{X}$ is modified to transpose the adaptive curiosity to classification problems. The objective is to estimate sub-model's progresses in each zone using a measure of performance. The area under ROC curves [Ext-110] (AUC) is used to evaluate performances of sub-models on labelled examples which belong to the zone (l).

Measure of performances : ROC curves plot the rate of good predictions against the rate of bad predictions on a two dimentional space. These curves are build sorting instances of test set according to the output of the model. ROC curves are usually built considering a single class. Consequently, $|\mathbb{Y}|$ ROC curves are considered. AUC is computed for each ROC curve, and the global performance of the model is estimated by the mathematical expected value of AUC, over all classes : $AUC_{global} = \sum_{i=1}^{|\mathbb{Y}|} P(y_i).AUC(y_i)$

Measure of progress : Progresses of sub-models are estimated on a temporal window which is constituted by two successive iterations. Progresses are defined as follow, with $l \in L$ the subset of labelled examples : $\mathcal{P}rogress(l) = AUC^t_{global}(l) - AUC^{t-1}_{global}(l)$

8.3.2 Experimental conditions

Stochastic strategy

The "stochastic" strategy handles a global model and uniformly selects examples according to their probability distribution. This strategy plays a role of reference and is used to measure the contribution of adaptive curiosity.

Toy example

The toy example is a binary classification problem in a two dimensional space $\mathbb{X} = x \times y$. We consider two classes that are separated by the boundary $y = sin(x^3)$, on intervals $x \in [-2, 2]$ and $y \in [-2, 2]$ (see

figure 8.4). In the following experiments, we use 2000 training examples (Φ) and 30000 test examples that are uniformly generated over the space $\mathbb{X}$.

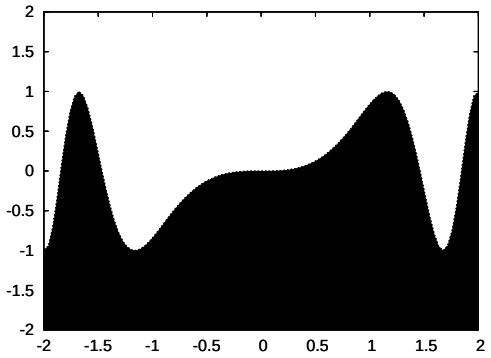

FIG. 8.4 – Toy example : $sin(x^3)$. White area corresponds to class "0" and black area corresponds to class "1".

Protocol

Beforehand, data is normalized using mean and variance. At the beginning of experiments, the training set contains only two labelled examples which are randomly chosen among available data. At every iteration, a single example is drawn in the current zone to be labelled and added to the training set. Active learning stops when 250 examples are labelled[5].

The used model is a logistic regression implemented by a neural network (section 8.3.1). Two criteria (section 8.3.1) evaluating zones, which are respectively based on the mean square error and the empirical risk, are tested during two series of experiments. Adaptive curiosity is compared to stochastic strategy (section 8.3.2) in a third serie of experiments.

These experiments evaluate the average performance of the system, according to the number of labelled examples. Each experiment has been done ten times in order to obtain an average provided with its variance, for every point of results curves.

8.3.3 Results and discussion

Performances

The criterion which is used below to evaluate strategies on the test sets is the AUC (see section 8.3.1).

In this part, performances of the global model (m_* on figure 8.3) are presented for adaptive curiosity approache. Figure 8.5 draws AUC of global model, against the number of labelled examples. Natches on curves represent variance of the 10 experiments ($\pm 2\sigma$). Perfomances of "stochastic" strategy also appears on figure 8.5. We notice that adaptive curiosity gives better performances than the stochastic strategy, nevertheless both strategies are very close. Results on figure 8.5 show this first implementation of adaptive curiosity does not improve significantly the quality of selected examples.

Selected examples

Figure 8.6 shows examples which have been selected during an experiment evaluating zones using AUC. The partitioning of input space and the choice of examples are relatively uniform ; even if a little more populated area can be noticed for each classes (at the top right and at the middle bottom of figure

[5] After 250 labelled examples, results does not vary significantly.

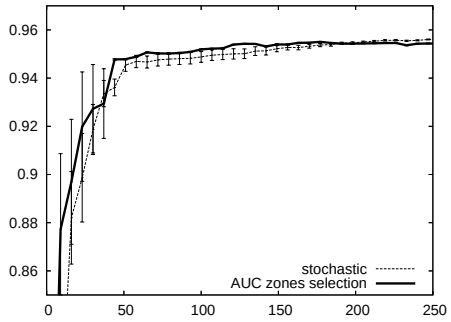

FIG. 8.5 – AUC versus number of examples

8.6). This strategy is unsatisfactory because areas which contain most labelled examples are not organized around the hidden pattern.

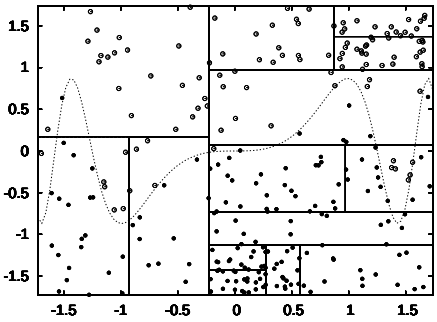

FIG. 8.6 – AUC zones selection in $\mathbb{X}$, with "$\circ$" points of first classe, and "$\bullet$" points of second classe

8.4 A new criterion of zones selection

Adaptive curiosity tries to deal with the dilemma exploration / exploitation drawing new examples in zones where progress is possible. To take in consideration this dilemma in a better way, a new criterion of zones selection is proposed in this section. The rest of the adaptive curiosity method is not modified. The new criterion is composed by two terms which respectively correspond to the exploitation and the exploration. A compromise between both terms is provided by the new criterion.

8.4.1 Exploitation : Mixture rate

Among existing splitting criteria [Ext-111], we use the entropy as a mixture rate. The function $\mathcal{M}ix\mathcal{R}ate(l)$ (equation 8.1) use labels of examples $l \subseteq L$ (which belong to the zone) to calculate the entropy over classes.

Part "A" of equation 8.1 corresponds to the entropy of classes that appear in a zone. Probabilities of classes $P(y_i)$ are empirically estimated by a counting of examples which are labelled with the considered class.

The entropy belongs to the interval $[0, \log|\mathbb{Y}|]$ (with $\mathbb{Y}$ the number of classes). Part "B" of equation 8.1 normalizes mixture rate in the interval $[0, 1]$.

$$\mathcal{M}ix\mathcal{R}ate(l) = \underbrace{-\sum_{y_i \in \mathbb{Y}} P(y_i) \log P(y_i)}_{A} \times \underbrace{\frac{1}{\log |\mathbb{Y}|}}_{B} \tag{8.1}$$

$$with\ P(y_i) = \frac{|x \in l,\ f(x) = y_i|}{|l|}$$

Mixture rate is the "exploitation" term of the proposed criterion. By choosing zones which have strongest entropy, the hidden pattern is locally clarified thanks to new labelled examples which are drawn in these zones. The model becomes very precise, on certain area of the space. Figure 8.7 shows an experiment which is realized on the toy example, using entropy to select interesting zones. Selected examples are grouped around the boundary, but there is a large part of the space which is not explored.

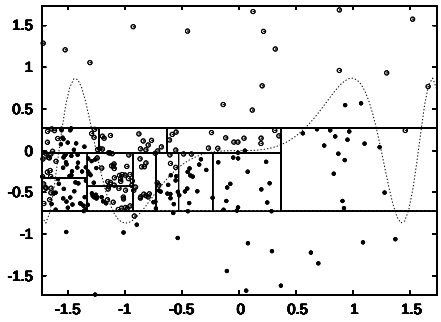

FIG. 8.7 – Selected examples using Mixture Rate only in $\mathbb{X}$, with "$\circ$" points of first classe, and "$\bullet$" points of second classe

8.4.2 Exploration : Relative density

Relative density is the proportion of labelled examples among available examples in the considered zone. Equation 8.2 expresses relative density, with $\phi \subseteq \Phi$ the subset of observable examples which belong to the zone. As mixture rate, relative density varies in the interval $[0, 1]$.

$$\mathcal{R}elative\mathcal{D}ensity(l, \phi) = \frac{|l|}{|\phi|} \tag{8.2}$$

Relative density is the "exploration" term of the criterion. The homogeneity of drawn examples over the input space is ensured by choosing zones which have lowest relative density. This strategy is different than a random sampling because homogeneity of drawn examples is forced. Figure 8.7 shows an experiment which is realized on the toy example, using relative density to select interesting zones. Input space partitioning and examples drawing are homogeneous.

8.4.3 Compromise Exploitation vs. Exploration

The criterion evaluates the interest of zones, taking into account both terms ; mixture rate and relative density. Equation 8.3 shows how each term is used. The parameter $\alpha \in [0, 1]$ corresponds to a compromise between exploitation of already known mixture zones and exploration of new zones.

$$\mathcal{I}nterest(l, \phi) = (1 - \alpha)\, \mathcal{M}ix\mathcal{R}ate(l) \tag{8.3}$$

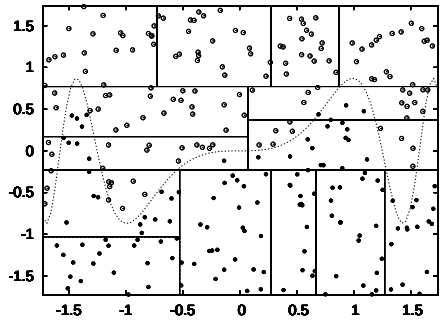

FIG. 8.8 – Selected examples using Relative Density only in $\mathbb{X}$, with "$\circ$" points of first classe, and "$\bullet$" points of second classe

$$+\alpha\,(1 - \mathcal{R}elative\mathcal{D}ensity(l, \phi))$$

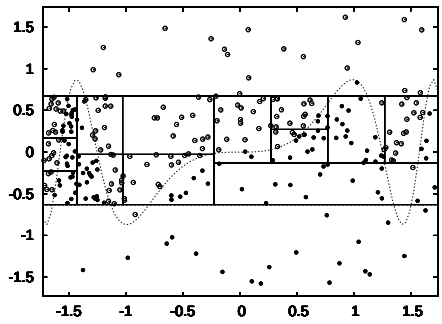

FIG. 8.9 – Selected examples with $\alpha = 0.5$ in $\mathbb{X}$, with "$\circ$" points of first classe, and "$\bullet$" points of second classe

The notion of progress is included in the criterion : the relative density (which increases at the same time new examples are labelled) forces the algorithm to leave zones in which mixture rate does not increase quickly. If there is no thing else to discover in a zone, the criterion naturally avoids it. In certain cases, the criterion prefers none mixed zones which are not enough explored. This criterion does not need a temporal window to evaluate the progress of sub-models (see paragraph 8.2.3). So its implementation is easier than original adaptive curiosity approach. Figure 8.9 shows an experiment which is realized on the toy example, using the criterion with $\alpha = \frac{1}{2}$. Input space partitioning and examples drawing are organized around the boundary without leaving any region of space.

8.4.4 Results and discussion

In this section, the toy example (section 8.3.2) as well as the experimental protocol (section 8.3.2) is re-used. Several series of experiments are realized for $\alpha = [0,\ \frac{1}{4},\ \frac{1}{2},\ \frac{3}{4},\ 1]$. The purpose of this part is to estimate the influence of this parameter on performances, and to compare the obtained results to "stochastic" strategy.

Figure 8.10 shows performances of the proposed strategy for various values of α. The first curve represents the "stochastic" strategy. When $\alpha = 0$ only mixture rate is considered by the criterion. In this

case, the observed performances are significantly lower than the "stochastic" strategy considering less than 100 examples. This phenomenon can be intuitively interpreted by a strong exploitation of detected mixture zones, to the detriment of the remaining space. When $\alpha = 1$ only relative density is considered. In this case, adaptive curiosity gives lower performances than the "stochastic" strategy considering less than 70 examples. The best performances are observed for $\alpha = 0.25$. In this case, the maximum AUC is reached very early (with 60 labelled examples). Observed performances are superior to "stochastic" strategy for all number of learnt examples. This value obviously offers a good compromise between exploration and the exploitation.

