Systèmes d'Information pour l'Aide à la Décision : Applications en Téléphonie Mobile et en Données Radar

Cédric Archaux *,**, Fabrice Pellen *, Brigitte Hoeltzener *, Ali Khenchaf *

* Laboratoire E^3I^2, ENSIETA 2, rue François Verny 29806 Brest Cedex 9
[archauce, pellenfa, hoeltzbr, khenchal]@ensieta.fr
http://www.ensieta.fr/e3i2/

** Bouygues Telecom, 20 quai du point du jour, 92100 Boulogne Billancourt
carchaux@bouyguestelecom.fr
http://www.bouyguestelecom.fr

Résumé. L'objectif de cet article est de présenter en première partie le processus ECD et son application dans le cadre de la téléphonie mobile où les travaux concernent la modélisation de la valeur des clients pour la mise en place d'indicateurs décisionnels dédiés aux experts *marketing*. Il est à préciser ici, que l'acquisition et la préparation des données en téléphonie sont considérés comme étant résolus par les standards et normes de qualité mis en œuvre en amont de la chaîne. La seconde partie à l'instar de la première, tient compte des possibles rétroactions en acquisition et préparation des données cela, de part le caractère physique du système d'acquisition et des possibles modélisations qui peuvent en découler (ex. : modélisation du bruit capteur, modélisation en reconstruction de signaux, modélisation de l'environnement et mise en place de procédures expérimentales d'acquisition en vue de la reconnaissance automatique de cibles radar). Cette partie intègre de plus une introduction aux indicateurs décisionnels prévus pour permettre une aide à la mise au point de la chaîne de reconnaissance voire une adaptation contrôlée en particulier du processus de préparation des données dédié à la reconnaissance.

1 Introduction

Cet article présente des travaux qui ont pour objectif à terme la mise en place de systèmes d'information décisionnels construits à partir de bases de données volumineuses. Les deux cadres d'application sont la téléphonie mobile prépayée et la reconnaissance automatique de cibles radar. Si la téléphonie mobile est un des domaines d'origine du data-mining, nous montrons ici une application au domaine de la téléphonie prépayée qui n'est pas l'objet de nombreuses publications. De plus, l'introduction de techniques d'analyse de survie n'est pas rencontré fréquemment dans ce domaine, de plus nous proposons un modèle concret d'estimation de la valeur client dans ce contexte spécifique. Parallèlement, le domaine radar a été la source de nombreux travaux et l'application de nombreuses techniques, cependant en approchant ces deux domaines d'un point de vue un peu plus global et en les considérant sur des aspects tels que le volume et la complexité des données brutes et les applications telles

que la classification pour la reconnaissance de cibles ou de profils de clients nous identifions une analogie certaine entre ces deux domaines que nous-nous proposons d'éclaircir ici.

La première partie de l'article concerne la téléphonie mobile et présente dans un premier temps le contexte de l'étude. Nous présentons ensuite les modèles de comportement client que nous utilisons, et plus particulièrement le modèle de survie, nous enchaînons sur l'introduction de modèle de valeur client pour finir sur une utilisation potentielle de ce modèle à des fins de prédiction et d'estimation des effets des campagnes marketing.

La seconde partie concernant le domaine radar met l'accent sur les différents types de données radar et leurs caractéristiques physiques, pour mettre en évidence la nature physique des informations à initialiser voire à extraire puis à exploiter le plus automatiquement possible, dans le processus global de reconnaissance. Nous présentons ensuite l'application du processus ECD à des sources de données mesurées physiquement et sur lesquelles il est possible d'agir.

2 Outil d'aide à la décision en téléphonie mobile prépayée : utilisation d'un modèle de valeur client

2.1 Présentation du contexte d'étude

Avec plusieurs millions d'utilisateurs en France aujourd'hui, les réseaux téléphoniques mobiles proposent une alternative au réseau téléphonique commuté. Après la très forte période de croissance de ces dernières années, le marché de la téléphonie mobile s'est stabilisé, il devient donc impératif de fidéliser les clients grâce à des méthodes de management de la relation client. Afin de quantifier les efforts qu'il est rentable de porter vers les clients, il est primordial pour le marketing de bénéficier d'une estimation de leur valeur. Dans le contexte spécifique de la téléphonie mobile prépayée, les clients ne sont pas engagés contractuellement avec leur opérateur. Les clients étant libres de cesser leur activité sans préavis, prévoir les dates de rechargement permet d'en déduire une durée de survie potentielle ce qui représente un fort enjeu.

Les appels passés par les clients, les services consommés et les données transférées transitent tous par le réseau téléphonique mobile, toutes ces consommations représentent un volume de données extrêmement important. A cause de l'importance de ce volume de données, il serait pratiquement impossible pour les acteurs du marketing de traiter individuellement leurs clients (ce qui les obligerait à prendre des décisions macroscopiques) sans méthodologie adaptée.

L'extraction de connaissances dans les bases de données présentée dans [Fayyad et al. 1996] est une méthode qui nous permet de faire face au volume grâce à un enchaînement de traitements (cf. Fig. 1).

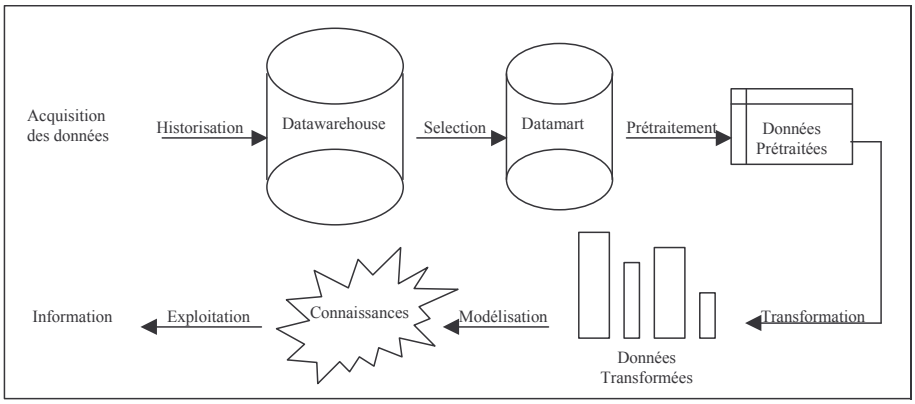

Fig 1 : *Processus d'ECD mis en oeuvre*

Cette partie présente une réalisation, suivant cette méthodologie, menée dans le contexte de la téléphonie mobile prépayée. Nous présentons la modélisation du délai de rechargement qui est ici assimilé à une durée de survie dans le domaine médical (pour la mise en place d'indicateur quant à la survie des clients). Le délai de rechargement est ainsi représenté par un modèle paramétrique, puis la probabilité de rechargement est calculée par la méthode de Kaplan-Meier [Kaplan et Meier, 1958]. Ceci permet de déduire une probabilité de survie du client. Ensuite, nous présentons le modèle de valeur de client et présentons son utilisation. L'objectif de cette partie est de présenter la conception d'un système d'information décisionnel dédié au calcul de la valeur client en téléphonie mobile prépayée.

