Data Mining 2 – Modélisation Statistique et Apprentissage
Master Statistique & Econométrie
UT1 Sciences Sociales, Toulouse
E-thème 1 **08 janvier 2018**
Enseignant: Gilles Cohen

Data Mining

1 Introduction

Le data mining est une discipline relativement nouvelle, et en cours d'évolution. La première conférence internationale sur la découverte de connaissances et le Data Mining ("KDD") a eu lieu en 1995. Selon une grande entreprise américaine de conseil et de recherche dans le domaine des techniques avancées le Gartner Group [1], "le Data Mining (fouille de données) est un procédé de découvertes de corrélations significatives, de règles et de tendances en parcourant de grands volumes de données stockées dans des référentiels en utilisant des technologies de reconnaissance de formes mais également des techniques statistiques et mathématiques".

En 2001 le Data Mining était prédit selon le magasine en ligne ZDNET News [5] comme étant "un des développements technologiques les plus révolutionnaires des dix prochaines décennies". La prestigieuse revue Technology Review du MIT (Massachusetts Institute of Technology) [6] a choisie le Data Mining comme l'une des dix technologies émergentes qui vont changer le monde.

De nombreuses autres définitions du data mining ont été données parmi lesquelles on peut citer :

- "Extraction d'informations intéressantes (non triviales, implicites, préalablement inconnues et potentiellement utiles) à partir de grandes bases de données.".
- "Le Data Mining est l'extraction d'informations ou de connaissances originales, auparavant inconnues, potentiellement utiles à partir de gros volumes de données" (d'après Frawley et Piateski-Shapiro) [2].
- "Il s'agit du processus de séléction, exploration, modification et modélisation de grandes bases de données afin de découvrir des relations entre les données jusqu'alors inconnues" (selon SAS-INSTITUTE) [2].
- "Le Data Mining est l'analyse d'un ensemble (souvent important)

d'observations qui a pour but de trouver des relations insoupçonnées et résumer les données d'une nouvelles manières, de façon qu'elles soient plus compréhensibles et utiles pour leurs détenteurs" (Hand et al. [3]).

— "Le Data Mining est un domaine pluridisciplinaire qui regroupe des techniques d'apprentissage automatique, de la reconnaissance de forme, des statistiques, des bases de données et de la visualisation pour apporter une réponse à l'extraction d'information provenant de base de données de grande taille" (Evangelos, Simoudi in Cabena et al. [4]).

La première définition est assez concise et capture l'essence du data mining.

Pour résumer le Data Mining correspond donc à l'ensemble des techniques et des méthodes qui à partir de données permettent d'obtenir des connaissances exploitables. Son utilité est grande dès lors que l'entreprise possède un grand nombre d'informations stockées sous forme de bases de données. Une distinction plus précise s'établit autour du concept de KDD (Knowledge Discovery in Database ou Découverte de Connaissances dans les Bases de données) et celui de Data Mining (Fouille de données). En effet, ce dernier n'est que l'une des étapes du processus de découverte de connaissances correspondant précisément à l'extraction des connaissances à partir des données (ECD). Avant de réaliser une étude Data Mining, il faut donc procéder à l'élaboration d'un data Warehouse (Entrepôt de Données).

2 Origine du Data Mining

De grandes quantités d'informations sont quotidiennement collectées. Cependant, qu'est ce que l'on exploite de ces données ? Quelles connaissances en sont réellement tirées ?