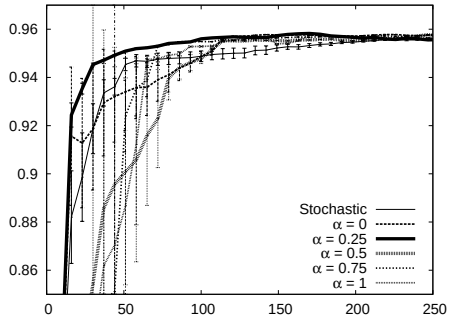

FIG. 8.10 – AUC vs number of examples

These results show that adaptive curiosity can be beneficially used in active learning framework, with the proviso of using an adapted zones selection strategy. Moreover, the new strategy of zones selection is only based on data typology. Sub-models are only used to carry out the partitioning and not to choose interesting zones.

8.5 Comparison with two active strategies

The objective of this section is to compare the previously obtained results with active learning approaches which come from the literature. Two active strategies are considered in this paper : "uncertainty sampling" and "error reduction sampling".

8.5.1 Uncertainty sampling

Uncertainty sampling is an active learning strategy [Ext-112] which is based on a confidence measure associated by the model to its prediction. The used model must be able to produce an output and to estimate the relevance of its answers. The logistic regression estimates the probability of observing each class, given an instance $x \in \mathbb{X}$. The model selects the one that maximizes $\hat{P}(y_j|x)$ (with $y_j \in \mathbb{Y}$) among all possible classes. A prediction is considered as uncertain when the probability to observe predicted class is weak. This strategy of active learning selects unlabelled examples which maximize the uncertainty of the model. The uncertainty can be expressed as follows :

$$\mathcal{I}ncertain(x) = \frac{1}{argmax_{y_j \in \mathbb{Y}} \hat{P}(y_j|x)} \qquad x \in \mathbb{X}$$

8.5.2 Sampling by risk reduction

The purpose of this approach is to reduce the generalization error, $E(\mathcal{M})$, of the model [Ext-104]. It chooses examples to be labeled so as to minimize this error. In practice this error cannot be computed

because the distribution of instances in $\mathbb{X}$ is unknown. Nicholas Roy [Ext-104] shows how to bring this strategy into play since all the elements of $\mathbb{X}$ are unknown. He uses an uniform prior for $P(x)$:

$$\widehat{E}(\mathcal{M}^t) = \frac{1}{|L|} \sum_{i=1}^{|L|} \mathcal{L}oss(\mathcal{M}^t, x_i)$$

In this paper, one estimates the generalization error ($E(\mathcal{M})$) using the empirical risk [Ext-113] :

$$\hat{E}(\mathcal{M}) = R(\mathcal{M}) = \sum_{i=1}^{|L|} \sum_{y_j \in \mathbb{Y}} \mathbb{1}_{\{f(x_i) \neq y_j\}} \, P(y_j|x_i) P(x_i)$$

where f is the model which estimates the probability that an example belong to a class, $P(y_i|x_i)$ the real probability to observe the class y_i for the example $x_i \in L$, $\mathbb{1}$ the indicating function equal to 1 if $f(x_i) \neq y_i$ and equal to 0 else. Therefore $R(\mathcal{M})$ is the sum of the probabilities that the model makes a bad decision on the training set (L). Using a uniform prior to estimate $P(x_i)$, one can write :

$$\hat{R}(\mathcal{M}) = \frac{1}{|L|} \sum_{i=1}^{|L|} \sum_{y_j \in \mathbb{Y}} \mathbb{1}_{\{f(x_i) \neq y_j\}} \, \hat{P}(y_j|x_i)$$

In order to select examples, the model is re-trained several times considering one more “fictive” example. Each instance $x \in U$ and each label $y_j \in \mathbb{Y}$ can be associated to constitute this supplementary example. The expected cost for any single example $x \in U$ which is added to the training set is then :

$$\hat{R}(\mathcal{M}^{+x}) = \sum_{y_j \in \mathbb{Y}} \hat{P}(y_j|x) \hat{R}(\mathcal{M}^{+(x,y_j)}) \quad with \; x \in U$$

8.5.3 Results on the toy example

Once again, the same toy example (section 8.3.2) and the same experimental protocol (section 8.3.2) are used. Experiments bring into play active strategies which were presented in sections 8.5.1 and 8.5.2, using a global model. As shown on figure 8.11, our adaptive curiosity strategy (with $\alpha = 0.25$) is the best active learning strategy. The uncertainty sampling gives a very high variance (for a question of legibility, natches on curve represent $\pm \frac{\sigma}{5}$ only for uncertainty sampling). Moreover, the average performance of this approach is very low in comparison to stochastic sampling. So uncertainty sampling is a very bad strategy for the considered toy example. Sampling by error reduction gives better results than the other active strategy, but the observed performances are always lower than stochastic sampling and our adaptive curiosity strategy.

8.5.4 Results on real data

Experiments are conducted using also two public data files coming from the *"UCI repository"* [Ext-114]. The datasets used are the following :

Diabetes tracking : “Pima” data file deals with detection of diabetes problems for patients who are older than 21 years. The 786 subjects (Training : 354, Test : 354) of this dataset are characterized by 9 medical indicators such as blood pressure or body mass index. The considered problem is a binary classification between individuals who have (or not) diabetes problems. Figure 8.12 shows performances of different strategies on "Pima", according to the number of labelled examples. On this dataset, sampling by risk reduction gives the best results. The AUC values are highest, for all considered number of examples. Only one curve of adaptive curiosity is shown (this curve corresponds to the best value of α). In this case, adaptive

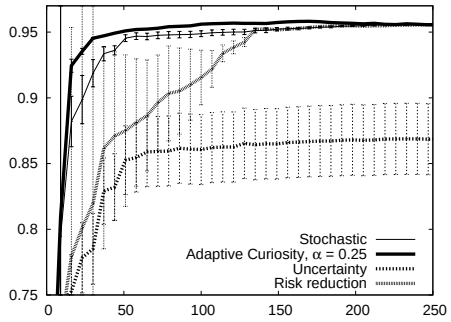

FIG. 8.11 – AUC of active learning methods

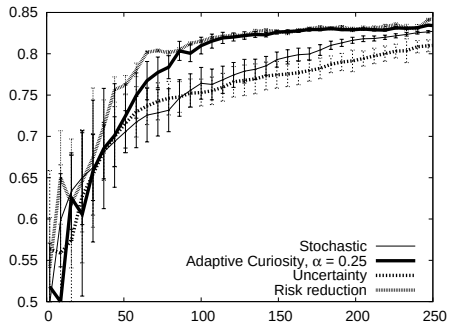

FIG. 8.12 – AUC on "Pima"

curiosity gives good performances very close to sampling by risk reduction. Moreover, adaptive curiosity gives very low variance. Finally, uncertainty sampling is the worse strategy, with AUC values which are largely lower than stochastic strategy.

Credit approval : "Australian" dataset concerns credit approvals. The 690 instances (Training : 345, Test : 345) of this dataset are defined by 14 attributes. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data. The considered problem is a binary classification on the acceptance of credits. Figure 8.13 shows performances of different strategies on "Australian". On this

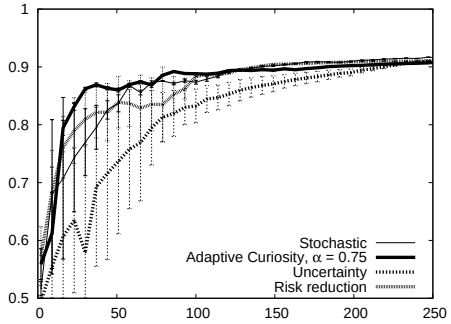

FIG. 8.13 – AUC on "Australian"

dataset, adaptive curiosity gives the best performances. The maximum AUC value (0.9) is reached with few labelled examples (about 80). When the number of labelled examples is greater than to 120, performances of "stochastic" strategy, sampling by error reduction and adaptive curiosity are very close. Once again, uncertainty sampling is the worse strategy.

Remarks : These results show that adaptive curiosity behaves similarly on the toy example and on real data. In both cases, the trend is the same : uncertainty sampling gives bad performances (worse than stochastic strategy) ; sampling by risk reduction and adaptive curiosity give close performances. However sampling by risk reduction generate a computing time 7 times higher than adaptive curiosity. Adaptive curiosity seems to be an efficient active learning strategy, with the proviso of properly adjusting the parameter α using a probabilistic estimation.

8.6 Conclusion

This paper shows that adaptive curiosity can be used as an active learning strategy in machine leaning framework. Adaptive curiosity is a strategy which is not dependent of the predictive model. This strategy can be applied on numerous real problems and is easy to use with existing systems.

We have defined a new zones selection criterion which gives good results on the considered toy example and on real data. However, this criterion balances exploitation and exploration using a parameter. Future works will be done to make the algorithm autonomous to adjust this parameter [Ext-115].

Adaptive curiosity was intially developped to deal with high dimensionality input spaces, where large parts are unlearnable or quasi-random. Future works will be realized to estimate the interest of our new criterion in such conditions. The influence of the complexity of the problem to be learnt (that is say, the number of examples necessary to solve it) will be studded to.

The partitioning step of adaptive curiosity has a $O(n^3)$ complexity and is prohibitive to treat high dimensionality datasets. Moreover, the cut criterion involves two parameters : the maximum number of labelled examples belonging to a zone, and the maximum balance rate of labelled examples subsets of a zone split. The use of non parametric discretization method [Ext-108] could be an efficient way to decide "when" and "where" a zone has to be split. This aspect will be considered in future works.

“Curriculum Vitae”

Annexe A

CV

A.1 Titres universitaires

- 1989 - BTS Électronique
- 1994 - Licence Ingénierie Électrique (Orléans)
- 1995 - Maîtrise EEA (Créteil - Paris XII)
- 1996 - DEA Robotique (Jussieu - Paris VI), Stage sur l'estimation de mouvement 3D à l'aide de vision active (IRISA - Rennes) sous la responsabilité de François Chaumette
- 1999 - Thèse de l'université de Paris VI, "Application réseaux de neurones artificiels à la gestion d'un réseau ATM" encadrant : Daniel Collobert, Directeur de thèse : Maurice Milgram

A.2 Parcours

Ma formation et mon parcours professionnel sont constitués de 4 grandes phases (voir tableau A.1) :

- 1985-1989 : une première formation et expérience professionnelle qui ont été orientées autour du métier de technicien en ingénierie électrique et/ou électronique ;
- 1990-1996 : une deuxième formation et expérience professionnelle qui ont été consacrées à devenir enseignant en deuxième cycle ;
- 1996-2002 : une troisième formation et expérience professionnelle qui ont été inspirées par le souhait de devenir ingénieur recherche ;
- depuis 2002 : enfin la suite de mon parcours a été dédiée à l'amélioration de mes compétences, à l'encadrement et à la prise de responsabilités.

Année	Formation	Profession	Thèmes de recherche (principaux)
1985-1986	Bac F2		
1987-1989	BTS Électronique	Électricien industriel	
1989-1990		Service militaire	
1990-1991		Professeur d'électronique (Dreux)	
1991-1992		Professeur d'électronique (Dreux)	
1992-1993	BTS Informatique industriel	Professeur d'électronique (Montargis)	
1993-1994	Licence Ingénierie Électrique (Orléans)	Professeur d'électronique (Montargis)	
1994-1995	Maîtrise EEA (Créteil - Paris XII)	Professeur de technologie (Orléans)	
1995-1996	DEA Robotique (Jussieu - Paris VI)	-	Stage sur l'estimation de mouvement 3D à l'aide de vision active (IRISA - Rennes) (encadrant François Chaumette)
		-	
		-	
1996-1999	Thèse de Paris VI		Application réseaux de neurones artificiels à la gestion d'un réseau ATM (encadrant Daniel Collobert) (directeur de thèse : Maurice Milgram)
1999-2001	Ingénieur R&D France Télécom		1999-2001 Data Mining Client
2002-2002	Chercheur Invité - Idiap (Institut de recherche Suisse)		2002-2002 Analyse de données
2003-2008	Ingénieur R&D - France Télécom (Orange Labs) :		
2003-2008	2003 - Nomination (dossier et oral) en tant que expert senior recherche 2003 - Responsable axe KDD (voir section A.5.5) 2006 - Renouvellement (dossier et oral) en tant que expert senior recherche 2006 - Responsable axe Apprentissage (voir section A.5.5)		2003-2004 Analyse Exploratoire 2004-2006 Sélection de Variable 2006-2008 Interprétation de modèle 2006-2008 Sélection d'instances

FIG. A.1 – Parcours

A.3 Enseignement

Avant mon doctorat

J'ai enseigné pendant 5 ans avant le début de ma thèse (10/1990 à 10/1995) en tant que professeur dans un lycée. Les deux premières années (10/1990-10/1992) j'ai enseigné l'électronique dans le lycée professionnel Edouard Branly de Dreux (28) à plein temps soit 23h de cours par semaine. Les deux années suivantes (10/1992-10/1994) j'ai enseigné l'électronique dans le lycée professionnel "Château Blanc" de Chalette sur Loing près de Montargis (45) là encore à plein temps soit 23h de cours par semaine. La cinquième année (10/1994-10/1995) j'ai enseigné la matière "technologie" dans le collège Château Blanc d'Orléans cette fois-ci à mi-temps soit 10h de cours par semaine.