Nous allons dans un premier temps introduire la spécificité du système d'acquisition des données, ensuite nous décrirons les modèles de rechargement et de valeur client mis en place, nous conclurons sur les perspectives.

2.1.1 Système d'acquisition des données

Le système d'acquisition des données est composé de deux environnements distincts. L'environnement de production est la composante réseau par laquelle transitent les communications mobiles, l'environnement décisionnel permet d'historiser et traiter l'information issue de l'environnement de production afin qu'elle soit exploitable. La figure 2 (système d'acquisition des données) présente les deux environnements du système celui de la production et celui décisionnel.

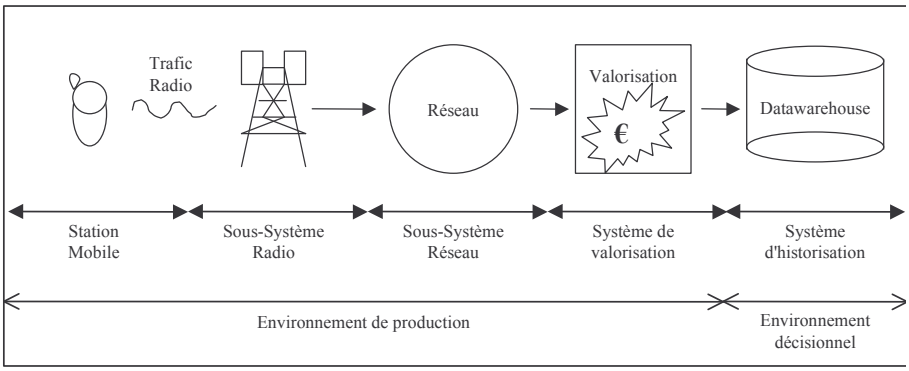

Fig 2 : *Système d'acquisition des données*

2.1.2 Système de valorisation des appels prépayés

Le système informatique de gestion des lignes prépayées permet de gérer les rechargements qu'effectuent les clients pour créditer leur compte de consommation, ainsi que les caractéristiques des lignes: autorisations d'appels, services souscrits, réserve de communication, etc. Ce système de gestion des lignes prépayées est couplé à un système de valorisation qui permet de débiter sur la réserve de communication des clients le montant correspondant aux appels qu'ils passent. Pour comparaison, le système de valorisation équivalent en téléphonie mobile post-payée correspond à la facturation mensuelle des clients. Ici les clients paient à l'avance leurs communications, et peuvent utiliser leur crédit pendant une durée limitée. Le débit du montant correspondant aux appels passés a lieu en temps réel sur le crédit de communication prépayé.

L'unité d'information au cœur du réseau téléphonique est le ticket de taxe, qui caractérise un appel en indiquant sa durée, ainsi que de nombreuses caractéristiques du réseau au moment de l'appel. Les tickets de taxes décrivent une durée d'appel fixe ; un appel téléphonique passé par un client est constitué d'un ou plusieurs tickets de taxes. Le système de valorisation des appels agrège les tickets de taxe correspondant à un appel pour obtenir la durée totale. Le client est donc ainsi débité du montant correspondant à son appel.

Le nombre de clients des opérateurs de téléphonie mobile se comptant en millions, chaque client passant en moyenne plusieurs appels par jour, les tickets de taxe générés par le réseau prennent très rapidement une volumétrie considérable, ce qui rend leur traitement difficile. C'est pourquoi, le système d'information décisionnel est alimenté en sortie du système de valorisation des appels.

2.1.3 Historisation des données acquises

Le réseau et le système de valorisation sont des systèmes informatiques opérationnels qui ont de très fortes contraintes de qualité de service fourni aux utilisateurs du réseau. Il est complètement impossible d'entreprendre des traitements à fin d'analyse (processus décisionnel) directement sur les bases de données associées à ces systèmes sans en pénaliser le fonctionnement. Un entrepôt de données alimenté directement par les données générées par le réseau permet de les historiser et de les rendre disponible à tous les acteurs du décisionnel au niveau de l'entreprise : cette mémoire de l'entreprise est appelé datawarehouse.

2.2 Indicateur décisionnel de comportement client et modélisation de la survie

L'objectif de cette section est d'estimer la probabilité de rechargement des clients dans les six mois qui suivent une date de référence donnée : la date de modélisation. La méthode que nous utilisons s'organise en deux étapes [Hill et al. 1990] [Kalbfleisch et Prentice 2002] [Lawless 2002] Une première étape consiste à modéliser et prévoir le délai de rechargement de chaque client, à partir des rechargements effectués durant la période d'apprentissage. Nous effectuons dans la deuxième étape une analyse de survie, qui permet d'affecter à chaque client sa probabilité de rechargement dans les six mois suivant la date de modélisation. Afin d'évaluer la qualité de la modélisation de la première partie, nous comparons les délais prévus à ceux des rechargements effectués lors de la période de test.

Certains rechargements ne sont pas observés durant la période d'apprentissage, mais on sait que les délais concernant ces rechargements sont au moins supérieurs aux durées séparant les derniers rechargements de la fin de la période d'apprentissage. Ces délais de rechargement non observés sont dits censurés :

Lorsque le délai exact d'occurrence de l'événement associé à la donnée de survie n'est pas observé alors que l'on possède des informations partielles sur ce délai, on dit que la donnée est censurée. Si nous appelons T, la variable aléatoire correspondant à la donnée de survie et C le délai observé, nous définissons les trois types de censures par :

Censure à droite : $T > C$

Censure à gauche : $T < C$

Censure par intervalle : $C_1 < T < C_2$

Dans le cadre de notre propos, nous sommes confrontés à des données censurées à droite à la fin de la période d'apprentissage. C'est la date de censure qui fixe l'échéance de prévision du délai de rechargement.

2.2.1 Estimation des délais de rechargement

Comme nous l'avons vu en introduction de cette partie, nous modélisons les délais de rechargement par un modèle paramétrique en ajustant une loi paramétrique à la variable aléatoire décrivant la donnée de survie. Nous utilisons une loi gamma [Saporta, 1990] qui est traditionnellement [Kalbfleisch et Prentice 2002] [Lawless, 2002] utilisée dans les études de survie. Nous cherchons ici un modèle qui soit de plus hétérogène possible, qui tienne compte de l'influence des covariables sur la distribution de la donnée de survie.

Les modèles hétérogènes utilisés sont appelés « Modèles à temps accéléré » (SAS 2000). La raison de cette appellation est expliquée par la forme du modèle, qui s'écrit généralement : $y = X\beta + \sigma\varepsilon$

où, $y = \log(T)$ est le logarithme du temps de survie,

X est la matrice des covariables (variables explicatives),

β est un vecteur des paramètres inconnus de régression,

σ est un paramètre d'échelle,

ε est un vecteur d'erreur distribué selon une loi connue.

Le modèle à temps accéléré suppose que l'effet des variables indépendantes sur une distribution de délais de rechargement est multiplicatif. Généralement, la fonction d'échelle

est $\exp(x'\beta)$, où x' désigne le vecteur transposé de x représentant les valeurs des covariables et β est un vecteur de paramètres inconnus.