Dès 1984, dans son livre Megatrends [7], John Naisbitt constatait le fait que "nous sommes noyés dans l'information mais assoiffés de connaissance". La remarquable croissance actuelle du Data Mining et la découverte de connaissances ont été alimentées par la confluence heureuses d'une variété de facteurs [1] :

— la croissance explosive de la collecte des données (par exemple les scanners dans les supermarchés)

— le stockage des informations dans les entrepôts de données qui permettent à toute l'entreprise d'accéder à une base de données fiable

— la disponibilité croissante d'accès aux données à partir d'Internet et

d'intranets
— la pression concurrentielle pour accroître la part de marché dans un marché mondialisé
— le développement de logiciels de Data Mining prêts à l'emploi
— l'énorme croissance de la puissance informatique et de la capacité de stockage

Le Data Mining a vu le jour dans les années 80, quand les professionnels ont commencé à se soucier des grands volumes de données informatiques inutilisables tels quels par l'entreprise. Le Data Mining d'alors consistait essentiellement à extraire de l'information de gigantesques bases de données de la manière la plus automatisée possible ; contrairement à aujourd'hui où le Data Mining consiste à l'analyse qui suit l'extraction. Le Data Mining s'est donc dissocié du Data Warehousing.

Le data mining se situe à la confluence des statistiques et de l'apprentissage automatique. Une variété de techniques d'exploration de données et de construction de modèles existent depuis longtemps dans le monde des statistiques - la régression linéaire, la régression logistique, l'analyse discriminante et l'analyse en composantes principales, en sont des exemples.

3 La croissance rapide du Data Mining

Certainement le facteur le plus important contribuant à la croissance du data mining est la croissance des données. Le détaillant en gros Walmart en 2003 a capturé 20 millions de transactions par jour dans une base de 10 térabytes (1 térabyte équivaut à 1000000 mégabytes). En 1950, la plus grande entreprise n'avait suffisamment de données que pour occuper, sous forme électronique, plusieurs dizaines de mégaoctets. Lyman and Varian (2003) estimait que 5 hexabytes d'information était produit en 2002, le double de ce qui était produit en 1999 (un hexabyte correspond à un millionde terabytes). La croissance des données est entraîné non pas simplement par une économie en expansion et la base de connaissances, mais par la baisse du coût et la disponibilité croissante des mécanismes de saisie automatique des données. Non seulement il y a plus d'événements qui sont enregistrés, mais il y a également plus d'informations par événement. La croissance de l'Internet a créé une vaste arène pour la génération de nouvelles informations. En marketing, un changement d'orientation des produits et services vers une focalisation sur le client et ses besoins a créé une demande pour des données détaillées sur les clients. Les bases de données opérationnelles utilisées pour enregistrer les transactions individuelles en appui à l'activité métier de

routine peut traiter des requêtes simples, mais ne sont pas adéquates pour une analyse plus complexe et globale. Les données de ces bases de données opérationnelles sont donc extraites, transformées et exportées vers un entrepôt de données - une grande installation intégrée de stockage de données qui relie les systèmes de prise de décision d'une entreprise. Bon nombre des techniques exploratoires et analytiques utilisées dans l'exploration des données ne seraient pas possible sans la puissance de calcul d'aujourd'hui. Le coût en constante diminution du stockage de données a permis de construire les installations nécessaires pour stocker et mettre à disposition de vastes quantités de données. En bref, l'amélioration rapide et continue de la capacité de calcul est un moteur essentiel de la croissance de l'extraction de données.

4 Activités principales du data mining

1. Analyse exploratoire.
2. Modélisation descriptive : Cette activité permet de réaliser des "vues" de haut niveau d'un ensemble de données. On peut distinguer par exemple les tâches suivantes :
   (a) Détermination des distributions de probabilité des données (appelé parfois estimation de densité) ;
   (b) Elaboration de modèles décrivant la relation entre les variables (appelé parfois modélisation des dépendances) ;
   (c) Partitionnement des données en groupes, par analyse de groupes (clustering) ou par segmentation. Les algorithmes de clustering essaient de trouver des "groupes naturels", l'utilisateur peut spécifier que tous les cas doivent être répartis dans x groupes. Pour la segmentation, le but est de trouver des groupes homogènes liés à la variable à modéliser (e.g., segments comme les gros dépensiers).
3. Modélisation prédictive : Classification et Régression : Le but ici est de construire un modèle où la valeur d'une variable peut-être prédite à partir des valeurs d'autres variables. La classification est utilisé pour des variables "catégorielles" (e.g., variables Oui/Non ou réponses à choix multiples). La régression est utilisée pour des variables "continues" (e.g., âge d'une personne, pression sanguine).
4. Découverte de motifs et de règles : Cette activité consiste par exemple à trouver des combinaisons d'éléments qui se produisent fréquemment ensemble dans des bases de données transactionnelles ((e.g., produits qui sont généralement achetés ensemble, au même moment,