En tant que professeur de lycée j'ai formé avant tout des jeunes à un métier. Mon quotidien était à la fois ancré dans une pratique professionnelle et tourné vers l'avenir. Avec ses ateliers et ses nombreux travaux pratiques, le lycée professionnel est le lieu d'apprentissage des savoir-faire. L'enseignant maîtrise les techniques liées à un métier, ou à une série de métiers, dans un secteur comme pour moi l'électronique. J'ai du savoir utiliser les installations et les outils liés à la discipline que j'enseignais et me tenir régulièrement informé des évolutions dans ce domaine.

Si transmettre des connaissances constitue la première tâche de tout enseignant, le rôle social de l'établissement tend à être un peu plus marqué en lycée technique et, surtout, en lycée professionnel que dans un établissement généraliste. Ma liberté était grande dans l'organisation de mes cours. Notamment dans une discipline comme l'électronique où il n'existait pas à l'époque de livre de cours. Le cours était donc construit en décortiquant un objet existant et en l'appuyant sur les points à faire travailler aux élèves d'après le bulletin officiel de la filière (compétences à acquérir pour l'examen). Mais je collaborais aussi avec les autres membres de l'équipe éducative.

Le professeur de lycée professionnel "devait" ces années là 23h de cours par semaine. S'y ajoutait la correction des copies et travaux pratiques (TP) ainsi que les préparatifs d'atelier, qui faisait passer ma semaine de travail à au moins trente-cinq heures.

Pendant mon doctorat

Après cinq années à enseigner à plein temps dont 3 réalisées en même temps que mes études personnelles je n'éprouvais pas le désir, ni le besoin, d'enseigner pendant mon doctorat. Néanmoins suite à la demande d'André Thépaut professeur de l'Ecole Nationale Supérieure des Télécoms (ENST) de Brest j'ai dispensé chaque année 3h de cours sur la thématique "Applications industriels des réseaux de neurones artificiels" et plus particulièrement sur des problèmes liés au télécoms.

Après mon doctorat

J'ai continué pendant 2 années à dispenser à l'ENST les 3h de cours sur la thématique "Applications industriels des réseaux de neurones artificiels" et plus particulièrement sur des problèmes liés aux télécoms.

Pendant les années 2000 à 2003 j'ai aussi dispensé 20h de cours par an sur la thématique "Techniques Statistiques Avancées" à l'école nationale de la statistique et de l'analyse de l'information (ENSAI) suite à la demande de Laurence Duval responsable de la filière informatique.

J'enseigne depuis 2004 les cours "Réseaux de neurones artificiels" et "Apprentissage Statistique une introduction" à l'ENSAI.

Je m'occupe également des séminaires d'équipes et j'ai proposé plusieurs formations en interne. Ces propositions ont débouché deux fois sur des journées de tutoriaux : 1) le premier sur les différentes formes d'apprentissage statistique (85 participants internes à France Télécom) et 2) le collège Scientifique apprentissage organisé le 29 janvier 2006 (130 participants en interne et quelques invités en externe).

La majorité des supports que j'utilise actuellement sont disponibles sur le nouveau site de l'AFIA.

1990-1992	Professeur d'électronique (Dreux) - 25h par semaine
1992-1994	Professeur d'électronique (Montargis) - 25h par semaine
1994-1995	Professeur de technologie (Orléans) - 10h par semaine
1999-2005	Intervenant extérieur à l'ENSTB, pour 'Applications au télécoms des réseaux de neurones artificiels' (3h)
2000-2003	Intervenant extérieur à l'ENSAI, pour 'Techniques statistiques pour le datamining' (20h)
2004-2009	Intervenant extérieur à l'ENSAI, pour 'Méthodes d'apprentissage' (3h) + 'Réseaux de neurones artificiels' (15h)

FIG. A.2 – Enseignement

A.4 Activités liés à l'administration

A.4.1 Le grade (interne à France Télécom) d'expert senior recherche

Chaque année, la division R&D de France Télécom organise des académies de reconnaissance d'experts. Des jurys analysent si le candidat possède la maîtrise totale de son domaine technique ou de son métier. Chaque jury a à sa disposition un dossier écrit et auditionne chaque candidat ; chaque jury est composé de membres internes à la division, internes au Groupe et extérieurs.

L'expert doit mettre à profit ses connaissances pour construire son propre parcours professionnel, pour aider à faire progresser la division et le Groupe et pour être visible et reconnu à France Télécom et auprès de ses partenaires. L'expert met en oeuvre les solutions les mieux adaptées à son profil, une dispersion importante ne serait pas une bonne démarche. Un échange et un accord avec le manager définissent les meilleures conditions de succès du rôle que jouera l'expert, pour lui-même et pour France Télécom.

Rôle et missions d'un expert

Les missions attendues d'un expert sont classées selon 4 catégories, détaillées mais non exhaustives, les items se voulant être des exemples : la 1ère catégorie comporte des items quasi-obligatoires puisque la liste reprend partiellement les critères de nomination d'un expert. Chacun pondère ensuite les autres items et peut les modifier selon sa situation et les termes de son évaluation annuelle :

- maintenir et développer son expertise :
  - posséder la maîtrise complète de son métier ou de son domaine de spécialisation technique,
  - innover et faire évoluer son domaine pour donner à France Télécom une avance compétitive durable,
  - produire des connaissances nouvelles pour le Groupe,
  - connaître l'état de l'art et alerter si besoin ;

- grâce à sa maîtrise reconnue dans son domaine, l'expert a un rôle de conseil dans le Groupe, auprès des managers notamment, et doit participer activement à la vie de la division et du Groupe :
  - se rendre mobilisable dans des projets stratégiques,
  - être capable de répondre à toute division du Groupe qui ferait appel à lui,
  - représenter France Télécom dans les organismes internationaux, face ou avec des partenaires,
  - alerter des signaux faibles, détecter les ruptures porteuses de risques ou de valeurs pour le Groupe,

  - analyser les évolutions de son domaine pour en évaluer l'impact stratégique pour France Télécom, contribuer aux orientations,
  - être acteur de la stratégie (notamment les experts émérites),
  - participer aux jurys d'experts (s'ils ne sont pas eux-mêmes candidats),
  - identifier les forces, les faiblesses et les besoins en compétences de sa filière (notamment les experts émérites),
  - contribuer à la montée en compétences à la division R&D ;

- capitaliser et transmettre ses connaissances :
  - candidater comme auditeur ou évaluateur de projets (internes ou externes) ou comme membre des comités de validation du Groupe,
  - intervenir dans des conseils scientifiques et techniques (ou équivalents),
  - valoriser ses résultats, les protéger,
  - faire évoluer les métiers,
  - diffuser les bonnes pratiques, alerter sur les mauvaises pratiques,
  - être référent ou tuteur (conseil, assistance, aide auprès de collègues candidats experts, accompagnement) selon le cas,
  - encadrer des jeunes (en particulier, mais pas uniquement),
  - avoir un rôle actif sur les formations (les construire, les dispenser) ;

- fonctionner dans des réseaux :
  - avoir un rôle actif et reconnu dans des actions de coopérations internationales, représenter France Télécom dans les instances internationales du domaine (normalisation, forums) pour y défendre les intérêts du Groupe,
  - animer des travaux et des réseaux d'experts dans son domaine et avec des partenaires,
  - créer des réunions d'échange ouvertes à d'autres métiers,
  - partager l'expertise, faire bénéficier de son expertise la division en participant aux communautés, pôles et réseaux,
  - se confronter et être reconnu par ses pairs dans et en dehors du Groupe.

Gains pour l'expert

Grâce à la reconnaissance d'expert acquise devant un jury, l'expert bénéficie d'ouverture et de situations appropriées :

- travailler sur son sujet "préféré",
- participer à des missions intéressantes, pointues sur la compétence déclarée,
- se consacrer à un travail d'approfondissement sur son domaine d'expertise,
- monter en compétence dans une filière donnée,
- favoriser son évolution au sein de son parcours professionnel, développer son employabilité,
- être reconnu et visible dans la division et le Groupe,
- disposer de temps pour poursuivre et développer son expertise,
- être ou devenir expert représente un argument fort pour l'expert dans sa gestion de carrière et son évolution auprès de son manager, notamment en termes de promotion.

La motivation personnelle est aussi une nécessité, charge au candidat de démontrer au Groupe qu'il maintient et amplifie ses compétences.

Gains pour France Telecom

La reconnaissance de l'expertise, témoin des savoirs acquis, doit être tournée vers l'avenir et transformée

au quotidien vers ses collègues, vers la division et le Groupe. France Télécom attend de ses experts qu'ils contribuent à :

- transformer les résultats de l'expert en "valeur" pour le Groupe, valoriser, communiquer,
- détecter les opportunités de brevets et les pistes de valorisation externe, notamment en normalisation et sur les domaines déclarés porteurs de valeur,
- posséder une avance technologique compétitive durable et la maîtrise des meilleures solutions,
- disposer des meilleurs professionnels (experts "métier") pour mener à bien les projets,
- acquérir des connaissances nouvelles adaptées aux besoins du Groupe,
- alerter de manière précoce sur la perception de signaux faibles,
- avoir des experts qui se consacrent à un travail d'approfondissement sur leur domaine d'expertise,
- participer aux jurys d'expert et apporter des propositions d'évolution.

A.4.2 Membre de conseils, commission de spécialistes, ...

De part mon expertise j'ai participé au groupes de travail suivants :

- 2008 : Définition et élaboration du contenu d'un white paper et position paper pour le groupe France Telecom ;
- 2008 : Membre de plusieurs groupes d'expertise (de plusieurs domaines de recherche) à France Télécom ;
- 2007 : Membre du groupe de travail "Contexte" - Rédaction d'un white paper interne ;
- 2005 : Membre du groupe de travail sur la rédaction des questions dures en recherche pour France Télécom (avant la réorganisation de la recherche dans le groupe) ;

A.5 Activités liées à la recherche

A.5.1 Prix déjà reçus pour un article ou la thèse

Des nominations parmi les meilleurs papiers de conférences mais de prix.

A.5.2 Participation à des comités, jurys, Éditorial boards, organisation de colloques, séminaires etc.

Comités Scientifiques ou Comité de direction

- Comité scientifique de la conférence "Système d'Information et Intelligence Economique (SIIE) 2008, 2009 ;
- Comité scientifique de l'atelier "Analyse de Traces" (à venir en mars 2009) ;
- Membre du bureau de l'Association Française d'Intelligence Artificielle (AFIA) ;
- Membre du comité de direction de la Coordinate Action de l'UE - Ubiquitous KD ;
- 2009-... : Responsable d'un champ de recherche du Pôle de recherche '(interne à Orange) "Network Optimisation and Decision Engineering" ;
- 2006-2008 : Responsable d'un champ de recherche du Pôle de recherche (interne à Orange) "Knowledge Science" ;
- 2003-2006 : Responsable d'un champ de recherche du Pôle de recherche (interne à Orange) "Dataledge".