Ainsi, si T_0 est le délai de rechargement correspondant aux valeurs nulles des covariables, le modèle accéléré de temps d'échec nous donne : $T = \exp(x'\beta)T_0$.

Si on pose $y = \log(T)$ et $y_0 = \log(T_0)$, on obtient $y = x'\beta + y_0$.

Il s'agit d'un modèle linéaire avec y_0 comme terme d'erreur. En termes de probabilité de survie, ce modèle devient $\Pr(T > t \mid \mathrm{x}) = \Pr(T_0 > \exp(-x'\beta)t)$

La probabilité de gauche de l'équation est évaluée pour une valeur de covariables x donnée. Le côté droit est calculé en utilisant la distribution de probabilité à une valeur d'échelle près de l'argument. Le côté droit de l'équation représente la valeur de la distribution de la fonction de survie évaluée à $\exp(-x'\beta)t$.

Un paramètre d'interception et un paramètre d'échelle peuvent être utilisés dans le modèle. En termes de délai de survie, les effets du terme d'interception et du terme d'échelle sont respectivement l'échelle du délai de rechargement et la puissance du délai de rechargement.

C'est-à-dire, si $\log(T) = \mu + \sigma \log(T_0)$, alors $T = \exp(\mu)(T_0)^{\sigma}$

2.2.2 Estimation des paramètres

La technique d'estimation utilisée est la méthode du maximum de vraisemblance. Nous estimons à la fois les paramètres de la distribution de Y_0 et du vecteur b . Pour f désignant la densité de la variable aléatoire Y_0 , la vraisemblance s'écrit :

$$L(t_1, t_2, t_3, .., t_n, \theta) = \prod_i [f(w_i, \theta)] \prod_j [S(w_j, \theta)] \prod_k [S(w_k, \theta) - S(w_{k+1}, \theta)] \prod_l [1 - S(w_l, \theta)]$$

avec $w_i = Y_i - \sum_m b_m x_{im}$

$\prod_i [f(w_i, \theta)]$ la vraisemblance des défaillances observées,

$\prod_j [S(w_j, \theta)]$ les défaillances censurées à droite,

$\prod_k [S(w_k, \theta) - S(w_{k+1}, \theta)]$ les défaillances censurées par intervalle,

$\prod_l [1 - S(w_l, \theta)]$ les défaillances censurées à gauche.

Les paramètres de la loi de Y_0 ainsi que le vecteur b sont estimés par le maximum de vraisemblance à partir de la vraisemblance finale. S'il n'existe pas de solution analytique au problème de maximisation, une méthode algorithmique est employée (méthode de Newton Raphson (SAS 2000)).

2.2.3 Comportement client et Estimation de la probabilité de rechargement

L'objectif de ce paragraphe est d'obtenir la fonction de survie, qui correspond dans notre application à la probabilité de rechargement. La signification des variables explicatives introduites dans le modèle est vérifiée par un test du chi2.

Nous discrétisons le délai prévu dans la section 2.2.1 pour le prochain rechargement des clients en dix segments homogènes numérotés de 1 à 10. Chaque individu est alors affecté à un segment. Nous effectuons une analyse de survie en stratifiant la variable segment. Nous modélisons le délai de rechargement observé par l'estimation de Kaplan-Meier qui est une méthode d'estimation des modèles non paramétriques de données de survie qui permet de tenir compte des données censurées. Dans le cadre de cette estimation, les délais de rechargement sont classés en ordre croissant et les quantités suivantes sont définies :

d_j est le nombre de défaillances à t_j,

n_j est le nombre d'individus soumis à risque à t_j,

c_j est le nombre de sujets censurés entre t_j et t_{j+1}.

La fonction de survie est estimée par :

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right)$$

2.3 Indicateur décisionnel pour l'aide à la mise en place de campagnes marketing et modélisation de la valeur client

Notre but ici est de modéliser la valeur d'un client caractérisé par ses covariables $x_1,...,x_n$ et de présenter comment ce modèle peut aider le marketing et les ventes à constituter de meilleures cibles de campagnes. Nous développons ici la définition de la fonction de valeur, et les modèles qui sont exploités.

2.3.1 Ciblage de Clients pour les Campagnes Marketing

Les campagnes de ventes ont été le domaine d'application de nombreuses techniques de statistiques et d'extraction de connaissances à partir de données, et un bon nombre d'éléments théoriques sont maintenant disponibles pour estimer les bénéfices d'une campagne ciblée de vente. [Piatetsky-Shapiro et Masand, 1999] présente la manière d'estimer le bénéfice d'une campagne proposant une offre si nous appelons :

N le nombre total de clients,

T la fraction des clients ciblés qui ont le comportement désiré (c'est-à-dire réponse à la campagne),

B le bénéfice de l'acceptation de l'offre A par un client correctement identifié,

C le coût de faire une proposition A à un client, que ce soit une cible ou pas.

On définit ***Profit(P)*** le bénéfice de faire la proposition à P pour cents de tous les clients (où $0 < P < 1$). Le bénéfice de faire une proposition à tous les clients est ***Profit(1.0)*** $= NTB - NC = N(TB - C)$. Selon cette définition nous notons rapidement qu'il y a un fort besoin d'identifier les clients qui répondront à la proposition pour maximiser le bénéfice des campagnes. [Rosset et al., 2001] présentent l'évaluation des modèles de

prévision pour des campagnes de vente indépendamment des méthodes sous-jacentes au modèle lui-même. Cette fonction de bénéfice est conçue pour aider les décideurs à constituer les cibles campagnes permettant d'en tirer le meilleur profit. Les experts disposent d'une connaissance, et s'il pourrait être intéressant (bénéfice immédiat) de cibler certains clients, il peut être plus avantageux de les contacter plus tard afin de préserver leur valeur à moyen terme. Nous identifions donc le besoin de mesurer la valeur à terme de clients.

2.3.2 Indicateur de valeur à terme des clients

Afin d'avoir une estimation de la valeur économique à gagner tant que les clients sont actifs, il est possible d'utiliser l'indicateur de valeur à terme des clients (LTV). Cet indicateur est défini dans [Rosset et al., 2003] par : $LTV = \int_0^{+\infty} S(t)v(t)D(t)dt$ où les trois facteurs sont :

La valeur du client $v(t)$ au temps t pour $t \geq 0$, et $t = 0$ l'origine,

La durée de service (LOS), décrivant la probabilité de survie du client,

Un facteur d'actualisation $D(t)$, qui décrit combien vaut aujourd'hui chaque euro gagné à une future date t.

Détaillons ci-dessous ces fonctions.

- **Fonction de valeur**

Le calcul de la valeur du client est habituellement un calcul direct fondé sur l'information récente concernant le client : données d'usage, plan tarifaire, rechargements, appels vers les centres d'appels, etc.. Ce calcul est présenté par [Rosset et al., 2002] comme une combinaison de prévisions, d'analyse de la tendance et de séries chronologiques. Dans nos expériences la fonction de valeur des clients prend en compte la valeur moyenne des clients.