par un client chez un fournisseur, etc.) ou à trouver en astronomie des regroupements d'étoiles, peut-être de nouvelles étoiles, ou encore pour trouver des profils génétiques dans les essais de puces à ADN. Des analyses de ce genre peuvent être utilisés pour générer des règles d'association ; e.g., si une personne va au magasin pour acheter du lait, il va également acheter du jus d'orange. Le développement de règles d'association est pris en charge par des algorithmes dans de nombreux produits commerciales de data mining.

5. Recherche par le contenu : Ce type d'activité consiste à rechercher dans un nouvel ensemble de données des modèles similaires à un un motif connu. Cette approche de la reconnaissance de formes est le plus souvent utilisée avec des données textuelles (e.g., documents écrits, contenu de pages Web pages) ou avec des ensembles de données constitués d'images.

La figure suivante (fig.1 résume les différentes activités du processus d'extraction de connaissances dans les données.

5 Les étapes du Data Mining

En 1996, Osama Fayyad a proposé un processus pour la fouille de données qui a bien répondu aux besoins d'entreprises, et qui est devenu rapidement très populaire. Knowledge Discovery in Databases (KDD) a comme but l'extraction des connaissances, des motifs valides, utiles et exploitables à partir des grandes quantités de données, par des méthodes automatiques ou semi-automatiques. On peut remarquer sur la figure 2, ci-dessous , que le Data Mining(DM) représente seulement une étape dans le processus de KDD. Le processus de KDD est itératif et interactif.

Les neuf étapes de KDD sont les suivantes :

1. Analyse du problème d'application
   - choisir un problème précis, des objectifs tangibles et quantifiables
   - définir la manière dont la solution sera déployée
   - spécifier la solution
2. Obtenir les données qui seront utilisées dans l'analyse.
   - évaluer la qualité des données, détecter leurs insuffisances et pathologies
   - visualiser, analyser les distributions et les regroupements
3. Prétraitement des données