Séminaires Organisés

- 2004 - 2008 : Groupe de lectures à Orange Labs Lannion ;
- Mai 2006 : Séminaire sur les différentes formes de l "Apprentissage Statistique" à Orange ;

– Janvier 2006 : Collège Scientifique “Apprentissage” à Orange.

Séminaires Invités
– 2008 : Séminaire “Apprentissage” à l’école nationale de la statistique et de l’analyse de l’information (ENSAI) ;
– 2008 : Séminaire “Analyse exploratoire de données” à l’Université d’Angers ;
– 2007 : Plenary Talk “Self Organizing Map” à “Gesellschaft für Klassifikation” (GFKl) ;
– 2006 : 50 ans de l’association Française pour l’Intelligence Artificielle (AFIA) ;
– 2006 : “Self Organizing Map” aux Journées inter-association à Lyon.

Reviews pour des conférences ou des revues
– Journal of Machine Learning Research (JMLR) : 2007, 2008 ;
– Industrial Conference on Data Mining (ICDM) : 2007, 2008 ;
– Neural Processing Letter (NPL) : 2007, 2008 ;
– Système d’Information et Intelligence Economique (SIIE) : 2008, 2009 ;
– International Joint Conference on Neural Network (IJCNN) : 2007 ;

A.5.3 Programmes d’échanges, collaborations, réseaux internationaux, projets nationaux et européens

Réseaux européens
– 2006 - 2008 : Membre du comité de direction de la Coordinate Action de l’UE - Ubiquitous KD (voir `http://www.kdubiq.org/kdubiq/control/index`).

Collaborations avec le monde universitaire :
– 2008 - 2009 : Elaboration et suivi du contrat de recherche externe entre Orange d’une part et l’INRIA et l’université de Strasbourg d’autre part : “Apprentissage sur données relationnelles.” (Responsable INRIA : Michèle Sebag, Responsable Strasbourg : Nicolas Lachiche)
– 2005 - 2008 : Elaboration et suivi du contrat de recherche externe entre Orange d’une part et l’université d’Angers (S. Loiseau) sur la thématique de l’apprentissage actif.
– 2000 - 2002 : Elaboration et suivi du contrat de recherche externe entre Orange et l’université de Laval : “Analyse de Churn à l’aide de réseaux Bayésiens”

Projets nationaux
– 2008 - 2011 : Elaboration et participation au projet ANR “Madspam” (30 mois à partir du 01/01/2008)

A.5.4 Actions de valorisation, brevets, logiciels, matériels diffusés, autres réalisations.

L’ensemble de mes activités sont de part mon appartenance à Orange des actions de valorisations de mes activités de recherche ou de ceux des membres de mon équipe, de mon laboratoire, ... Tous les méthodes que j’ai proposé ont été appliquées à un moment dans une étude ou un produit de France Télécom. Certaines de ces valorisations sont décrites ci-dessous.

Brevets -
– “Procédé et dispositif d’interprétation d’un exemple résultant de l’application d’un modèle prédictif de classification ou de régression”, V. Lemaire, R. Féraud, F. Clérot. Déposé à l’INPI le 07/09/2006

sous le numéro 06 53609

– “Procédé de modélisation de HRTF pour l’interpolation et l’individualisation des HRTF”, Rozenn Nicol, Sylvain Busson and Vincent Lemaire - Déposée à l’INPI le 10/01/2005, sous le numéro 05 00218

– “Modélisation HRTF BEM”, Rozenn Nicol, Sylvain Busson and Vincent Lemaire - Déposée à l’INPI le 27/10/2005, sous le numéro 05 10995

– “Mesure de l’importance des variables ayant servi à l’élaboration d’une modélisation”, Vincent Lemaire and Fabrice Clérot - Déposé le 27/01/2004 sous le No 04 00736

Logiciels - Netmetrics - Mon action s’est ici portée non pas sur la programmation mais sur la valorisation des méthodologies développées autour des cartes de Kohonen par moi même et d’autres membres de mon équipe. J’ai de nombreuses fois permis ou suggéré des présentations orales et écrites de ces méthodologies auprès des “décideurs” de la Recherche et du Développement de France Télécom. Cette dissémination (ou “lobying”) nous a permis de trouver des soutiens et a contribué à l’implémentation des cartes de Kohonen dans Netmetrics.

L’outil Netmetrics (outils interne à Orange destiné à l’analyse de logs) est un outil de mesure de différents indicateurs d’audience. Les cartes de Kohonen sont une méthode d’analyse exploratoire de données qui permettent de regrouper des observations similaires et de visualiser de manière très intuitive le résultat du regroupement sur un plan. Une méthodologie basée sur les cartes de Kohonen a été développée dans le cadre du pôle Knowledge Sciences et implémentée dans Netmetrics. Il s’agit d’une méthodologie à plusieurs niveaux qui exploite la nature hiérarchique des données (un client est décrit par ses sessions, une session est décrite par le nombre de pages vues sur différents thèmes) et construit de manière incrémentale de la connaissance qualitative et quantitative sur les clients. Le module carte de Kohonen de Netmetrics propose différentes fonctionnalités et interfaces de visualisation : statistiques sur les groupes obtenus, profils, carte des groupes, carte des variables de description.

“Add-on” Logiciel (en cours) - Le logiciel Khiops [1] est utilisé dans la phase de préparation des données du Data Mining. Il permet d’évaluer l’importance prédictive des variables et d’expliquer leur influence en les segmentant de façon intelligible. Il permet également de produire un modèle de scoring efficace de manière entièrement automatique.

Khiops s’appuie sur une méthode de discrétisation, qui segmente les valeurs d’une variable numérique en une série d’intervalles, et sur une méthode de groupage, qui regroupe les valeurs d’une variable symbolique en ensembles cohérents. Ces prétraitements fournissent alors une version synthétique des variables, dont l’importance prédictive est évaluée. Khiops s’appuie sur une méthode statistique permettant une évaluation optimale des variables et ne nécessite aucun paramétrage. Le logiciel Khiops est particulièrement optimisé pour les bases de données volumineuses. Il est couramment utilisé pour analyser des échantillons de données comportant des centaines de milliers d’instances et des milliers de variables.

L’Add-on à Khiops réalise l’interprétation des scores d’un prédicteur Bayesien naïf pour des problèmes de classification. L’outil est disponible en mode interface ou en mode batch et peut donc facilement être intégré à un autre projet de modélisation de données.

Ce prototype pourra être utilisé par la cellule Score Orange ainsi que par l’équipe du moteur Orange. Dans le premier cas, l’usage classique serait de rechercher les leviers d’action pour renforcer la classe de référence. Dans le second cas, il s’agirait d’exhiber les variables les plus importantes pour la probabilité qu’un site soit par exemple spam ou non.

Cet outil en est à sa phase d’évaluation par les unités d’affaire de France Télécom. Après avoir pris en compte les retours clients, ce prototype pourra ensuite être intégré dans le logiciel Khiops afin d’être diffusé

[1] voir `perso.rd.francetelecom.fr/boulle`

plus largement. Certaines évolutions sont dès à présent envisagées.

Méthodologie : - L'ensemble des études, publications scientifiques [MP-5] [MP-11] [MP-10] [MP-6] [MP-12] [MP-13] sur l'analyse de sensibilité et l'interprétation d'une modélisation conduit à la création d'une entité "Personnalisation" au sein de la cellule score d'Orange.

A.5.5 Administration liée à la recherche (coordinateur de projet, chef d'équipe, chef de laboratoire, etc.)

Coordinateur de projet

Ne sont décrites ici que les parties dont j'ai obtenu l'autorisation pour publication de la part de France Télécom

Jusqu'en **2003** la recherche à France Télécom était organisée dans des projets nommés "Programmes Vision". En 2003 elle a été réorganisée. La partie recherche a été placée dans des projets nommés "pôle de recherche" et l'anticipation a été placée dans des projets nommés "Domaines". L'un de ces pôles de recherche s'appelait Dataledge dirigé par Alain Léger.

L'activité de Recherche du Pôle scientifique **Dataledge** était au carrefour des communautés scientifiques et techniques de l'Intelligence Artificielle, des sciences cognitives et de l'informatique avancée. J'ai pris en charge l'axe KDD (Analyse des Données et Apprentissage Automatique) équivalent à une équipe de 10 personnes.

Je pense que mon apport s'est situé dans la remise à plat de cette activité. Dans le contexte difficile d'un champ scientifique assez ancien (Analyse des données statistiques : Benzécri, Saporta, Lebart, Diday, ... 1960-80) et de technologies aujourd'hui matures et donc déployées très largement, j'ai essayé de mettre à plat les forces et faiblesses de cette activité à France Télécom, pour en dégager les grands axes de recherches les plus critiques et les plus spécifiques d'un opérateur intégré . Ce travail de clarification et de diversification de nos recherches technologiques sur ce domaine au profit du groupe se voit aujourd'hui fortement sollicité pour répondre aux enjeux vitaux de l'opérateur que sont la connaissance très pertinente des usages et besoins de nos clients et donc de leur confiance renouvelée, mais aussi pour répondre à une bien meilleure efficacité de nos outils de soutien décisionnel.

Après la construction du contenu de cet axe de recherche et de certaines des roadmaps le concernant mon rôle a été d'administrer ce champ de recherche, d'y apporter une vision claire et de le faire évoluer. Le lot KDD-ML est présenté dans l'une des sections du mémoire (voir le Bulletin hors série de l'AFIA IA & Entreprises).

En **2006** les pôles de recherche et les domaines d'anticipation de France Télécom ont été réorganisés. Les pôles de recherches ont été réorganisés en Macropôles (plus gros et moins nombreux) et les domaines d'anticipation en Programme de Recherche. Le pôle de recherche Dataledge a été supprimé. Une nouvelle activité a été construite au sein du Macropôle nommé "**Knowledge Science**" et plus particulièrement de l'axe Apprentissage.

Le macropôle de recherche "Knowledge Science" comportait 4 autres axes de recherche : "Raisonnement", "Circulation", "Langues" et "Structuration". C'est cette activité, que j'ai grandement participé à construire, qui est décrite succinctement dans l'une des sections du mémoire (équivalent plein temps de 12 personnes).

Après la construction du contenu de cet axe de recherche et de certaines des roadmaps le concernant mon rôle a été d'administrer ce champ de recherche, d'y apporter une vision claire et de le faire évoluer.

L'activité concernée est décrite dans la section 1.1.1 de ce mémoire.

A.6 Encadrement

Le rôle d'un "senior research" même au sein d'une entreprise comme Orange est aussi à l'occasion d'encadrer des doctorants, de développer des cours et de développer ses projets propres. J'ai encadré de manière ponctuelle et non officielle (au sens administratif du terme) des doctorants, des post doctorants, des stagiaires. Ces encadrements ponctuels, ou aides ponctuelles, se sont parfois traduits par des publications ou brevets en communs (voir la liste des publications et des brevets).

Mes encadrements officiels sont listés ci-dessous.

- Thèse soutenue (encadrement à 100%) - (2006 :2008) - Alexis Bondu pour sa thèse en apprentissage actif - Université d'Angers
  → Publication(s) : [MP-14] [MP-15] [MP-16] [MP-8] [MP-17] [MP-18] [MP-3] [MP-19] ;
  → Brevet(s) : en cours, en lien avec les centres d'appels ;
  → Logiciel : en cours, Khiops "semi-supervisé" version $\alpha.0$ prévue fin 2008

- Thèse soutenue (collaboration importante) - (2002 :2005) - Sylvain Busson [Ext-116]
  → Publication(s) : [MP-3] [MP-20] [MP-21]
  → Brevets(s) : [MP-22] [MP-23])
  → Logiciel(s) : -

- Postdoc (encadrement à 50%) - (2007 :2008) Jonathan Milgram pour son postdoc sur l'étiquetage interactif
  → Publication(s) : en cours un article de revue et un article de conférence ;
  → Brevet(s) : -
  → Logiciels : prototype d'un logiciel d'étiquetage interactif.

- Master Recherche

  - 2007 - Maxime Chesnel (encadrement à 100%)
    → 2è année de Master "INSA" (`http://www.insa-rennes.fr/`);
    → Sujet : "Réglage d'une fenêtre de Parzen dans le cadre d'un apprentissage actif et d'un problème de classification" ;
    → Publication(s) : [MP-16].