- **LOS**

Comme nous l'avons vu au paragraphe 2.2.3 la durée de service est modélisée ici par la technique d'analyse de survie paramétrique qui présente l'avantage d'expliquer la valeur de survie en fonction des valeurs des covariables. La durée de service peut être modélisée par des approches non paramétriques d'analyse de survie, car les modèles paramétriques purs supposent que la fonction de survie a une forme paramétrique qui ne tient pas compte de certaines irrégularités porteuses de sens. Le modèle à risque proportionnel de Cox présenté dans [Cox et Oakes, 1984] peut être employé. Pour une fonction de survie donnée nous définissons la probabilité "instantanée" que le client arrête son activité au temps t (phénomène de dénommé "*churn*" dans la littérature anglophone) par $f(t) = -dS/dt$ et la fonction de risque $h(t) = f(t)/S(t)$: Le modèle à risque proportionnel de Cox suppose un modèle pour la fonction de risque $h(t)$ de la forme $h_i(t) = \frac{f_i(t)}{S_i(t)} = \lambda(t)\exp(\beta' x_i)$, ou alternativement $\log(h_i(t)) = \log(\lambda(t)) + \beta' x_i$.

Il y a un effet linéaire paramétrique fixe pour tous les covariables excepté le temps, qui est expliqué dans le risque plancher $\lambda(t)$.

♦ **Facteur d'actualisation**

Le facteur d'actualisation utilisé dans nos expérimentations est fourni par des experts de l'activité, son mode de calcul n'est pas détaillé ici. Deux modèles populaires présentés dans (Rosset et al. 2002) sont

le modèle exponentiel : $D(t) = \exp(-\alpha t)$ avec $\alpha \geq 0$ ($\alpha = 0$ signifiant pas d'actualisation)

la fonction seuil $D(t) = I\{t \leq T\}$ pour $T > 0$ environ (où I est la fonction indicatrice)

2.4 Indicateur décisionnel pour l'aide à la prédiction et Estimation des effets des campagnes

Comme nous l'avons vu plus haut, de façon opérationnelle la conduite de campagnes fait appel à la connaissance du domaine détenue par les acteurs du marketing. Le modèle de valeur à terme des clients peut être utilisé pour estimer si le ciblage d'un client pour une campagne donnée est bénéfique pour l'entreprise. En effet, si nous notons G le coût encouru, et $v^{(i)}(t)$ et $S^{(i)}(t)$ le changement de valeur et de la durée de service si le client accepte la proposition. Nous pouvons estimer le changement de LTV des clients par

$$LTV^{(i)} - LTV = T.\left(\int_0^{\infty}\left[S^{(i)}(t)v^{(i)}(t) - S(t)v(t)\right]D(t)dt - G\right) - C$$

Les informations concernant les modifications de valeur et de durée de service de chaque proposition sont dans l'état actuel estimés par les experts du domaine. Ces paramètres précis doivent pouvoir être obtenus par l'étude des résultats d'offres comparables proposées auparavant, cependant il faut s'assurer de considérer ces changements toutes choses égales par ailleurs, ce qui est difficile à réaliser dans le domaine hautement concurrentiel de la téléphonie mobile. L'évaluation du modèle de valeur se fait à posteriori en comparant la valeur constatée du client sur une période de temps aux valeurs prévues par le modèle. Vu l'aspect temporel très important dans le modèle prédictif, nous identifions que la moindre erreur d'estimation la valeur au temps t_0 prend rapidement des dimensions importantes à cause de la projection au cours du temps, ceci nous amène à envisager des techniques plus performantes d'analyse de survie.

2.5 Conclusion

Nous avons présenté les travaux d'initialisation d'un système d'information décisionnel qui sont en cours de développement. Le modèle de valeur à terme du client intègre des composantes que nous estimons pour certaines et qui nous sont fournies pour d'autres. En ce qui concerne la modélisation de la durée de service, des techniques alternatives d'intelligence artificielle sont également appliquées. [Mani et al., 1999] utilisent un réseau de neurones semi-paramétrique où chaque survie possible correspond à son propre neurone de sortie (la survie est discrétisée au niveau mensuel). Ils illustrent qu'un modèle élaboré de réseau de neurones exécute mieux que le modèle à risque proportionnel sur leurs données. [Biganzoli et al., 2002] présentent comment les réseaux de neurones peuvent être appliqués à l'analyse statistique des données temporelles censurées. Les réseaux de neurones possèdent le défaut d'être difficilement interprétables, ce qui ne correspond pas forcément à une attente d'explication des décisions, nos perspectives de travail concernent donc le test de cette famille de modèles pour l'estimation des fonctions de survie, ainsi que la finalisation et le déploiement du Système d'Information Décisionnel

3 Processus d'ECD pour l'aide au développement d'une chaîne automatique de reconnaissance de cibles radar

3.1 Introduction

La classification automatique de cibles aériennes, et plus spécialement d'avions, a une longue histoire, parfois tragique, comme par exemple l'erreur de classification d'un avion civile ayant entraîné la perte de nombreuses vies il y quelques années.

L'absence de radar commerciaux ou militaires comportant un sous-système de classification suffisamment robuste justifie l'intérêt et la difficulté d'une activité de recherche dans ce domaine.

Le problème traité dans cette section s'insère donc dans le cadre général de l'identification d'une cible aérienne non-coopérative à partir de la rétrodiffusion d'un signal radar multifréquentiel et l'approche proposé par notre laboratoire consiste à robustifier la classification en utilisant deux types d'information (à 1 et 2 dimensions) comme nous le verrons par la suite.

Le cadre général étant fixé, l'objectif de cette partie est d'utiliser le processus d'extraction des connaissances à partir des données (ECD) et l'ingénierie système comme des outils d'aide à la conception de la chaîne complexe de reconnaissance de cible radar que nous souhaitons mettre en place.

La suite de ce chapitre est organisé de la façon suivante : dans un premier temps nous allons présenter le processus d'ECD particularisé aux données radar avant de rentrer plus dans le détail en décrivant l'architecture logique de la fonction que nous souhaitons réaliser et nous mettrons en évidence les points important relatifs aux données et aux différents traitements envisagés, afin de mettre en évidence le lien direct entre cette chaîne et le processus d'ECD, enfin nous introduirons un point important pour l'aide à la conception de la fonction globale à réaliser, à savoir la prise en compte d'indicateurs décisionnels tels que, la qualité des données et les indicateurs de complexité.

3.2 Processus d'ECD particularisé aux données radar

L'objectif de cette section est de présenter la chaîne d'extraction de connaissances appliquée aux données radar.

Le terme originel anglais pour extraction des connaissances à partir des données (ECD) est « Knowledge Discovery in Databases » et a été introduit par [Fayyad et al., 1996].

L'extraction des connaissances est un processus interactif et itératif, constitué de cinq phases allant de l'acquisition et la préparation des données (pré-traitement et transformation des données) jusqu'à l'interprétation et l'évaluation des résultats, en passant par la phase de recherche des informations : le « data mining ». Les cinq étapes de l'ECD illustrées par la figure 3 sont développées ci-dessous.

