Université Lumière Lyon 2

THESE pour obtenir le grade de docteur En Informatique

présentée et soutenue publiquement par

Gaëlle LEGRAND

Le 20 Décembre 2004

Approche Méthodologique de Sélection et Construction de Variables pour l'Amélioration du Processus d'Extraction des Connaissances à partir de Grandes Bases de Données

Faculté des Sciences Economiques et de Gestion
Informatique et information pour la société (INSA, ECL, LYON 1, LYON2)
Préparée au sein du laboratoire ERIC Equipe de Recherche en Ingénierie des Connaissances
Informatique
Sous la direction de Nicolas Nicoloyannis

COMPOSITION DU JURY M. Younès BENANI M. Gilles VENTURINI M. Yves LECHEVALLIER M. Gilbert RITSCHARD M. Djamel Abdelkader ZIGHED M. Nicolas NICOLOYANNIS

Table des matières

à M., T. et G.

Remerciements

En premier lieu, je tiens à remercier Nicolas NICOLOYANNIS sans qui cette thèse n'aurait pas vu le jour. Il m'a accordé sa confiance, son amitié, son soutien et ses connaissances tout au long de ces trois années.

Je remercie également l'ensemble des membres du laboratoire ERIC qui ont permis que ce travail s'effectue dans une ambiance chaleureuse. J'ai une pensée particulière pour Valérie et Lydie qui m'ont grandement aidé et facilité toutes les démarches administratives.

J'aimerais remercier Gilles VENTURINI et Younès BENNANI qui ont accepté de rapporter ce mémoire ainsi que Djamel Zighed, Yves Lechevallier et Gilbert Ritschard qui ont accepté d'être membre du jury.

Je remercie chaleureusement mes amis Mihaela, Marian, Tiffany, Pierre, Jérémy et Cécile pour leur soutien, leur aide et leur amitié de tous les instants. Un merci à Laurent d'avoir toujours été là depuis bien longtemps.

Je tiens à remercier mes parents et mon frère à qui je dois ce que je suis et qui m'ont toujours soutenu quelques soient mes choix. Merci pour tout ce que vous m'apportez !!

J'ai une pensée affectueuse pour mes grands-parents.

J'ai une pensée amoureuse pour stéphane. Je le remercie pour sa patience, sa tolérance et son affection. Il a réussi à supporté mes nombreux sautes d'humeur de ces derniers mois !!

Enfin je remercie parents, amis et tous ceux que je n'ai pas cités nommément.

Introduction Générale

L' Extraction de Connaissances à partir de Données (ECD) consiste à parcourir d'immenses volumes de données contenus dans une base, à la recherche de connaissances. C'est une discipline qui se situe à l'intersection de différents domaines tels que l'informatique, l'intelligence artificielle, l'analyse de données, les statistiques, la théorie des probabilités, l'optimisation, la reconnaissance de formes, les bases de données et l'interaction Homme-Machine,… Fayyad [1] donne une définition de l'ECD que la communauté scientifique francophone traduit de la manière suivante : L'ECD est le processus non trivial, interactif et itératif qui permet d'identifier des modèles valides, nouveaux, potentiellement utiles et compréhensibles à partir de bases de données massives.

Le terme processus signifie que l'ECD se décompose en plusieurs opérations, figure 1. Ces opérations peuvent être regroupées en cinq phases majeures :

- **La phase de compréhension du domaine étudié** : Lors de cette phase, une analyse du problème et des contraintes qui lui sont attachées doit permettre la collecte de données brutes. Ces données se composent d'individus ou objets et des variables qui leurs sont associés et qui doivent permettre de décrire au mieux le problème traité. L'utilisateur ne sait pas encore si les données qu'il a réunies seront toutes adaptées à son problème ni si ces données seront suffisantes. Nous sommes en présence des données initiales.
- **La phase de prétraitement** : Lors de cette phase, un prétraitement est effectué à la

fois sur les individus et sur les variables. Cette phase de prétraitement consiste à nettoyer les données, les mettre en forme, traiter les données manquantes, échantillonner les individus, sélectionner et construire des variables. On obtient ainsi un ensemble de données cibles. Cette phase a une place importante au sein du processus d'ECD car c'est elle qui va déterminer la qualité des modèles construits lors de la phase de fouille de données. Elle peut prendre jusqu'à 60% du temps dédié au processus d'ECD.