  - 2006 - Gana Diagne (encadrement à 100%)
    → 3è année de Master "UVSQ-TRIED" (`http://www.tried.uvsq.fr/`);
    → Sujet : "Sélection de variables et méthodes d'interprétation des résultats obtenus par un modèle boite noire" ;
    → Publication(s) : - .

  - 2005 - Patrick Vovor (encadrement à 100%)
    → 3è année de Master "IRCAM-ATIAM" (`http://recherche.ircam.fr/`);
    → Sujet : "Application de techniques d'apprentissage statistiques à la prédiction d'HRTF" ;
    → Publication(s) : [MP-21].

  - 2004 - Vincent Choqueuse (encadrement à 100%)
    → 3è année de Master "UTT-OSS" (`http://www.utt.fr/admission/FiliereOSS.php?rub=02&m=01&sm=01`);
    → Sujet : "Utilisation d'outils statistiques pour l'individualisation des HRTF" ;

→ Publication(s) : [MP-20] [MP-21]
→ Brevet(s) : [MP-23].

A.7 Mes “pétales” applicatives

De par mon appartenance à un organisme de recherche industriel de larges périodes de mon emploi du temps sont consacrées à appliquer les méthodes, les méthodologies, ..., que je développe dans ma partie “coeur” de recherche décrite section précédente. J’ai choisi ici de présenter deux de ces “pétales” applicatives (d’autres valorisations sont présentées section A.5.4). Un résumé des publications liées à d’autres pétales applicatives est présenté section A.8 dans le tableau résumé.

Spatialisation sonore - La synthèse binaurale permet de reproduire des scènes sonores spatialisées en 3 dimensions à partir d’un casque d’écoute. Les applications possibles touchent des domaines tels que les télécommunications, la réalité virtuelle, le domaine militaire et le "design" sonore.

La connaissance des mécanismes mis en jeu dans la localisation auditive a permi l’élaboration des techniques de synthèse binaurale. Dans une situation d’écoute réelle : lorsqu’une onde sonore incidente se propage jusqu’au tympan, elle est diffractée par le corps de l’auditeur, en particulier par sa tête et son torse. Le son incident arrive alors aux tympans transformé, livrant ainsi au système auditif des indices caractéristiques de sa position. Pour chaque incidence, ces transformations peuvent être capturées, et implémentées sous forme de filtres audionumériques, qualifiés de Head-Related Transfer Functions (HRTF). Ces filtres constituent des empreintes spatiales qu’on appliquera au canal monophonique à spatialiser.

Ces techniques suscitent un grand intérêt du fait qu’elles se proposent de restituer aux tympans de l’auditeur le champ sonore qu’il aurait perçu dans une situation d’écoute réelle. Cependant elles présentent certaines difficultés pour une utilisation "grand public". Ces difficultés résident essentiellement dans la lourdeur des traitements audio-numériques, et la variabilité des HRTF en fonction de l’individu. En effet, la perception d’une scène sonore spatialisée avec les HRTF d’un individu se dégrade considérablement pour un individu différent de celui-ci. Une conséquence directe de cette dépendance individuelle est que les systèmes de reproduction sonore 3D (utilisant les techniques de synthèse binaurale) à usage "grand public" nécessitent un grand nombre de mesures d’HRTF (au moins 1000) à effectuer par individu. Les études développées s’inscrivaient dans un objectif d’individualisation de la synthèse binaurale.

Cette activité s’est déroulée sur 3 années avec Rozenn Nicol (permanente à Orange Labs), Sylvain Busson (en thèse à l’époque à Orange Labs), Vincent Choqueuse (stagiaire M2 à Orange Labs la première année), Patrick Vovor (stagiaire M2 à Orange Labs la deuxième année). Les résultats de cette activité ont été :

- publications scientifiques [MP-3], [MP-20], [MP-21] ;
- l’ensemble du chapitre 4 de la thèse de Sylvain Busson [Ext-116] ;
- les deux rapports de stage ;
- deux brevets : [MP-23] [MP-24] ;

Analyse clients - L’analyse prédictive étudie les données et les caractéristiques comportementales des personnes (B2C) ou des entreprises (B2B) pour en tirer des modèles prédictifs en vue d’optimiser la relation avec les clients. Elle recouvre des technologies et des pratiques comme les statistiques, l’analyse de données ou le data mining.

A partir de l’historique des informations disponibles sur les clients, l’analyse prédictive détermine, par une analyse statistique des relations entre les données disponibles, si elles sont de nature à prédire, avec la meilleure fiabilité possible, le futur comportement d’un client.

Très schématiquement : l'analyse exploratoire aura pour but de décrire le client et l'analyse prédictif détectera des prospects potentiels pour une campagne.

Certains des résultats (publiés) de cette activité ont été :

- publications scientifiques [MP-2], [MP-4], [MP-25], [MP-26], [MP-9] ;
- rapport de recherche [MP-27]
- note technique : "Analyse de la fraude sur la carte France Télécom (1999) - non disponible à l'externe ;

A.8 Publications depuis 1999

A.8.1 Tableau récapitulatif

Publication	Article ayant passé un comité de lecture	Nombre de pages	Taux acceptation (si connu)
Livres ou chapitres de livres			
[MP-5]	oui	8 pages (SC)	
[MP-2]	oui	13 pages (SC)	
Revues internationales			
[MP-28]	oui	30 pages (SC)	
[MP-4]	oui	12 pages (SC)	
[MP-29]	oui	10 pages (SC)	
Revues nationales			
[MP-14]	oui	19 pages (SC)	
[MP-30]	oui	25 pages (SC)	
Conférences internationales			
[MP-6]	oui	6 pages (DC)	40%
[MP-15]	oui	7 pages (DC)	40%
[MP-16]	oui	10 pages (SC)	
[MP-17]	oui	10 pages (SC)	
[MP-18]	oui	14 pages (SC)	
[MP-25]	oui	12 pages (SC)	28%
[MP-11]	oui	10 pages (SC)	30%
[MP-31]	oui	10 pages (SC)	19%
[MP-32]	oui	10 pages (SC)	30%
[MP-3]	oui	14 pages (DC)	
[MP-26]	oui	8 pages (DC)	
[MP-20]	oui	6 pages (DC)	40%
[MP-10]	oui	6 pages (DC)	40%
[MP-9]	oui	5 pages (DC)	
[MP-33]	oui	10 pages (SC)	
[MP-34]	oui	3 pages (DC)	#40%
Conférences nationales			
[MP-12]	oui	8 pages (SC)	40%
[MP-13]	oui	2 pages (SC)	40%
[MP-19]	oui	6 pages (SC)	40%
[MP-21]	oui	6 pages (SC)	

FIG. A.3 – Publications : SC (Simple Colonne), DC (Double Colonne)

A.8.2 Tableau récapitulatif par thème

- **Active Learning** [MP-14], [MP-15], [MP-8], [MP-17], [MP-18], [MP-19]
- **Model Interpretation** [MP-6], [MP-12], [MP-13]
- **Variable Selection** [MP-5], [MP-11], [MP-10]
- **Sound 3D** [MP-3], [MP-20], [MP-21]
- **Fraud or Behavior Analysis** [MP-2], [MP-4], [MP-25], [MP-26], [MP-9]
- **Image** [MP-28], [MP-31], [MP-32], [MP-34]
- **Cost Fonction** [MP-30], [MP-29], [MP-33]
- **Parzen Window** [MP-16]

A.8.3 Livres ou chapitres de livres avec comité de lecture

[MP-5] “Feature extraction, foundations and applications”, Chapter “An Input Variable Importance Definition based on Empirical Data Probability” (2005). V. Lemaire and F. Clérot

[MP-2] “Classification and Clustering for Knowledge Discover”, Series : “Studies in Computational Intelligence” (2005), Vol. 4, Editeurs : Halgamuge, Saman K. ; Wang, Lipo (Eds.), Approx. 300 p., Hardcover, ISBN : 3-540-26073-0, Chapter : “The many faces of Kohonen Map”. V. Lemaire and F. Clérot

A.8.4 Articles dans des revues internationales avec comité de lecture

[MP-4] “Combining several SOM approaches in data mining : application to ADSL customer behaviours analysis” (2008), F. Fessant, V. Lemaire and F. Clérot, in the Springer series “Studies in Classification, Data Analysis, and Knowledge Organization” (DAKO).

[MP-28] “Learning invariants to illumination changes typical of indoor environments : application to image color correction” (2007), B. Bascle, O. Bernier and V. Lemaire, International Journal of Imaging Systems and Technology , special issue on “Applied Color Image Processing” (IJIST)

[MP-29] “A new method to increase the margin of multilayer perceptrons” (2000), V. Lemaire, O. Bernier, F. Clérot and D. Collobert, in Neural Processing Letter (NPL).

A.8.5 Articles dans des revues nationales avec comité de lecture

[MP-14] “Etat de l’art sur les methodes statistiques d’apprentissage actif” (2008) A. Bondu, V. Lemaire, to be published in a special number of the review ’Revue des Nouvelles Technologies de l’Information (RNTI).

[MP-30] “Une nouvelle fonction de coût régularisante dans les réseaux de neurones artificiels. Application aux réseaux discriminants” (2000), V. Lemaire, in Revue d’Intelligence Artificielle (RIA).

A.8.6 Articles dans des conférences internationales avec comité de lecture

[MP-6] “Contact Personalization using a Score Understanding Method” (2008), V. Lemaire and R. Féraud and N. Voisine, in the International Joint Conference on Neural Networks (IJCNN)

[MP-15] “Adaptive Curiosity for Emotions Detection in Speech” (2008), A. Bondu and V. Lemaire, in the International Joint Conference on Neural Networks (IJCNN)

[MP-16] “Réglage de la largeur d’une fenêtre de Parzen dans le cadre d’un apprentissage actif : une évaluation” (2008), V. Lemaire, A. Bondu and M. Chesnel, in Proceeding of 1st International Conference on Information Systems and Economic Intelligence (SIIE)

[MP-8] “Active Learning using Adaptive Curiosity” (2007), A. Bondu, V. Lemaire, 7th International Conference on Epigenetic Robotics : Modeling Cognitive Development in Robotic Systems (ICER)

[MP-17] “Purchase of data labels by batches : study of the impact on the planning of two active learning strategies” (2007), V. Lemaire, A. Bondu and F. Clérot, 14th International Conference on Neural Information Processing (ICONIP)

[MP-18] “Active Learning Strategies : a case study for detection of emotions in speech” (2007), A. Bondu, V. Lemaire and B. Poulain - in “Advances in Data Mining : Theoretical Aspects and Applications” P. Perner (Ed.), Industrial Conference on Data Mining 200), LNAI 4597, Springer Verlag (ICDM)

[MP-25] “Som based data mining approach to ADSL customer behavior analysis” (2007) F. Fessant, V. Lemaire and F. Clérot, invited paper at The 31st Annual Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications (GFKL)

[MP-11] “Driven Forward Features Selection : a comparative study on Neural Networks” (2006), V. Lemaire and R. Féraud, 13th Int. Conf. on Neural Information Processsing (ICONIP)

[MP-31] “Illumination-invariant Color Correction” (2006), B. Bascle, O. Bernier and V. Lemaire, International Workshop on Intelligent Computing in Pattern Analysis/Synthesis (IWICPAS)

[MP-32] “A statistical approach for learning invariants : application to image color correction and learning invariants to illumination” (2006), B. Bascle, O. Bernier and V. Lemaire, 13th Int. Conf. on Neural Information Processsing (ICONIP)

[MP-3] “Looking for a relevant similarity criterion for HRTF clustering : a comparative study” (2006), V. Lemaire - R. Nicol, S. Busson and A. Bondu, 120th Audio Engineering Society Convention (AES120)

[MP-26] “Effective Organization and Visualization of Web Search Results” (2006), N. Bonnel, V. Lemaire, A. Cotarmanac’h and A. Morin, Internet and Multimedia Systems and Applications (Euro IMSA).