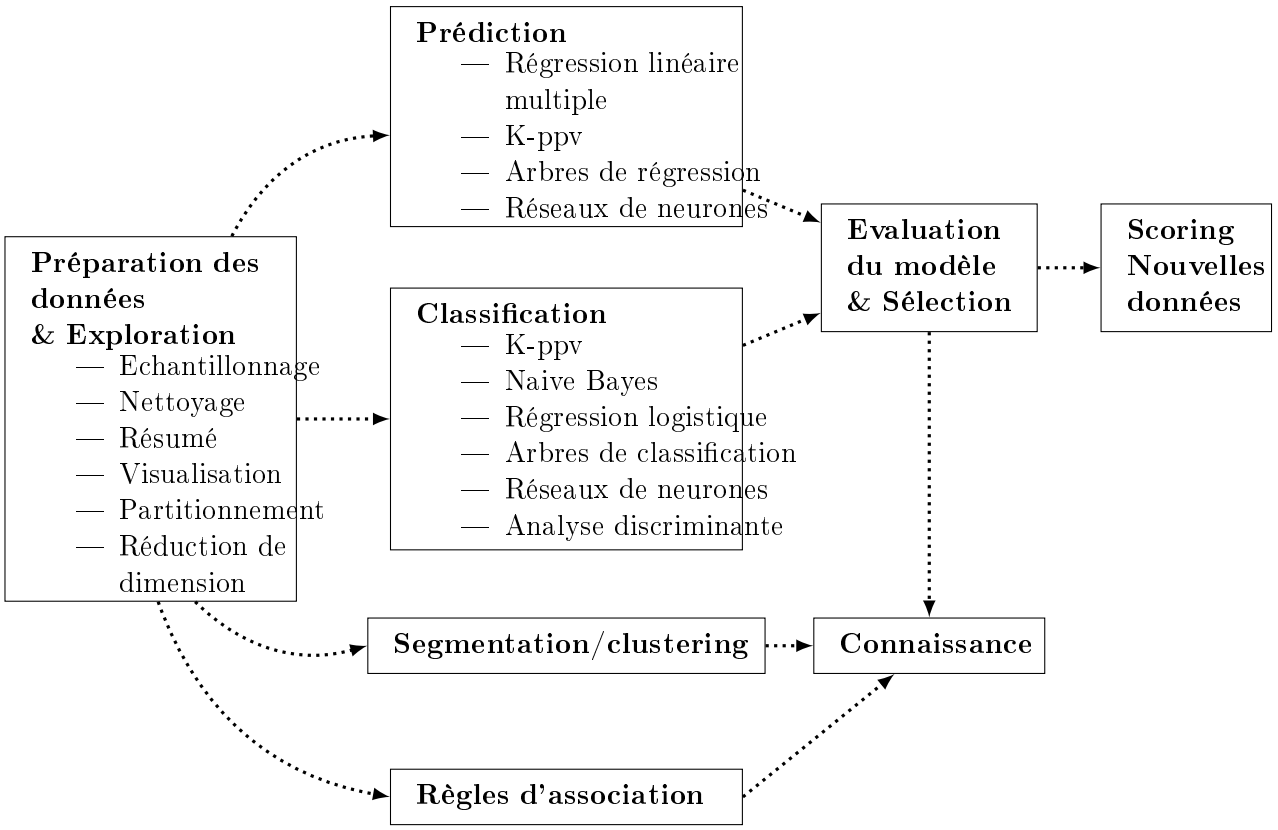

FIGURE 1 – Processus d'extraction de connaissances dans les données.

- nettoyage : suppression du bruit, valeurs manquantes ou abérrantes
- réduction des données
  - sélection des instances
  - sélection, extraction, combinaison des variables
- transformation des données
  - discrétisation des variables continues
  - numérisation des variables nominales
  - invention de nouvelles variables

4. Déterminer la tâche de data mining (classification, prédiction, clustering, etc...).
5. Choisir la technique de data mining à utiliser (régression, réseaux de neurones, clustering hiérarchiques...) en fonction du problème/des

données.

6. Utiliser l'algorithme pour réaliser la tâche. C'est typiquement un rocessus itératif.
7. Evaluation et interprétation des résultats
   — évaluation quantitative indispensable
   — compréhensibilité souvent capitale (ex. applications médicales)
8. Déploiement du modèle. Ceci implique l'intégration du modèle dans un système opérationnel et le faire tourner sur des données réelles afin de produire des décisions ou des actions.

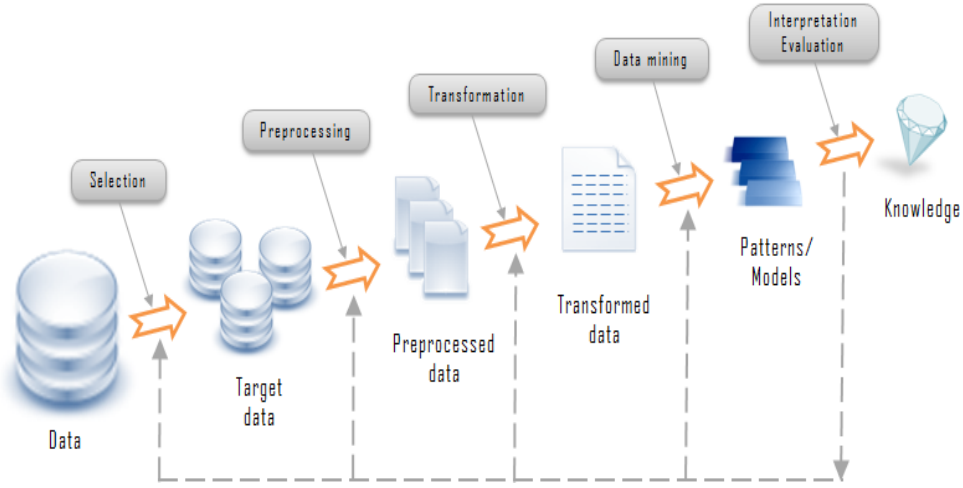

FIGURE 2 – Le processus de découvertes de connaissances dans les données - Fayyad. [11]