- **La phase de fouille de données** : Cette phase intègre le choix de la méthode d'apprentissage qui va être employée et son paramétrage. Ces choix doivent tenir compte des contraintes liées au domaine étudié ainsi que des connaissances que les experts du domaine peuvent nous fournir. L'algorithme sélectionné est alors appliqué aux données cibles dans le but de rechercher les structures sous-jacentes des données et de créer des modèles explicatifs ou prédictifs.
- **La phase de post-traitement** : Cette phase consiste en l'évaluation et la validation des modèles construits lors de la phase précédente. Ce n'est qu'après cette phase que les données et l'information que l'on en a tirées deviennent des connaissances.
- **La phase d'interprétation et d'exploitation des résultats** : L'interprétation des résultats qui sont sous forme de modèles ou de règles permet d'obtenir des connaissances. Ce sont ces connaissances qui seront fournies à l'utilisateur.

La finalité de l'ECD est de pouvoir traiter des données brutes et volumineuses, et à partir de ces données d'établir des connaissances directement utilisables par un expert ou un non expert du domaine étudié.

Les techniques d'ECD deviennent de plus en plus prisées au sein du monde industriel. En effet, les promesses de l'ECD en terme de valorisation de l'information ne peuvent laisser insensibles les acteurs industriels. Tout d'abord parce que l'information apparaît, de nos jours, comme un élément stratégique déterminant. Ensuite parce que les avancées technologiques en informatique permettent d'augmenter les capacités de stockages et de calculs. Ainsi, si l'on considère comme exemple l'ensemble des tickets de caisse d'un supermarché sur une période 10 ans, il est aisé d'imaginer la quantité de données présentes, la diversité des caractéristiques, et donc la difficulté conséquente d'une exploitation de l'information présente. Pourtant, on dispose là d'une immense source d'information, à savoir une quantité suffisamment importante de données pour établir une classification pertinente de la clientèle ainsi que son comportement typique. Le processus d'ECD résout de manière efficace ces difficultés et fournit les connaissances attendues.

Cependant, le processus d'ECD ne se passe pas sans encombre. La taille des bases de données étant de plus en plus importante, l'amélioration de la qualité de représentation des données est devenu un problème majeur de l'extraction des connaissances à partir des données.

L'une des difficultés principales liée à la représentation des données est la dimension des données. Le problème de la dimension des données concerne le nombre et la qualité des variables descriptives caractérisant chacun des individus. Ce problème peut se

résumer par la phrase de Liu et Motoda, [2], Less is more qui signifie que si l'on désire extraire de l'information utile et compréhensible à partir de nos données, il convient en premier lieu de retirer les parties non pertinentes.

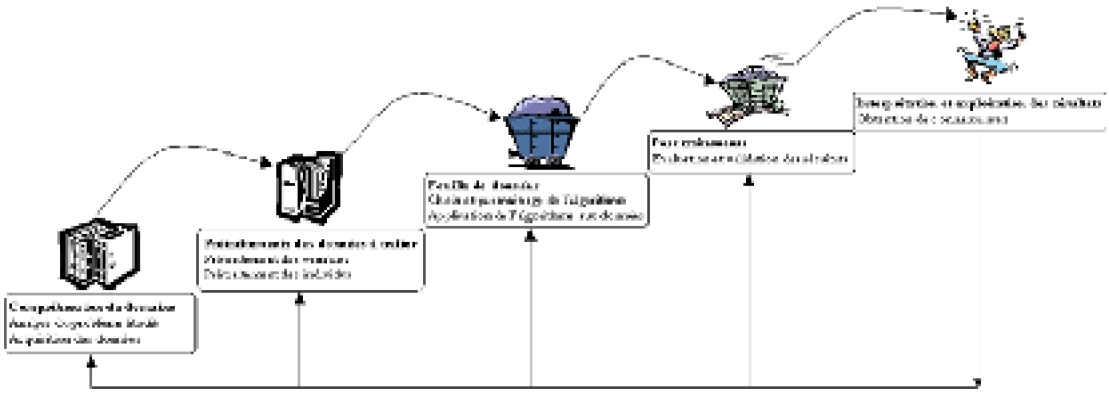

Figure 1 Processus de l'Extraction de Connaissances à partir des Données.