[MP-20] “Individualized HRTFs From Few Measurements : a Statistical Learning Approach” (2005), V. Lemaire, F. Clerot, S. Busson, R. Nicol and V. Choqueuse, International Joint Conference on Neural Networks (IJCNN)

[MP-10] “An Input Variable Importance Definition based on Empirical Data Probability and Its Use en Variable Selection” (2004), V. Lemaire and F. Clérot, International Joint Conference on Neural Networks (IJCNN)

[MP-9] “SOM-based clustering for on-line fraud behaviour classification : a case study” (2002), V.Lemaire and F. Clérot, Fuzzy Systems and Knowledge Discovery (FSKD).

[MP-33] “Estimation of the blocking probabilities in a ATM Network node using Artificial Neural Networks for Connection Admission” (1999), V. Lemaire and F. Clerot, in International Teletraffic Congress (ITC16).

[MP-34] “MULTRAK : a system for Automatic Multiperson Localization and tracking in real-Time” (1998), O. Bernier, M. Collobert, R. Feraud, V. Lemaire, J.-E Viallet, D. Collobert, in International Conference on Image Processing (ICIP)

A.8.7 Articles dans des conférences nationales avec comité de lecture

[MP-12] “Extraction d’information via une méthode d’interprétation de scores” (2007) V. Lemaire and R. Féraud, Workshop ’Data mining dans la banque, l’assurance et la finance’ joint to the conference (EGC).

[MP-13] “Interprétation de scores” (2007) V. Lemaire and R. Féraud, ’Extraction et Gestion des Connaissances (EGC)’

[MP-19] “Apprentissage actif d’émotions dans les dialogues Homme-Machine” (2007), A. Bondu, V. Lemaire and B. Poulain, Extraction et Gestion des Connaissances (EGC)

[MP-21] “Modélisation de HRTF par réseaux de neurones - Classification : Métrologie et traitement du signal” (2006), S. Busson, R. Nicol, V. Choqueuse, P. Vovor, V. Lemaire, F. Clérot, Congrès de la société française d’acoustique (CFA).

A.8.8 Brevets

[MP-35] “Procédé et dispositif d’interprétation d’un exemple résultant de l’application d’un modèle prédictif de classification ou de régression”, V. Lemaire, R. Féraud, F. Clérot. Déposé à l’INPI le 07/09/2006 sous le numéro 06 53609

[MP-22] “Procédé de modélisation de HRTF pour l’interpolation et l’individualisation des HRTF”, Rozenn Nicol, Sylvain Busson and Vincent Lemaire - Déposée à l’INPI le 10/01/2005, sous le numéro 05 00218

[MP-23] “Modélisation HRTF BEM”, Rozenn Nicol, Sylvain Busson and Vincent Lemaire - Déposée à l’INPI le 27/10/2005, sous le numéro 05 10995

[MP-24] “Mesure de l’importance des variables ayant servi à l’élaboration d’une modélisation”, Vincent Lemaire and Fabrice Clerot - Déposé le 27/01/2004 sous le No 04 00736

A.8.9 Rapports de recherche

[MP-36] “Bagging Using the VMSE Cost Function” (2002), V. Lemaire, IDIAP-RR 02-27.

[MP-27] “Analyse, par réseaux de neurones, des données relatives aux clients de FTM Étude de la fidélisation” (1999) Clérot, F. and Collobert, D. and Féraud, R. and Lemaire, V.

A.9 Mon projet d'HDR

A.9.1 Introduction

Mes perspectives de recherche sont en adéquation avec la suite de la feuille de route que j'ai réalisé en 2002. Une première partie est la continuité de ce que j'ai réalisé chapitres 4 et 5 de la partie scientifique de ce mémoire. Cette suite est décrite brièvement section A.9.2 ci-dessous. Une seconde partie est l'application de la connaissance que j'ai acquise en apprentissage actif mais aussi des travaux réalisé au sein de la thèse d'Alexis Bondu (encadré de 2005 à 2008), tel que décrit section A.9.3 ci-dessous. Enfin un troisième sujet dont les résultats sont à attendre à un peu plus long terme sont décrits section A.9.4.

Ces trois perspectives sont donc :
- Section A.9.2 : une perspective méthodologique (sans encadrement) ;
- Section A.9.3 : une perspective applicative (encadrement d'un post-doc) ;
- Section A.9.4 : une perspective prospective (encadrement d'un apprenti).

Je compte aussi encadrer à nouveau une thèse fin 2009[2] en apprentissage actif, après avoir bien tirer tous les enseignements des travaux en cours.

En termes d'animation de la recherche, je compte :
- continuer ma participation à la vie de l'AFIA ;
- continuer à reviewer des articles pour des conférences ou des revues ;
- continuer à participer à au moins un comité de conférence par an ;
- continuer à participer à l'animation de la recherche à Orange, suite à la dernière réorganisation de la recherche en 2008 au sein des Oranges Labs ;
- continuer à donner des cours dans la imite de mes journées de congés disponibles pour cela.

Les actions dans lesquelles je suis déjà engagé pour les années à venir seront autant de support de réalisation :

Comités Scientifiques ou Comité de direction
- Comité scientifique de la conférence "Système d'Information et Intelligence Economique (SIIE) : 2009 ;
- Comité scientifique de l'atelier "Analyse de Traces" (à paraître) ;
- Membre du bureau de l'Association Française d'Intelligence Artificielle (AFIA) ;
- 2009-... : Responsable d'un champ de recherche du Pôle de recherche '(interne à Orange) "Network Optimisation and Decision Engineering" ;

Séminaires Organisés
- 2009 : Groupe de lectures à Orange Labs Lannion ;
- Décembre 2009 : Sémininaire (interne à Orange) sur les fonctions de classement ;
- 2009 : Collège Scientifique "Recommandation" interne à Orange.

[2] après le post-doc et l'apprenti car il m'est difficile d'encadrer plus de 2 personnes

Reviews pour des conférences ou des revues
- Industrial Conference on Data Mining (ICDM) : 2009 ;
- Système d'Information et Intelligence Economique (SIIE) : 2009 ;
- International Joint Conference on Neural Network (IJCNN) : 2009 ;
- Atelier fouille de Traces : 2009

Réseaux européens
- Indéterminé à ce jour.

Collaborations avec le monde universitaire :
- 2008 - 2009 : Suivi du contrat de recherche externe entre Orange d'une part et l'INRIA et l'université de Strasbourg d'autre part : "Apprentissage sur données relationnelles." (Responsable INRIA : Michèle Sebag, Responsable Strasbourg : Nicolas Lachiche).

Projets nationaux
- 2008 - 2011 : Participation au projet ANR "Madspam"
- Eventuellement discussion autours d'un projet sur la thématique de "l'apprentissage et incertitude".

Brevets
- Indéterminé à ce jour.

Logiciels
- Grâce aux travaux d'Alexis Bondu une version semi-supervisé du logiciel Khiops verra le jour en 2008. Les années qui suivront permettront son amélioration et de franchir des étapes intéressantes vers un logiciel d'apprentissage actif performant.

"Add-on" Logiciel
- L'Add-on à Khiops qui réalisera l'interprétation des scores d'un prédicteur Bayesien naïf pour des problèmes de classification sera disponible fin 2008. Il sera employé dans le projet Madspam et agrémenté d'une visualisation. Cette visualisation sera réalisée en collaboration avec la société Kartoo (l'un des partenaires du consortium de Madspam).

A.9.2 Recherche de parangons

Lors de la phase de déploiement d'un processus de data mining, le modèle construit préalablement doit être appliqué à toute la population concernée, afin de produire un score pour chaque instance. Toutes les variables explicatives pour toutes les instances doivent être construites. Cette étape est potentiellement très coûteuse lorsque le nombre d'instances et de variables explicatives est important.

Pour faciliter le déploiement des modèles, Féraud et al. [Ext-100] ont proposé une méthode permettant d'extraire d'une base complète, une table réduite de parangons. Cette table de parangons est constituée des seules variables explicatives pertinentes au sens du score construit et des instances les plus représentatives des variables sélectionnées. La table de parangons est reliée à la base complète par un index construit automatiquement. Toute l'information produite sur la table des parangons peut être déployée par une simple jointure sur l'ensemble des instances.

Pour faciliter le déploiement d'un modèle, une table de parangons est construite. La table des parangons contient les individus représentatifs des variables explicatives utilisées par le modèle. Les parangons sont reliés par un index à toute la population. Les scores produits par l'application du modèle sur la table des parangons sont déployés sur toute la population par une simple jointure entre la table des parangons et l'index. Chaque instance de l'entrepôt de données se voit ainsi attribuer le score de son parangon. Cette méthode de déploiement est particulièrement efficace lorsque le modèle à déployer est récurent. Par exemple pour des campagnes marketing mensuelles, seule la table réduite des parangons est construite chaque mois pour produire le score de toute la population. Cette approche permet d'augmenter considérablement le nombre de scores récurrents pouvant être produits sur une même architecture technique.

La table des parangons est déterminante pour la performance finale du système. Une table de parangons peu représentative pourrait conduire à la construction de scores inefficace sur l'ensemble de la population. A contrario, une base de parangons de très grande taille diminuerait sensiblement l'intérêt de l'utilisation des parangons. Il s'agit donc de gérer au mieux le compromis entre la réduction de volumétrie et la représentativité de la base. Pour sélectionner efficacement les instances, Féraud et al utilisent un algorithme d'échantillonnage en une seule passe optimisant un critère de représentativité de l'échantillon.

En plus du critère de représentativité, la complexité algorithmique doit être prise en compte afin de rester dans des temps de calcul acceptables, les algorithmes fonctionnant en une seule passe sont les seuls candidats possibles. Les algorithmes considéré devront avoir une complexité inférieure ou de l'ordre de o(n). Alors et seulement c'est la représentativité de l'échantillon qui permettra de faire les choix entre les différents algorithmes.

Une perspective loin de l'état de l'art

L'approche d'échantillonnages actuellement codée dans la plate-forme est basée sur une méthode d'échantillonnage [Ext-117]. L'échantillonnage est la sélection d'une partie dans un tout. On citera uniquement à titre d'exemple les échantillonnages de type probabiliste qui reposent sur de l'aléatoire. Chaque unité a une chance d'être incluse dans l'échantillon. Connaissant les probabilités d'inclusion des éléments, des estimations d'erreurs d'échantillonnages peuvent être produites tout comme des inférences sur la population totale. Les méthodes probabilistes les plus simples reposent sur un tirage aléatoire des individus, avec ou sans remise où chaque individu de la base au moment du tirage a la même chance d'être tiré que ses camarades. Le cadre "sans remise" nous intéresse plus particulièrement puisque nous voulons que chaque parangon soit unique. D'autre type de tirages ne se font pas directement sur la base mais sur des strates de celle-ci. Les individus sont réunis dans des strates mutuellement exclusives et une proportion d'individu est tirée dans chaque strate. Les proportions d'individus provenant des strates peuvent soit respecter les proportions d'origine soit adopter d'autres proportions. Les strates assurent la représentativité de toutes les catégories, et permettent d'insister sur certaines strates à faibles populations qu'un tirage aléatoire pourrait sous représenter.

A.9.3 Systèmes de recommandation

Les systèmes de recommandation de contenus [Ext-118] connaissent un intérêt croissant depuis 20 ans. Le but de tels systèmes est d'aider des utilisateurs à trouver des produits intéressants (pour eux) au sein d'un catalogue (généralement très grand). Pour réaliser cela trois types d'approches sont couramment utilisées :

- le filtrage collaboratif ;
- le filtrage basé sur les contenus ;
- les approches hybrides ;

Ces trois approches utilisent des techniques plus ou moins sophistiquées d'apprentissage automatique ; bien que les "lazy learner" soient plus répandus.

Comme pour tout système d'apprentissage automatique, l'efficacité des systèmes de recommandation dépend des données disponibles. La particularité dans le cadre de la recommandation réside dans le fait que l'espace des données est naturellement creux : les utilisateurs ne consomment, ne consultent, qu'un nombre restreint de produits. Les données sont fournies par des utilisateurs réels. Il faut donc mettre en place une interaction conviviale avec l'utilisateur et sans que ce dernier passe beaucoup de temps à préciser, indiquer, ses préférences. La mise en place d'une méthode[3] efficace d'aide à la création de profil d'un nouveau utilisateur (ou mise à jour) est donc nécessaire. Cette problématique renvoie naturellement aux techniques d'apprentissage actif [MP-1].