Une méthodologie méritant d'être également présentée et qui a été développé en 1996 pour répondre aux besoins des projets industriels de Data Mining est CRISP-DM [10] (CRoss-Industry Standard for Data Mining). Cette méthodologie fournit un aperçu du cycle de vie d'un projet de data mining. Elle identifie clairement les principales phases de ce processus au travers de tâches et des relations entre ces tâches. Même si le modèle ne le spécifie pas explicitement, il y a des relations possibles entre toutes les tâches en fonction des objectifs d'analyse et des données qui sont analysées.
CRISP-DM est décrit comme un processus hiérarchique constitué par plusieurs tâches, avec quatre niveaux d'abstraction : la phase, la tâche générique, la tâche spécialisée et l'instance du processus
CRISP-DM contient un cycle de six étapes : Business understanding (La

compréhension du business), Data understanding (La compréhension des données), Data preparation (La préparation des données), Modeling (La modèlisation), Evolution (L'évolution) et Deployment (Déploiement).

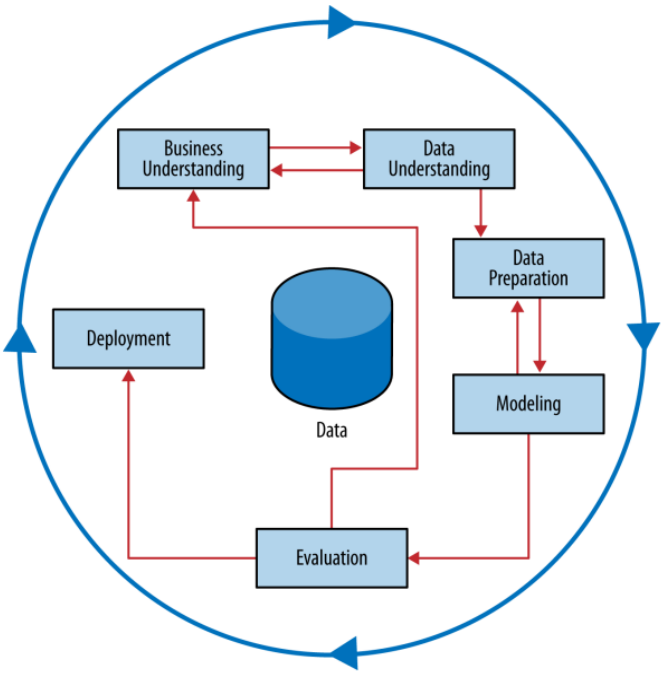

FIGURE 3 – Le cycle de vie du CRISP-DM.

1. Business understanding - cette phase initiale porte sur la compréhension des objectifs et des exigences du projet. Les connaissances acquises vont définir la problématique et le plan préliminaire pour accomplir ces objectifs.
2. Data understanding - elle commence avec une collection de données initiales et continue avec des activités afin de se familiariser avec les données, d'identifier les problèmes de qualité des données, de détecter des sous ensembles afin de construire des hypothèses pour l'information cachée.
3. Data preparation - cette phase contient toutes les activités nécessaires afin de construire la base de données finale.
4. Modeling - dans cette phase sont sélectées et appliquées plusieurs techniques sur les données et leur paramètres sont calibrés avec des valeurs optimales. Comme il y a plusieurs formes spécifiques de construc-

tion des données, la plupart du temps il est nécessaire de revenir une étape en arrière.