L'utilisateur qui veut couvrir tous les aspects existants d'un problème particulier et obtenir une connaissance compréhensible définit et considère un grand nombre de variables. Or, parmi ces variables certaines seront inutiles. En effet, il est souvent difficile voire impossible de discerner les variables pertinentes des variables non pertinentes ce qui pousse l'utilisateur à s'emparer de toutes les variables disponibles. De plus, les sources de données peuvent être multiples et la fusion des données issues de chacune de ces sources conduit à la création d'un ensemble contenant des variables inutiles et redondantes.

La solution que l'on peut apporter à cette difficulté est la sélection d'un sous-ensemble de variables. La sélection de variables est un processus permettant l'élimination de variables inutiles et/ou redondantes et l'élimination du bruit pouvant être généré par certaines variables. Le processus d'apprentissage est accéléré et la précision prédictive des algorithmes d'apprentissage peut être améliorée. Aucune nouvelle variable n'est générée et certaines variables sont éliminées, ainsi l'espace de représentation des données est réduit.

La deuxième difficulté est la qualité des données brutes. La qualité d'apprentissage est fortement liée à la présence de variables discriminantes. Les variables composant l'espace de représentation des données ne sont pas forcément les mieux adaptées pour décrire le problème. Or, en l'absence de nouvelles informations disponibles, il convient de créer de nouvelles variables qui permettront d'expliciter l'espace de représentation.

La construction de variables permet de créer de nouvelles variables. Elle est le processus qui découvre les informations manquantes dans une relation entre variables et qui augmente l'espace des variables en créant des variables supplémentaires. Après le processus, un certain nombre de variables supplémentaires sont disponibles. Par exemple, les variables longueur et largeur dans un problème à deux dimensions peuvent entraîner la création de la variable aire . La construction de variables est l'application d'un ensemble d'opérateurs booléens ou numériques à un ensemble de variables existantes, débouchant sur la construction d'une ou plusieurs nouvelles variables destinées à être utilisées pour la description de la variable endogène. La

construction de variables augmente l'espace des variables en créant des variables supplémentaires. Cependant, aucune information extérieure à l'ensemble d'apprentissage n'est ajoutée lors du processus de construction.

Les travaux de cette thèse se situent au centre des préoccupations de l'ECD. Ils touchent essentiellement les problèmes liés à la phase de prétraitements des données du processus d'ECD. Nous nous intéressons aux problèmes de dimension et de qualité des données. Nous voulons grâce aux processus de sélection et construction de variables modifier l'espace de représentation des données afin d'en améliorer sa qualité et déterminer le moment où ces processus sont nécessaires.

Il nous semble important de tenir compte dans nos travaux de deux facteurs :

- L'applicabilité des méthodes proposées à la réalité : nos travaux tiennent compte des contraintes liées aux caractéristiques des problèmes industriels telles que la limitation des coûts calculatoires, la limitation des coûts de stockage, la conformité à des exigences issues de la distribution de l'information,…
- L'utilisabilité des méthodes : nous désirons tenir compte au sein de nos travaux de l'interaction entre Homme et Machine en intégrant l'expertise et les connaissances humaines aux méthodes développées.

Nous nous plaçons dans un cadre d'apprentissage supervisé qui consiste à déterminer sur une base d'un nombre fini d'individus, la relation entre un ensemble de variables exogènes et une variable endogène.

Ces travaux ont été effectués dans le cadre d'un projet en collaboration avec France Telecom. Le but principal de ce projet était un tour d'horizon et une comparaison de l'ensemble des méthodes de sélection et de construction de variables. Aussi, cette thèse comporte un état de l'art pour chacun de ces processus particulièrement développé. Ces travaux ont été validés par la publication d'un ensemble d'articles (Voir Annexe).

Ce document s'organise de la manière suivante : Le premier chapitre introduit la formalisation et les définitions qui seront utilisées tout au long du document. Le chapitre 3 est entièrement consacré au processus de sélection de variables tandis que le chapitre 4 est dédié au processus de construction de variables. Le chapitre 5 aborde le problème lié à la détermination du moment où les processus de sélection et/ou de construction sont nécessaires.

Nous tenons également à préciser que les différentes expérimentations proposées dans ce rapport ont été possibles grâce à l'utilisation des logiciels libres : Sipina développé au laboratoire ERIC [3], Weka, [4], de l'Université de Waikato en Nouvelle-Zélandeet d'un logiciel mis au point au cours de ces travaux.