Mes perspectives dans ce domaine sont d'encadrer un post-doctorant sur ce sujet. Il s'agira d'exploiter :

- les résultats de la thèse d'Alexis Bondu en apprentissage actif et de la connaissance que j'ai personnellement acquise sur ce sujet lors de l'encadrement d'Alexis.
- les résultats du post-doc de Laurent Candillier en "recommandation" [Ext-119; Ext-120] et de la connaissance que j'ai personnellement acquise sur ce sujet lors de sa présence dans mon équipe.

A.9.4 Apprentissage Autonome

L'apprentissage autonome[4] permet des scénarios d'autodétermination et d'autorégulation. C'est une école de pensée qui voit les apprenants comme des individus qui doivent être autonomes, c.-à-d. responsable de leur propre apprentissage, opinion, vision, caractère pratique et "liberté". Ces attributs servent à aider l'apprenant lors des ses phases d'autonomie.

Quand on parle d'apprentissage autonome on parle alors souvent d'apprentissage par renforcement et/ou comportement adaptatif. L'apprentissage par renforcement fait référence à une classe de problèmes d'apprentissage automatique, dont le but est d'apprendre, à partir d'expériences, ce qu'il convient de faire en différentes situations, de façon à optimiser une récompense numérique au cours du temps. Un paradigme classique pour présenter les problèmes d'apprentissage par renforcement consiste à considérer un agent autonome, plongé au sein d'un environnement, et qui doit prendre des décisions en fonction de son état courant. En retour, l'environnement procure à l'agent une récompense, qui peut être positive ou négative. L'agent cherche, au travers d'expériences itérées, un comportement décisionnel (appelé stratégie ou politique, et qui est une fonction associant à l'état courant l'action à exécuter) optimal, en ce sens qu'il maximise la somme des récompenses au cours du temps. L'apprentissage autonome est sujet très intéressant pour un opérateur qui a un réseau à gérer [Ext-122].

Dans mes perspectives je suis pour ce domaine dans une phase prospective. Les seules étapes réalisées sont la recherche d'un modèle simple sans paramètres [MP-16], capable de s'auto-adapter. Les politiques

[3] Ensemble des techniques nommées : "Method for cold-start recommendation" par Google ou Yahoo.

[4] Pour plus de détails on pourra lire les pages du laboratoire dédié à ce type d'études : http ://www-anw.cs.umass.edu/index.shtml The Autonomous Learning Laboratory (ALL), Department of Computer Science University of Massachusetts, Amherst (ou encore [Ext-121])

d'auto-adaption non pas encore été explorées. Sur ce sujet de recherche j'espère pouvoir encadrer en 2008-2009 un ingénieur en cursus d'apprentissage dans le but de faire d'autres avancées prospectives.

Bibliography

Mes publications citées (après 1999)

[MP-1] A. Bondu and V. Lemaire. Apprentissage actif : un état de l'art. *RNTI*, 2007.

[MP-2] V. Lemaire and F. Clérot. The many faces of kohonen map. In *Classification and Clustering for Knowledge Discover*, volume 4 of *Studies in Computational Intelligence*. Halgamuge, Saman K. and Wang, Lipo, 2005. ISBN : 3-540-26073-.

[MP-3] V. Lemaire, R. Nicol, S. Busson, and A. Bondu. Looking for a relevant similarity criterion for hrtf clustering : a comparative study. In *120th Audio Engineering Society Convention (AES120)*, 2006.

[MP-4] F. Fessant, V. Lemaire, and F. Clérot. Combining several som approaches in data mining : application to adsl customer behaviours analysis. *Studies in Classification, Data Analysis and Knowledge Organization (DAKO)*, 2008.

[MP-5] I. Guyon. *Feature extraction, foundations and applications*, chapter An Input Variable Importance Definition based on Empirical Data Probability. Elsevier, 2005.

[MP-6] V. Lemaire, R. Féraud, and N. Voisine. Contact personalization using a score understanding method. In *in International Joint Conference on Neural Networks (IJCNN)*, 2008.

[MP-7] V. Lemaire, R. Féraud, and F. Fessant. A naïve understanding of the naive bayes classifier. Poster submited to ECAI 2008.

[MP-8] A. Bondu and V. Lemaire. Active learning using adaptive curiosity. In *International Conference on Epigenetic Robotics : Modeling Cognitive Development in Robotic Systems (ICER)*, 2007.

[MP-9] V. Lemaire and F. Clérot. Som-based clustering for on-line fraud behaviour classification : a case study. In *Fuzzy Systems and Knowledge Discovery (FSKD)*, 2002.

[MP-10] V. Lemaire and F. Clérot. An input variable importance definition based on empirical data probability and its use in variable selection. In *International Joint Conference on Neural Networks IJCNN*, 2004.

[MP-11] V. Lemaire and R. Féraud. Driven forward features selection : a comparative study on neural networks. In *International Conference on Neural Information Processing*, Hong-Kong, October 2006.

[MP-12] V. Lemaire and R. Féraud. Extraction d'information via une méthode d'interprétation de scores. In *Workshop 'Data mining dans la banque, l'assurance et la finance' joint to the conference (EGC)*, 2007.

[MP-13] V. Lemaire and R. Féraud. Interprétation de scores. In *Extraction et Gestion des Connaissances (EGC)*, 2007.

[MP-14] A. Bondu and V. Lemaire. Etat de l'art sur les methodes statistiques d'apprentissage actif. *Revue des Nouvelles Technologies de l'Information (RNTI)*, 2008.

[MP-15] A. Bondu and V. Lemaire. Adaptive curiosity for emotions detection in speech. In *in International Joint Conference on Neural Networks (IJCNN)*, 2008.

[MP-16] V. Lemaire, A. Bondu, and M. Chesnel. Réglage de la largeur d'une fenêtre de parzen dans le cadre d'un apprentissage actif : une évaluation. In *International Conference on Information Systems and Economic Intelligence (SIIE)*, 2008.

[MP-17] V. Lemaire, A. Bondu, and F. Clérot. Purchase of data labels by batches : study of the impact on the planning of two active learning strategies. In *International Conference on Neural Information Processing (ICONIP)*, 2007.

[MP-18] A. Bondu, V. Lemaire, and B. Poulain. Active learning strategies : a case study for detection of emotions in speech. In P. Perner, editor, *Industrial Conference on Data Mining (ICDM)*, Advances in Data Mining : Theoretical Aspects and Application. Springer Verlag, 2007.

[MP-19] A. Bondu, V. Lemaire, and P. Poulain. Apprentissage actif d'émotions dans les dialogues homme-machine. In *Extraction et Gestion des Connaissances (EGC)*, 2007.

[MP-20] V. Lemaire, C. Clérot, S. Busson, R. Nicol, and V. Choqueuse. Individualized hrtfs from few measurements : a statistical learning approach. In *International Joint Conference on Neural Networks (IJCNN)*, 2005.

[MP-21] S. Busson, R. Nicol, V. Choqueuse, P. Vovor, V. Lemaire, and F. Clérot. Modélisation de hrtf par réseaux de neurones - classification : Métrologie et traitement du signal. In *Congrès de la société française d'acoustique (CFA)*, 2006.

[MP-22] R. Nicol, S. Busson, and V. Lemaire. Procédé de modélisation de hrtf pour l'interpolation et l'individualisation des hrtf. Brevet déposé à l'INPI le 10/01/2005, sous le numéro 05 00218.

[MP-23] R. Nicol, S. Busson, and V. Lemaire. Modélisation hrtf bem. Brevet déposée à l'INPI le 27/10/2005, sous le numéro 05 10995.

[MP-24] V. Lemaire and F. Clérot. Mesure de l'importance des variables ayant servi à l'élaboration d'une modélisation. Brevet déposé à l'INPI le 27/01/2004 sous le numéro 04 00736.

[MP-25] F. Fessant, V. Lemaire, and F. Clérot. Som based data mining approach to adsl customer behavior analysis. In *Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications (GFKL)*, 2007.

[MP-26] N. Bonnel, V. Lemaire, A. Cotarmanac'h, and A. Morin. Effective organization and visualization of web search results. In *Internet and Multimedia Systems and Applications (Euro IMSA)*, 2006.

[MP-27] F. Clérot, D. Collobert, R. Féraud, and V. Lemaire. Analyse, par réseaux de neurones, des données relatives aux clients de ftm étude de la fidélisation (churn). Technical report, CNET, 1999.

[MP-28] B. Bascle, O. Bernier, and V. Lemaire. Learning invariants to illumination changes typical of indoor environments : application to image color correction. *International Journal of Imaging Systems and Technology (IJIST), special issue on "Applied Color Image Processing*, 2007.

[MP-29] V. Lemaire, O Bernier, F. Clérot, and D. Collobert. A new method to increase the margin of multilayer perceptrons. *Neural Processing Letter*, 1999.

[MP-30] V. Lemaire. Une nouvelle fonction de coût régularisante dans les réseaux de neurones artificiels. application aux réseaux discriminants. *Revue d'Intelligence Artificielle*, 2000.

[MP-31] B. Bascle, O. Bernier, and V. Lemaire. Illumination-invariant color correction. In *International Workshop on Intelligent Computing in Pattern Analysis/Synthesis (IWICPAS)*, 2006.

[MP-32] B. Bascle, O. Bernier, and V. Lemaire. A statistical approach for learning invariants : application to image color correction and learning invariants to illumination. In *Int. Conf. on Neural Information Processsing (ICONIP)*, 2006.

[MP-33] V. Lemaire and F. Clérot. Estimation of the blocking probabilities in a atm network node using artificial neural networks for connection admission. In *International Teletraffic Congress (ITC16)*, 1999.

[MP-34] O. Bernier, M. Collobert, R. Feraud, V. Lemaire, J.-E Viallet, and D. Collobert. Multrak : a system for automatic multiperson localization and tracking in real-time. In *International Conference on Image Processing (ICIP)*, 1998.

[MP-35] V. Lemaire, R. Féraud, and F. Clérot. Procédé et dispositif d'interprétation d'un exemple résultant de l'application d'un modèle prédictif de classification ou de régression. Brevet déposé à l'INPI le 07/09/2006 sous le numéro 06 53609.

[MP-36] V. Lemaire. Bagging using the vmse cost function. Technical Report RR 02-27, IDIAP, 2002.

[MP-37] R. Féraud and V. Lemaire. Parangons selection to faster model deployment. to be submited to Intelligent Data Analysis end of 2008.

Publications externes citées

[Ext-38] B. J. Han and M. Kamber, editors. *Data Mining*. Elsevier, seconde edition, 2006.

[Ext-39] B. Liu, editor. *Web Data Mining*. Springer, 2006.

[Ext-40] M. W. Berry, editor. *Survey of Text Mining*. Springer, 2003.

[Ext-41] D. J. Cook and L. B. Holder, editors. *Mining Graph Data*. Wiley, 2007.

[Ext-42] M. Garofalakis, J. Gehrke, and R. Rastogi, editors. *Data Stream Management : Processing High-Speed Data Streams*. Springer, 07/2008.

[Ext-43] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*, chapter Unsupervised Learning and Clustering. Wiley Interscience, 2001.

[Ext-44] O. Chapelle. *Semi-supervised Learning*. MIT Press, 2006.

[Ext-45] S. Kotsiantis. Supervised machine learning : A review of classification techniques. *Informatica Journal*, 31 :249–268, 2007.

[Ext-46] R. Sutton and A. G. Barto. *Reinforcement Learning - An Introduction*. - web, 1998.

[Ext-47] Q Yang. 10 challenging problems in data mining research. Slides prepared for ICDM 2005. `www.slidefinder.net/1/10_challenging_problems/data_mining_research/2033475`.

[Ext-48] T. Kohonen. Self-organizing maps. In *Springer Series in Information Sciences*, volume 30. Springer, Berlin, Heildelberg, 1995.

[Ext-49] J. Vesanto. Som-based data visualization methods. *Inteligent Data Analysis*, 3(2) :111–126, 1999.

[Ext-50] J. Lampinen and E. Oja. Clustering properties of hierarchical self-organizing maps. *Journal of Mathematical Imaging and Vision*, 2(3) :261–272, 1992.