Précisons que les types de traitements utilisés dans la chaîne sont particularisés à l'application de reconnaissance automatique de cibles dans le domaine radar qui sera présentée au paragraphe 3.3.

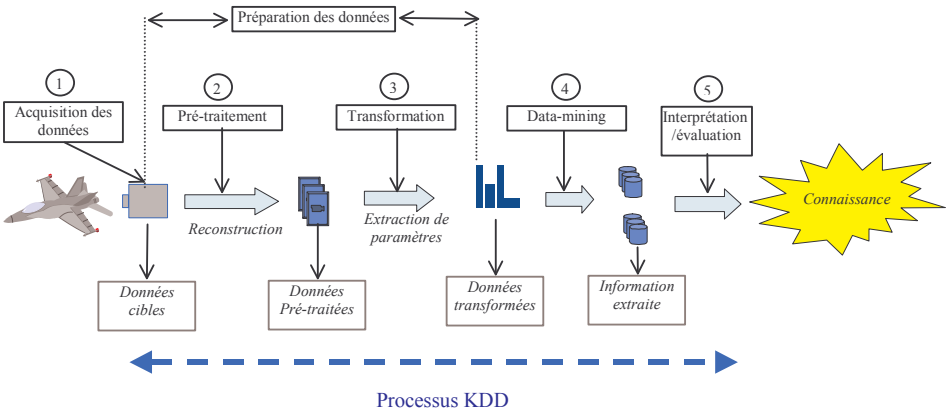

Fig 3 : *Les cinq étapes de la chaîne d'ECD*

Les éléments constituants de la chaîne d'extraction dans le cas du traitement des données radar sont :

- *L'acquisition des données*, première étape du processus, particulièrement importante dans le processus d'extraction étant donné que le système d'acquisition a une influence non négligeable sur le qualité des données et donc sur les performances globales de la fonction à réaliser.
- Le *pré-traitement*, deuxième étape du processus est une phase de transformation des signaux fréquentiels en profils distance (par transformée de Fourier inverse ou méthodes super-résolution), ou en image radar à ouverture de synthèse inverse (ISAR).
- La *transformation des variables*, troisième étape du processus, présente les données sous la forme exigée par l'algorithme d'extraction de connaissances. Paramètres extraits dans notre application : positions des pics correspondants à la projection des points brillants sur les profils de distance, et extraction de contours ou reconnaissance de forme sur les images ISAR.
- *L'application d'un algorithme d'extraction de connaissances*, quatrième étape du processus, met en évidence les informations sous-jacentes qui structurent les données. C'est le cœur du système d'extraction de connaissances
- *L'interprétation et l'évaluation des informations extraites*, dernière étape du processus, a pour objectif de produire de la connaissance sur le domaine d'étude en s'assurant que les conclusions émises correspondent à des phénomènes réels. C'est la suppression des connaissances inutiles ou redondantes, et la transformation des connaissances intéressantes en connaissances compréhensibles par l'utilisateur. C'est aussi le retour possible à n'importe quelle étape du processus. Cette étape peut amener à des modifications structurelles (choix des méthodes de n'importe quelle étape du processus).

Les étapes concernant le pré-traitement et la transformation des variables constituent la phase de préparations des données.

Dans ce qui suit, le processus ECD qui décrit une succession d'étapes fonctionnelles sera particularisé et détaillé en fonction de traitements prévus pour réaliser la reconnaissance/identification de cibles radar. Ce niveau de détail conduit à définir une architecture logique c'est à dire celle des traitements qui seront à développer puis à tester sous différentes conditions environnementales et procédures d'acquisition.

3.3 La chaîne de reconnaissance automatique de cibles radar

Comme déjà indiqué précédemment, notre objectif est de concevoir une chaîne automatique de reconnaissance de cibles radar [Nebalin 1994] [Pearson, 1975] [Demeter, 1996] fondée sur la mesure des signatures radar à haute résolution à une dimension (profils distance) et à deux dimensions (images dites ISAR pour Inverse Synthetic Aperture Radar).

L'architecture logique de cette chaîne de reconnaissance est présentée sur la figure 4.

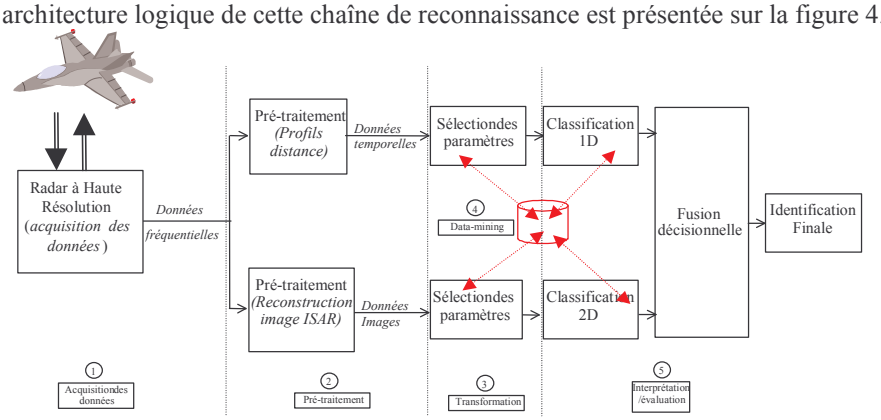

Fig 4 : *Architecture logique de la chaîne de reconnaissance automatique de cibles radar (à 1 et 2 dimensions)*

Notons dès à présent que la partie basée sur les données temporelles (profils distance) est finalisée [Radoi, 1999], par contre la partie concernant les données images est actuellement en cours de développement.

A ce grain de détail il semble important d'apporter quelques précisions sur les données et les traitements utilisés.

Les données expérimentales

Le signal complexe retrodiffusé (en quadrature et en phase) par la cible éclairée par le signal radar multifréquentiel (large bande) est mesuré à des fréquence discrètes et fourni la réponse fréquentielle de la cible c'est à dire la surface équivalente radar de la cible en fonction de la fréquence.

Données à une dimension

Cette représentation de la cible peut être transformée dans le domaine spatial au moyen de la transformée de Fourier [Mensa 1981], ou d'autres algorithmes d'analyse spectrale tels que les méthodes super-résolution [Walton 1987] [Kay, 1988]. Cette représentation spatiale constitue le *profil distance* de la cible.

Dans l'approximation haute-fréquence (c'est à dire quand la longueur d'onde du radar est très inférieur aux plus petits détails de la cible), les pics présents dans les profils distance correspondent aux points brillants dominants de la cible. Pour être plus précis, un profil distance correspond en réalité à la projection des points brillants dominants de la cible sur l'axe de visée du radar comme indiqué sur la figure 5.