5. Evaluation - à ce niveau le(s) modèle(s) sont évalué(s) et les étapes suivies pour la construction du modèle sont réévaluées pour s'assurer que le projet respecte les objectifs du business, définis au début du projet.
6. Deployment - La création du modèle ne représente pas la fin du projet. Même si le but initial du projet est d'augmenter les connaissances de données, les connaissances acquises ont besoin d'être organisées et présentées d'une manière utilisable par le client.

6 Le coeur du processus de Data Mining : l'apprentissage

6.1 Types d'apprentissage

6.1.1 Apprentissage supervisé

L'apprentissage supervisé consiste à inférer un modèle de prédiction à partir d'un ensemble d'apprentissage, c'est-à-dire plusieurs couples de la forme observation, étiquette, où chaque étiquette dépend de l'observation à laquelle elle est associée. La variable étiquette est appelée variable dépendante ou cible. Un algorithme d'apprentissage supervisé a pour but de déterminer une fonction s'approchant au mieux de la relation liant les observations et les étiquettes à partir de l'ensemble d'apprentissage uniquement. Cette fonction doit par ailleurs posséder de bonnes propriétés de généralisation et ainsi être capable d'associer une étiquette adéquate à une observation qui n'est pas dans l'ensemble d'apprentissage.

6.1.2 Apprentissage non supervisé

L'apprentissage non supervisé consiste à inférer des connaissances sur des classes sur la seule base des échantillons d'apprentissage, et sans savoir à priori à quelles classes ils appartiennent. Contrairement à l'apprentissage supervisé, on ne dispose que d'une base d'entrées et c'est le système qui doit déterminer ses sorties en fonction des similarités détectées entres les différentes entrées (règle d'auto organisation). Il n'y a pas de variable dépendante.

La figure 4 donne une liste des algorithmes du datamining que chaque data-scientist devrait connaître.

7 Plan du cours

1. Introduction Data Mining
2. L'apprentissage supervisé (classification, régression)
   - Partitionnement récursif : Arbres et règles de décision : CART
   - Approches numériques linéaires : discriminants linéaires
   - Apprentissage par estimation de densité : méthodes bayésiennes, KNN, noyaux
   - Approches numériques non linéaires : réseaux de neurones artificiels,
   - Machine à vecteurs support
3. Techniques d'évaluation et d'expérimentation
   - Mesures de performances et stratégies d'expérimentation
   - Comparaison et sélection de modèles
4. L'apprentissage non supervisé
   - Les règles d'association

Références

[1] Daniel T. Larose, Traduction et Adaptation de Thierry Vallaud, Des Données à la Connaissance, Vuibert, 2005.

[2] http ://www.web-datamining.net/forum/faq.asp

[3] David Hand, Heikki Mannila et Padhiac Smyth, Principles of Data Mining, MIT Press, Cambridge, MA, 2001.

[4] Peter Cabena, Pablo Hadjinian, Rolf Stadler, Jaap Verhees et Alessandro Zanasi, Discovering Data Mining : From Concept to Implementation, Prentice Hall, Upper Saddle River, NJ, 1998.

[5] Rachel Konrad, Data Mining : Digging user info for gold, ZDNET News, 7 février 2001, http ://zdnet.com.com/2100-11-528032.html ?legacy=zdnn

[6] The Technology Review Ten, MIT Technology Review, Janvier/février 2001.

[7] John Naisbitt, Megatrends, 6th Ed., Warner Books, New York, 1986.

[8] V. Vapnik. Statistical Learning Theory. Wiley, 1998.

[9] L. Devroye, L. Györfy, and G. Lugosi. A Probabilistic Theory of Pattern Recognition. Number 31 in Applications of mathematics. Springer, New York, 1996.

[10] Daimler-Chrysler Project Overview, CRISP-DM, 1996 http ://www.crisp-dm.org/Overview/index.html [Cited : 07.09.2010.]

[11] Fayyad, U.M : Data Mining in the KDD Environment, http ://www.data-mining-blog.com/data- mining/data-mining-kdd-environment-fayyad-semma- ?ve-sas-spss-crisp-dm/ [Cited : 07.09.2010.]

FIGURE 4 – Algorithmes datamining.