[Ext-51] Juha Vesanto. *Data Exploration Process Based on the Self-Organizing Map*. PhD thesis, Helsinki University of Technology, 2002.

[Ext-52] Juha Vesanto, Johan Himberg, Esa Alhoniemi, and Juha Parhankangas. SOM toolbox for Matlab 5. Report A57, Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, April 2000. `http://www.cis.hut.fi/projects/somtoolbox/`.

[Ext-53] S.M. Weiss and C.A. Kulikowski. *Computer Systems That Learn*. Morgan Kaufmann, 1991.

[Ext-54] A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2) :245–271, December 1997.

[Ext-55] P. Langley. Selection of relevant features in machine learning. In AAAI Press, editor, *AAAI Fall Symposium on Relevance*, New Orleans, 1994.

[Ext-56] C. Bishop. *Neural Network for Pattern Recognition.* Oxford University Press, 1996.

[Ext-57] J. R. Quinlan. *C4.5 : Programs for Machine Learning.* Morgan Kaufmann, 1993.

[Ext-58] DJ. Hand and K. Yu. Idiot's bayes - not so stupid after all ? *International Statistical Review*, 69(3) :385–399, 2001.

[Ext-59] F. Clérot and F. Fessant. From ip port numbers to adsl customer segmentation : knowledge aggregation and representation using kohonen maps. In *DATAMINING IV*, Rio de Janeiro, Brazil, 2003.

[Ext-60] F.L. Wightman and D.J. Kistler. Headphone simulation of free-field listening i : Stimulus synthesis. *JASA*, 98(5) :858–867, 1989.

[Ext-61] A. W. Bronkhorst. Localization of real and virtual sound sources. *JASA*, 1989.

[Ext-62] H. Moller, M. F. Sorensen, D. Hammershoi, and C. B. Jensen. Head related transfer functions of human subjects. *JAES*, 43 :300–321, 1995.

[Ext-63] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The cipic hrtf database. In *IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics*, pages 99–102, Mohonk Mountain House, New Paltz, NY, 2001.

[Ext-64] Y. Kahana. *Numerical modelling of the head related transfer function.* PhD thesis, University of Southampton, 2000.

[Ext-65] B. F. G. Katz. *Measurement and calculation of individual head related transfer functions using a boundary element model including the measurement and effect of skin and hair impedance.* PhD thesis, Pennsylvania State University, 1998.

[Ext-66] S. Shimada, N. Hayashi, and S. Hayashi. A clutering method for sound localization transfer functions. *JAES*, 42 :577–584, 1994.

[Ext-67] C. S. Fahn and Y. C. Lo. On the clustering of head-related transfer functions used for 3-d sound localization. *Journal of Information and Engineering*, 2003.

[Ext-68] W. M. Hartmann. *Signals, Sound, and Sensation.* Springer, 1998.

[Ext-69] C. Avendano, R. Duda, and V. Algazi. Modeling the contralateral hrtf. In *AES 16th International Conference on Spatial Sound Reproduction*, 1999.

[Ext-70] J. O. Smith. *Techniques for Digital Filter Design and System Identification with Application to the Violin.* PhD thesis, Elec. Engineering Dept., Stanford University (CCRMA), 1983.

[Ext-71] J. Blauert. *Spatial Hearing, the Psychophysics of human sound localization.* MIT Press, 1983.

[Ext-72] Leonard Kaufman and Peter J. Rousseeuw. Finding groups in data : An introduction to cluster analysis. In *Probability And Mathematical Statistics*, Wiley Series. John Wiley and Sons, 1989.

[Ext-73] J.O Smith. *Techniques for digital filtering design and system identification with the violin.* PhD thesis, CCRMA, Stanford, 1983.

[Ext-74] V. Choqueuse. Utilisation d'outils statistiques pour l'individualisation des hrtf. Master's thesis, UTC, 2004.

[Ext-75] J. A. Anderson. *An introduction to neural network.* MIT Press, 1995.

[Ext-76] H. Clement, D. Lautard, and M. Ribeyron. Adsl traffic : a forecasting model and the present reality in france. In *WTC (World Telecommunications Congress)*, 2002.

[Ext-77] J. Francois. Otarie : observation du traffic d'accès des réseaux ip en exploitation. Technical report, Orange Labs, 2002.

[Ext-78] J. Vesanto and E. Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3) :586–600, 2000.

[Ext-79] E. Oja and S. Kaski. *Kohonen maps.* Elsevier, 1999.

[Ext-80] F. Clérot and F. Fessant. From ip port numbers to adsl customer segmentation : knowledge aggregation and representation using kohonen maps. In *DATAMINING IV*, Rio de Janeiro, Brazil, 2003.

[Ext-81] JMLR. Special issue on variable and feature selection. *Journal of Machine Learning Research*, 2003.

[Ext-82] I. Guyon and A. Elisseef. An introduction to variable and feature selection. *JMLR*, 3(Mar) :1157–1182, 2003.

[Ext-83] R. Kohavi and G. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2), 1997.

[Ext-84] A. N. Réfénes, A. Zapranis, and J. Utans. Stock performance using neural networks : A comparative study with regression models. *Neural Network*, 7 :375–388, 1994.

[Ext-85] J. Moody. *Prediction Risk and Architecture Selection for Neural Networks*. From Statistics to Neural Networks-Theory and Pattern Recognition. Springer-Verlag, 1994.

[Ext-86] W. G. Baxt and H. White. Bootstrapping confidence intervals for clinical inputs variable effects in a network trained to identify the presence of acute myocardial infraction. *Neural Computation*, 7 :624–638, 1995.

[Ext-87] R. Féraud and F. Clérot. A methodology to explain neural network classification. *Neural Networks*, 15 :237–246, 2002.

[Ext-88] L. Breiman. Random forest. *Machine Learning*, 45, 2001. `stat-www.berkeley.edu/users/breiman/Breiman`.

[Ext-89] Greg Welch and Gary Bishop. SCAAT : Incremental tracking with incomplete information. In *SIGGRAPH*, Los Angeles, August 12-17 2001.

[Ext-90] A. N. Burkitt. Refined pruning techniques for feed-forward neural networks. *Complex Systems*, 6 :479–494, 1992.

[Ext-91] Y. Arcadius, J. Akossou, and R. Palm. Consequences of variable selection on the interpretation of the results in multiple linear regression. In *Biotechnol. Agron. Soc. Environ.*, volume 9, pages 11–18, 2005.

[Ext-92] J. P. Nakache and J. Confais. *Statistique explicative appliquée*. TECHNIP, 2003.

[Ext-93] J. J. Brennan and L. M. Seiford. Linear programming and l1 regression : A geometric interpretation. *Computational Statistics & Data Analysis*, 1987.

[Ext-94] S. Thrun. Extracting rules from artifcial neural networks with distributed representations. In MIT Press, editor, *Advances in Neural Information Processing Systems*, volume 7, Cambridge, MA, 1995. G. Tesauro, D. Touretzky, T. Leen.

[Ext-95] J. M. Benitez, J. L. Castro, and I Requena. Are artificial neural networks black boxes. *IEEE Transactions on Neural Networks*, 8(5) :1156–1164, 1997. Septembre.

[Ext-96] K. Främling. Explaining results of neural networks by contextual importance and utility. In *AISB*, 1996.

[Ext-97] M. Boullé. Khiops : a statistical discretization method of continuous attributes. *Machine Learning (ML)*, 55(1) :53–69, 2004.

[Ext-98] M. Boullé. A bayes optimal approach for partitioning the values of categorical attributes. *Journal of Machine Learning Research*, 2005.

[Ext-99] E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33 :1065–1076, 1962.

[Ext-100] R. Féraud, M. Boullé, F. Clérot, and F. Fessant. Vers l'exploitation de grandes masses de données. *Extaction et Gestion des Connaissances (EGC)*, 2008.

[Ext-101] M. Robnik-Sikonja and I. Kononenko. Explaining classifications for individual instances. *to appear in IEEE TKDE*, 2008.

[Ext-102] R. White. Motivation reconsidered : The concept of competence. *Psychological Review*, 66 :297–333, 1959.

[Ext-103] P-Y. Oudeyer and F. Kaplan. Intelligent adaptive curiosity : a source of self-development. In Luc Berthouze, Hideki Kozima, Christopher G. Prince, Giulio Sandini, Georgi Stojanov, G. Metta, and C. Balkenius, editors, *Proceedings of the 4th International Workshop on Epigenetic Robotics*, volume 117, pages 127–130. Lund University Cognitive Studies, 2004.

[Ext-104] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proc. 18th International Conf. on Machine Learning*, pages 441–448. Morgan Kaufmann, San Francisco, CA, 2001.

[Ext-105] A. Singh, R. Nowak, and P. Ramanathan. Active learning for adaptive mobile sensing networks. In *IPSN '06 : Proceedings of the fifth international conference on Information processing in sensor networks*, pages 60–68, New York, NY, USA, 2006. ACM Press.

[Ext-106] R. Castro, R. Willett, and R. Nowak. Faster rate in regression via active learning. In *NIPS (Neural Information Processing Systems)*, Vancouver, 2005.

[Ext-107] Y. Nagai, M. Asada, and K. Hosoda. Developmental learning model for joint attention. In *Proceedings of the 15th International Conference on Intelligent Robots and Systems (IROS)*, pages 932–937, 2002.

[Ext-108] M. Boullé. MODL : A bayes optimal discretization method for continuous attributes. *Machine Learning*, 65(1) :131–165, 2006.

[Ext-109] W. S. Sarle. Neural networks and statistical models. In *Proceedings of the Nineteenth Annual SAS Users Group International Conference, April, 1994*, pages 1538–1550, Cary, NC, 1994. SAS Institute.

[Ext-110] T. Fawcett. Roc graphs : Notes and practical considerations for data mining researchers. T. Fawcett. ROC Graphs : Notes and Practical Considerations for Data Mining Researchers. Technical Report HPL-2003-4, HP Labs, 2003., 2003.

[Ext-111] L. Breiman. Technical note : Some properties of splitting criteria. *Machine Learning*, 24(1) :41–47, July 1996.

[Ext-112] S. B. Thrun and K. Möller. Active exploration in dynamic environments. In John E. Moody, Steve J. Hanson, and Richard P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4, pages 531–538. Morgan Kaufmann Publishers, Inc., 1992.

[Ext-113] X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML (International Conference on Machine Learning)*, Washington, 2003.

[Ext-114] C. L. Blake D. J. Newman, S. Hettich and C. J. Merz. UCI repository of machine learning databases. http ://www.ics.uci.edu/∼mlearn/MLRepository.html, University of California, Irvine, Dept. of Information and Computer Sciences, 1998.

[Ext-115] T. Osugi, D. Kun, and S. Scott. Balancing exploration and exploitation : A new algorithm for active machine learning. In *Proceedings of the Fith IEEE International Conference on Data Mining (ICDM'05)*, 2005.

[Ext-116] S. Busson. *Individualisation d'indices acoustiques pour la synthèse binaurale*. PhD thesis, Université de la mediterranée Aix-Marseille II, 2005.

[Ext-117] R. Féraud, F. Clérot, and J. Dupont. Algorithmes d'échantillonage. Technical Report FT/RD/-TECH/05/12/242, Orange Labs, 2006.

[Ext-118] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17, 2005.

[Ext-119] L. Candillier, F. Meyer, and F. Fessant. Designing specific weighted similarity measures to improve collaborative filtering systems. In *Industrial Conference on Data Mining*, 2008.

[Ext-120] L. Candillier, F. Meyer, and M. Boullé. Comparing state-of-the-art collaborative filtering systems. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, 2007.

[Ext-121] Marie des Jardins. Pagoda : A model for autonomous learning in probabilistic domains. *AI Magazine*, 14(1) :75–76, 1993.

[Ext-122] A. Sohn, H. Kwak, and K. Chung. Autonomous learning of load and traffic patterns to improve cluster utilization. In *ARCS*, pages 224–239, 2007.

[Ext-123] J. Liangxiao, W. Dianhong, and C. Zhihua. Scaling up the accuracy of bayesian network classifiers by m-estimate. In *Advances Intelligent Computing Theories and Applications*, pages 475–484. 2007.