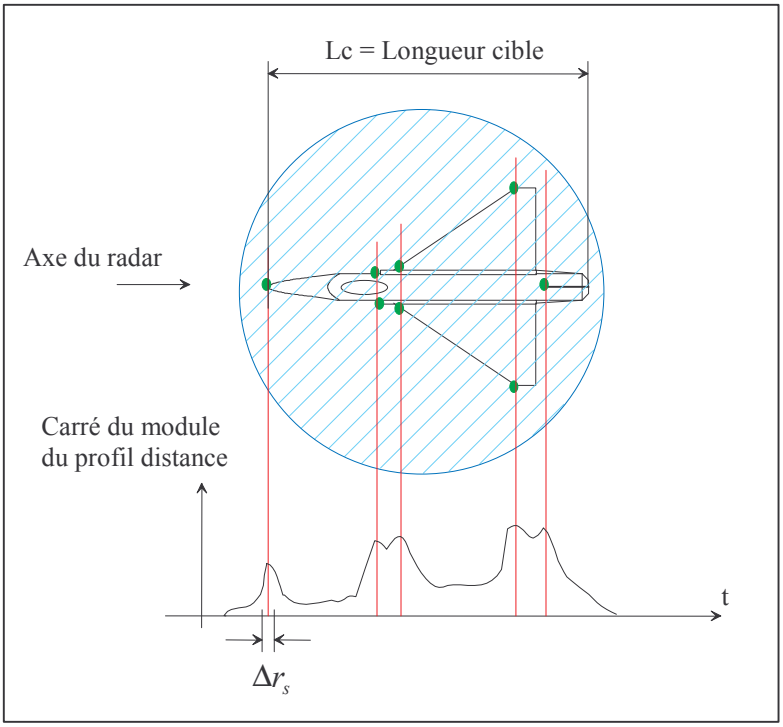

Fig 5 : *Obtention du profil distance d'une cible.*

Bien qu'il conserve uniquement une information partielle concernant la forme de la cible, le profil de distance est encore représentatif du type de l'objet, ce qui fait qu'il est considéré comme équivalent à la signature temporelle de la cible.

Données à deux dimensions

La formation d'images radar bi-dimensionnelle (images ISAR) pour la reconnaissance de cible a été étudié par plusieurs groupes [Compton, 1987] [Gupta, 1994] [Odendaal, 1994] et (a prouvé son intérêt) est très prometteuse pour améliorer la discrimination entre cibles. La raison en est évidente : en projetant les points brillants de la cible sur une seule dimension (profils distance), une part importante de l'information sur la cible est perdue.

Par contre, sur une image ISAR (bi-dimensionnelle), l'information transverse est préservée ceci améliore la capacité du système à faire une distinction fiable des cibles.

C'est pourquoi nous espérons que la fusion décisionnelle en sortie des classifieurs mono et bi-dimensionnels sera en mesure de rendre la classification plus robuste que les systèmes existants.

Précisons le traitement réalisé pour obtenir une images ISAR. La reconstruction ISAR consiste a restituer la réponse transverse de la cible. On entend par réponse transverse de la cible la projection des points brillants sur un axe perpendiculaire à l'axe de visée du radar (figure 5).

Cette réponse transverse est directement accessible à la mesure puisqu'elle est obtenue par analyse spectrale (Doppler) du signal reçu. Cette analyse s'effectuera pour chaque case distance des profils distance obtenus lors de l'éclairement de l'avion. Le moyen le plus généralement utilisé est une simple transformée de Fourier sur le temps d'illumination T_i. On obtiendra alors N_{CD} profils Doppler de la cible (NCD : nombre de cases distance). Lesquels vont nous permettre de positionner les points brillants de la cible suivant l'axe transverse.

La résolution ainsi obtenue suivant l'axe transverse est de :

$$\Delta f_d \approx \frac{1}{Ti}$$

A l'issu de ce traitement, en combinant les deux traitements (suivant l'axe radial et transverse), nous obtenons une image en deux dimensions de la cible suivant l'axe radial et transverse par rapport au radar. Cette image Distance/Doppler représente, en fait, la fonction bidimensionnelle de réflectivité de l'objet observé. La résolution de cette image est fixée par l'étendue du domaine d'observation en fréquence (radial) et en angle (transverse).

Le principe de construction d'une image ISAR à partir de l'enregistrement des réponses fréquentielles de la cible sous différents angles de présentation, est détaillé sur la figure 6.