Chapitre 1 Formalisations et définitions

legrand_g_chapitre01.pdf

Chapitre 2 La Sélection de variables

legrand_g_chapitre02.pdf

Chapitre 3 Construction de variables

legrand_g_chapitre03.pdf

Chapitre 4 Gestion de la phase de prétraitement

legrand_g_chapitre04.pdf

Conclusions et perspectives

La phase de prétraitement est une étape essentielle du processus d'Extraction des Connaissances à partir des Données. Elle permet d'extraire l'information utile des données. Les travaux de cette thèse se situe essentiellement au sein de la phase de prétraitement des données du processus d'Extraction de Connaissances à partir des Données. Nous nous sommes intéressés aux problèmes qui concernent les variables initiales appartenant à l'ensemble des données brutes collectées lors de la phase de la phase de compréhension du domaine étudié.

Tableau 1 Synthèse des contributions.

Problèmes liés à la phase de prétraitement des données	Solutions proposées
Dimension des données brutes	Méthode de sélection variables
Qualité des données brutes	Méthode de construction de variables
Détermination du moment où il est nécessaire prétraiter les données	Double indice

Le chapitre 2 répond au problème de dimension des données grâce au développement d'une méthode de sélection des variables. Nous sommes partis du constat que les méthodes myopes fournissaient des résultats aussi satisfaisants que les autres méthodes de sélection tout en étant très rapides et facilement utilisables. Le seul problème réside dans la forme des résultats que ces méthodes produisent.

Le principe sous-jacent de notre méthode de sélection est lié à l'agrégation des préférences. Nous avons voulu que l'utilisateur expert ou non expert du domaine étudié puisse utiliser cette méthode. Pour cette raison, notre méthode est sous la forme d'une méta-méthode . L'utilisateur a ainsi la possibilité de choisir un ensemble de critères myopes. S'il ne sait lesquels sélectionner, nous lui proposons un ensemble de dix critères qui nous avons employés lors des expérimentations. Les différents critères sont appliqués aux variables initiales. Chaque critère fournit une liste de variables classées en fonction de leur pertinence. Afin d'obtenir comme résultat un sous-ensemble de variables, nous utilisons une méthode d'agrégation des préférences basée sur une notion de préférence large. Nous sommes alors en présence d'une liste de sous-ensembles de variables classés en fonction de la pertinence des variables les composant. L'utilisateur est de nouveau libre de choisir la forme du résultat : soit il préfère conserver cette liste de sous-ensembles afin d'y apporter sa connaissance, soit il désire obtenir un sous-ensemble de variables optimal . Dans ce dernier cas, nous nous plaçons dans une approche de type enveloppe et utilisons l'algorithme d'apprentissage pour déterminer le sous-ensemble de variables considéré comme optimal par notre méthode.

Les expérimentations nous ont permis de conclure que notre méthode de sélection permet d'améliorer dans la plupart des cas la qualité d'apprentissage et de garder une certaine stabilité du modèle. La taille de l'espace de représentation se retrouve réduite. Ainsi notre méthode est rapide et efficiente.

Cependant, nous aimerions l'améliorer par la prise en compte des résultats d'autres méthodes de sélection. Ces méthodes devront de préférence avoir un fondement théorique différent de celui de notre méthode : nous pensons en priorité à des méthodes telles que MIFS, les méthodes utilisant les algorithmes génétiques ou les méthodes utilisant des réseaux de neurones. Les méthodes doivent être relativement rapides et s'adapter à tous types de problèmes. Et ainsi, les variables qui seront sélectionner devront être considérées comme pertinentes par l'ensemble des méthodes de sélection choisies.

Le chapitre 3 s'interroge sur les problèmes liés à la qualité des données en proposant une méthode de construction de variables. Cette méthode est basée sur le principe de l'analyse topologique des arbres d'induction.

L'ensemble des règles qui vont nous servir de base de construction sont générées par l'application de l'arbre d'induction ID3 pour lequel la contrainte liée au gain d'information minimal a été supprimée. Pour chaque règle, une variable intermédiaire est créée sous la forme d'une conjonction des éléments formant la prémisse de la règle. Les variables intermédiaires sont alors regroupées en fonction de la conclusion de la règle qui leur est associée. Les nouvelles variables qui seront ajoutées à l'espace de représentation des données peuvent maintenant être construites. Elles sont sous la forme de disjonctions des variables intermédiaires. Les variables construites sont de type booléen et leur nombre est égal au nombre de classes de la variable endogène.