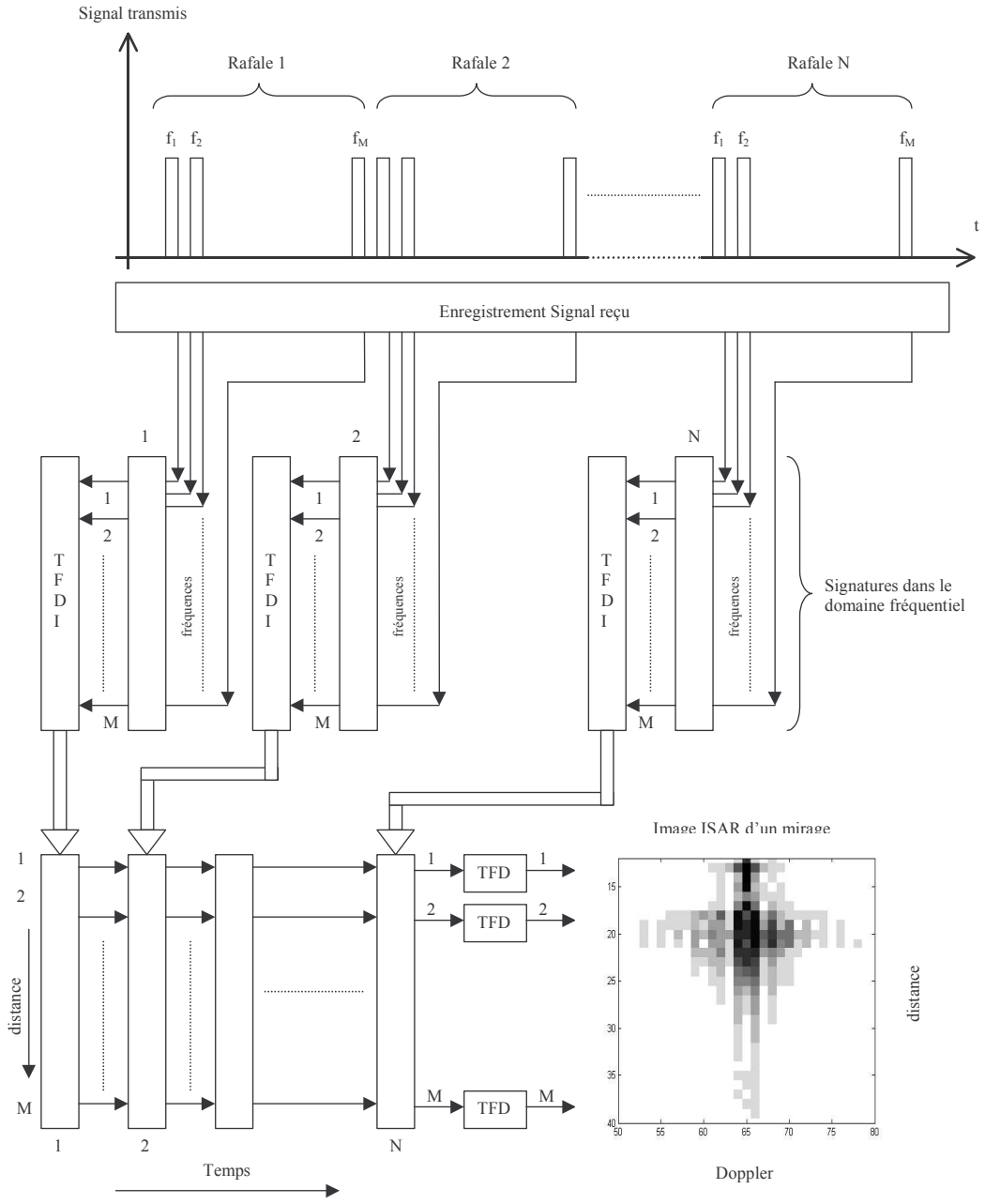

Fig 6 : *Principe d'obtention d'une image ISAR à partir d'un signal à sauts de fréquences.*

Notons que comme dans le cas des profils distance, il existe des méthodes plus performantes que la transformée de Fourier pour réaliser la reconstruction d'une images ISAR. Notons par exemple la généralisation des méthodes super-résolution à 2 dimensions [Odendaal 1994] [Radoi 1999].

Ainsi, le système de reconnaissance de cibles non coopérative que nous proposons (figure 4) consiste en un module d'acquisition de données (le signal fréquentiel (radar) rétrodiffusé), suivi des modules de pré-traitement constitués des algorithmes de formation des profils distance (1D) et des images ISAR (2D). Ces données mono et bi-dimensionnelles sont alors traitées par des modules d'extraction de caractéristiques (non présentés ici) produisant des vecteurs de caractéristiques qui sont classifiées dans les classifieurs 1D et 2D [Radoi 1999]. Une fusion appliquée sur le résultat de ces deux classifications a été retenue, elle devra démontrer son apport en amélioration de la fonction de reconnaissance notamment, pour le cas d'environnements incertains.

Cette solution (figure 4) en reconnaissance automatique de cibles radar étant clairement énoncée, il se pose le problème d'une *méthodologie adaptée*, permettant l'observation et éventuellement la remise en cause du système réalisé cela, par exploitation de connaissances qui seraient extraites au cours des différents tests de mise au point du système et des expérimentations en environnements réels. Ceci fait l'objet du paragraphe suivant.

3.4 Introduction à la modélisation de connaissances liées aux tests et à l'expérimentation de la fonction de reconnaissance

La modélisation des connaissances liée aux tests et à l'expérimentation de la fonction de reconnaissance a pour objectif de révéler des *influences* particulières entre les étapes du processus depuis l'acquisition des données fréquentielles jusqu'à l'évaluation et interprétation finale et qui ont pour effet de pénaliser le bon fonctionnement du système (ex. : modification particulière des conditions expérimentales de l'acquisition des données fréquentielles et chute de performance pour la fonction de reconnaissance). Pour cela, nous avons vu en section 3.2 et 3.3 qu'il était possible de caractériser une architecture logique en fonction des étapes suivies par le processus d'ECD. Cette caractérisation nous permet en première approximation, de prévoir au sein de l'architecture des points d'observation et d'évaluation partielle de la fonction de reconnaissance en cours de fonctionnement.

Les exemples d'influences sont en particulier :

L'influence de la mise au point du système d'acquisition sur la chaîne de traitements : évaluation des performances de l'identification

L'influence de la mise au point des systèmes couplés acquisition et préparation de données sur l'identification.

L'objectif retenu pour cela concerne l'estimation, l'analyse et l'exploitation d'indicateurs de qualité des données et de complexification des étapes amont du processus ECD (i.e. acquisition, préparation, transformation). Dans le contexte de la validation de chaînes de traitements une étude préliminaire a permis de mettre en évidence différents types d'indicateurs de complexité et leur possible exploitation pour l'aide à la conception et validation de chaînes de reconnaissance [Hoeltzener et al., 2003].

L'un des axes d'intérêt pour l'exploitation de tels indicateurs est celui de leur utilisation possible pour mieux contrôler le maintien d'une rétroaction sur l'une des étapes du processus ECD (ex. : préparation des données) ceci, sous la condition de vérifier par exemple, une amélioration significative voire, l'optimisation de la qualité des données en aval du processus ECD (ex. : réduction de l'incertain sur les données avant l'extraction des connaissances). Le maintien de cette rétroaction peut notamment dépendre de l'observation de la complexification de l'étape concernée (ex. : en temps, ressources, complexité algorithmique, nombre de tests) et cela relativisé aux autres étapes et à la performance attendue pour le système global de reconnaissance (ex. : rapidité, taux de reconnaissance, robustesse).

4 Conclusion

Les travaux de recherche présentés dans cet article sont orientés vers la définition et la mise en place de systèmes d'information dans le contexte d'extraction de connaissances et d'informations à partir de données volumineuses. Deux applications ont été présentées dans cet article. La première est réalisée dans le cadre de la téléphonie mobile prépayée avec comme objectif la mise en place d'un système d'information pour l'aide à la mise en place de campagnes de marketing. La seconde, réalisée dans le cadre des données radar a pour objectif de réaliser un outil d'aide à l'expérimentation et à la conception de chaînes de reconnaissance de cibles. Les volets qui ont été présentés dans le papier concernent,

- D'une part, la phase d'extraction des connaissances à proprement dite, vue au travers de la modélisation des connaissances et de la mise en place, en particulier d'indicateurs décisionnels en téléphonie mobile,
- D'autre part, la phase amont du processus ECD introduisant la possibilité de rétroagir à partir de la prise en compte de multiples modélisations et simulations physiques et de techniques de fusion, en particulier afin de réduire l'incertain et l'imprécis sur les données avant le passage par l'étape d'extraction des connaissances.

Références

[Alain, 2001], Alain J.M., Présentation du réseau GSM, http://www.lirmm.fr/~ajm/Cours/01-02/DESS_TNI/TER9/prercqu/fonction.htm, LIRMM, 2001.

[Baruch, 2003], Baruch B.W, Institute for marine biology and coastal research , the University of South Carolina, Colombia, http://inlet.geol.sc.edu/qaqchp/qaqchp.html, 2003.

[Bhattacharyya et Sengupta . 1991], Bhattacharyya A. et Sengupta D.L, Radar cross section analysis ans control, Artech House, 356 pages, ISBN 0-89006-371-0, 1991

[Biganzoli et al, 2002], Biganzoli E., Boracchi P., Marubini E., A General Framework for Neural Network Models on Censored Survival Data, Neural Networks Archive, vol 15, Issue 2, pp.209-218, 2002

[Briand. et Guillet 2001], Briand H. , Guillet F ;, Extraction des connaissances et apprentissage, Vol. 1 N° 1-2/2001, Eds Hermès.