Notre méthode est relativement rapide et efficiente. En effet, les expérimentations montrent que la qualité d'apprentissage est améliorée après le processus de construction grâce à une modification de l'espace de représentation. La taille de cet espace s'en

trouve augmenter mais de manière non exagérée. Les modèles conservent également une certaine stabilité.

Dans nos travaux futurs, nous pensons nous tourner vers une approche se rapprochant plus de l'utilisateur, c'est à dire lui laissant plus de liberté sur le choix des paramètres de notre méthode. Nous voudrions que l'utilisateur puisse choisir l'arbre d'induction qui sera utilisé pour la génération de la base de construction. Pour cela, nous lui proposerons une liste d'arbres que nous aurons au préalable paramétrés afin d'obtenir le même type de résultats fournis par ID3 libéré de sa contrainte du gain d'information minimum.

Le double indice permettant de déterminer le moment où il est nécessaire de construire des variables est le travail le moins abouti de cette thèse. Le chapitre 4 propose un double indice qui étudie respectivement l'apport informationnel des données en apprentissage supervisé et la structure intrinsèque des données en non supervisé. Ce double indice est basé sur le coefficient Kappa. Le coefficient Kappa permet de mesurer le degré d'accord entre un jugement théorique et un jugement observé. Nous utilisons le coefficient Kappa afin de comparer d'une part le jugement d'un algorithme d'apprentissage supervisé et celui de la variable endogène et d'autre part le jugement d'un algorithme d'apprentissage non supervisé et celui de la variable endogène. La confrontation de ces deux indices nous permet de déterminer si la construction de variables est indispensable.

Nous pouvons présenter un système permettant l' optimisation de la phase de prétraitement. Ainsi la phase de prétraitement s'organise de la manière suivante : le processus de sélection de variables est appliqué ; ensuite le double indice est calculé pour le sous-ensemble de variables sélectionnées. Selon les valeurs du double indice, le processus de construction peut être soit conseillé, soit recommandé, soit considéré comme indispensable ou inutile.

Pour l'instant, le processus de sélection de variables est toujours appliqué. Un indice permettant de savoir s'il est nécessaire de sélectionner et/ou de construire des variables permettrait d'optimiser et d'améliorer la phase de prétraitement et par la suite la phase de fouille de données.

Bibliographie

legrand_g_biblio.pdf

Annexe : Liste des publications

Conférences nationales avec comité de lecture et actes

Gaëlle Legrand et Nicolas Nicoloyannis : Nouvelle méthode de construction de variables, In Rencontres de la société française de classification, SFC04, Bordeaux, 2004.

Gaëlle Legrand et Nicolas Nicoloyannis : Sélection de variables et agrégation d'opinions, In Rencontres de la société française de classification, SFC04, Bordeaux, 2004.

Gaëlle Legrand et Nicolas Nicoloyannis : Construction de variables et arbre de décision, *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04), Clermont-Ferrand* , Janvier 2004; *Revue des Nouvelles Technologies de l'Information* , Vol. 2, 409-414.

Pierre-Emmanuel Jouve, Gaëlle Legrand et Nicolas Nicoloyannis : Sélection rapide en apprentissage supervisé, *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04), Clermont-Ferrand* , Janvier 2004; *Revue des Nouvelles Technologies de l'Information* , Vol. 2, 409-414.

Pierre-Emmanuel Jouve, Gaëlle Legrand et Nicolas Nicoloyannis : Chaos Game

Representation et traitement des séries temporelles. In Rencontres de la société française de classification, SFC03, Neuchâtel (Suisse), 2003, 409-414.

Gaëlle Legrand, Walid Erray, et Marc Boullé : Un survey des méthodes de sélection d'attributs dans le data mining. In Rencontres de la société française de classification, SFC02, Toulouse, 2002, 409-414.

Revue nationale avec comité de lecture

Gaëlle Legrand et Nicolas Nicoloyannis : Sélection de variables et agrégation d'opinions, Revue des Nouvelles Technologies de l'Information, 2004, Cépaduès.