Canada Centre for remote sensing, http://www.ccrs.nrcan.gc.ca/ccrs
[Compton, 1987], Compton R., Two-dimensional imaging of radar targets with the MUSIC algorithm. In: Technical Report 719267-14, Ohio State University Electroscience Laboratory, Electrical Engineering Department, 1987.
[Cox et Oakes ,1984], Cox D.R. et Oakes D., Analysis of Survival Data, CRC Press, 1984.
[Cox, 1972], Cox D.R, Regression Models and Life Tables, Journal of the Royal Statistical Society, B34: pp. 187- 220, 1972
[Demeter et Radoi, 1996],Demeter S., Radoï E., Radar target classification using neural networks, SBORNIK, 1/1996, pp. 45-52, Brno, République Tchèque, 1996
[Fayyad et al, 1996], Fayyad U., Piatetsky-Shapiro G., Smyth P., The KDD process for extracting useful knowledge from volumes of data, Communications of the ACM archive, Volume 39, Issue 11, pp. 27-34, 1996.
[Franklin, 2002], Franklin D., NOOA Satellite and Information Services, http://lwf.ncdc.noaa.gov/oa/climate/research/crn/crnprogress.html
Gemini Telescope on Mauna Kea, http://www.gemini.edu/sciops/sys-verif
[Gupta ,1994], Gupta I., High-resolution radar imaging using 2-D linear prediction. IEEE Transactions on Antennas and Propagation 42 1, pp. 31-37, 1994.
[Helsen et, Schmittlein 1993], Helsen K, Schmittlein D.C, Analyzing Duration Times in Marketing : Evidence for the Effectiveness of Hazard Rate Models, Marketing Science, Vol. 12, No. 4, pp. 395-414, 1993.
[Hill et al, 1990], Hill C., Com-Nougue C., Kramar A., Moreau T. O'Quigley J. Senoussi R., Chastang C., Analyse Statistique des Données de Survie, INSERM / Flammarion, 1990.
[Hoeltzener et al, 2003], Hoeltzener B. Pellen F., Khenchaf A, Analysis of complexity, different levels of exploitation : application to radar automatic target recognition, 16th conference in system engineering, Coventry UK, 2003.
[Hoeltzener et al, 2003], Hoeltzener B. Pellen F., Khenchaf A., Knowledge discovery from data process, fusion and supervised control on information extraction, SETIT'03, Sousse, Tunisie.
[Kalbfleisch et , Prentice 2002], Kalbfleisch J.D. et Prentice R.L, The Statistical Analysis of Failure Time Data, John Wiley & Sons; 2nd edition, ISBN: 047136357X, 2002.
[Kaplan et Meier 1958], Kaplan E.L, Meier R., Nonparametric Estimation From Incomplete Observations, Journal of the American Statistical Association, pp. 457-481, 1958.
[Kay, 1988], Kay S., Modern spectral estimation theory and application. , Prentice Hall, Englewood Cliffs, NJ, 1988.
[Krim et Viberg, 1996], Drim H., Viberg M., Two decades of array signal processing research, IEEE SP Magazine, Vol. 13, No 4, pp. 67-94, 1996.
[Lawless, 2002], Lawless J.F., Statistical Models and Methods for Lifetime Data, John Wiley & Sons; 2nd edition, ISBN: 0471372153, 2002.
[Mani et al, 1999], Mani D.R, Drew J., Betz A., Datta P., Statistics and Data Mining Techniques for Lifetime Value Modeling, Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 94-103, 1999.
[Mensa , 1981], Mensa D., High resolution radar imaging. , Artech House, Norwood, Massachusetts, 1981.
[Nebalin, 1994], Nebalin V.G., Methods and techniques of radar recognition, Artech House, London, 1994.

[Odendaal et al, 1994], Odendaal J; Barnard E. et Pistorius C., Two-dimensional superresolution imaging using the MUSIC algorithm. IEEE Transactions on Antennas and Propagation 42 10, pp. 1386-1391, 1994.
[Pearson et al, 1975], Pearson L.W., Van Blaricum M., Mittra R.,A new method for radar target recognition based on the singularity expansion for the target, IEEE International Radar Conference, Arlington, VA, pp. 452-56, 1975.
[Piatetsky-Shapiro et Masand, 1999], Piatetsky-Shapiro G., Masand B., Estimating Campaign Benefits and Modeling Lift, Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 185-193, 1999.
[Pipino et al, 2002], Pipino L., Lee W., Wang Y., Data Quality Assessment, Communication of the ACM Vol. 45, pp. 211-218, 2002.
[Radoi ,1999], Radoï E., Contribution à la reconnaissance des objets 3D à partir de leur signature Electromagnétique, Thèse de Doctorat, Université de Bretagne Occidentale, 18 juin 1999.
[Redman, 1996], Redman T., Data quality for the information age, Artech House Publishers, ISBN 0-89006-8836, 1996.
[Rosset et al, 2003], Rosset S., Neumann E., Eick U., Vatnik N., Lifetime Value Models for Decision Support. Data Mining and Knowledge Discovery Journal, Vol. 7, pp. 321-339, 2003.
[Rosset et al, 2001], Rosset S., Neumann E., Eick U., Vatnik N., Ydan Y., Evaluation of Prediction Models for Marketing Campaigns, proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp.456-461, 2001.
[Rosset et al, 2002], Rosset S., Neumann E., Eick U., Vatnik N., Ydan Y., Customer lifetime value modeling and its use for customer retention planning, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 332-340, 2002.
[Saporta ,1990], Saporta G., Probabilités, Analyse des données et statistique, Editions Technip, 1990.
SAS Institute Inc. (2000), SAS OnlineDoc®, Version 8, SAS institute Inc., 2000.
Space telescope science institute,http://www.stsci.edu/stsci/meetings/adassVII/ballesterp.htm
Spot System, http://www.spot.com/home/system/introsat/payload/vegetati/vegetati.htm
Statistics Canada, source officielle de statistiques et de produits sur la société et l'économie au Canada, http://www.statcan.ca/english/edu/power/ch3/quality/quality.htm
[Walton, 1987], Walton E., Comparison of Fourier and maximum entropy techniques for high-resolution scattering studies. Radio Science 22, pp.350-356, 1987.
[Wang, 1995], Wang R.Y.,A framework for analysis of data quality research, IEEE Transactions on Knowledge and Data Engineering, volume 7, no 4, pp.623-638, 1995.
[Wehner ,1987], Wehner R., High Resolution Radar, Artech House, Boston, 1987

Summary

The research tasks presented in this paper are oriented to the definition and the installation of information systems in the context of knowledge and information discovery in huge databases. The engineering and life cycle of the systems and the exploitation of knowledge bases are used to contribute with the design, the tests and the validation of data processing sequences leading to decision-making. Two significant applications are presented in this communication. The first one is realized within the framework of mobile telephony with the

objective of designing a tool allowing the estimation of the value of the customers in prepaid mobile telephony. The second, carried out in the framework of the data radar aims to produce a tool with experimentation and design of targets recognition chains. The principal goal of the two applications is the description of various decisional indicators put in perspective in the KDD process.