

STA211 : Introduction au data-mining

Audigier Vincent, Ndeye Niang-Keita

06 février, 2019

- 1 Introduction
- 2 Historique et définitions
 - 2.1 De la donnée à la connaissance
 - 2.1.1 Données
 - 2.1.2 Connaissances
 - 2.1.3 Découverte de connaissances dans les bases de données
 - 2.2 L'émergence du data-mining
 - 2.2.1 secteurs concernés
 - 2.2.2 les facteurs de l'émergence
 - 2.3 Une évolution au cours du temps
 - 2.3.1 Des données de plus en plus volumineuses
 - 2.3.2 Les âges de la statistique
- 3 Effectuer une étude de data-mining
 - 3.1 Le data-mining étape par étape
 - 3.1.1 Comprendre et analyser les objectifs de l'étude
 - 3.1.2 Création d'une base de données
 - 3.1.3 Prétraitement et nettoyage des données
 - 3.1.4 Réduction des données
 - 3.1.5 Segmenter la population
 - 3.1.6 Choix et validation du modèle
 - 3.1.7 Itérations du processus
 - 3.1.8 Déploiement des modèles
 - 3.2 Les outils du data-mining
 - 3.2.1 Méthodes non-supervisées
 - 3.2.2 Méthodes supervisées
 - 3.2.3 Méthodes avancées
 - 3.2.4 Des logiciels dédiés
- 4 Conclusion
- Références

1 Introduction

Littéralement *fouille de données*, le data-mining est *l'application des techniques de statistique, d'analyse de données et d'intelligence artificielle à l'exploration et l'analyse sans a priori de grandes bases de données informatiques, en vue d'en extraire des informations nouvelles et utiles pour le détenteur de ces données* (Tufféry (2007)). Le concept est assez vague, ce qui explique les nombreuses autres définitions, comme par exemple selon Fayyad, Piatetsky-Shapiro, and Smyth (1996) : *Data Mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data* ou selon D. J. Hand (2000) : *I shall define data-mining as the discovery of interesting, unexpected, or valuable structures in large data sets.*

Le data mining constitue un nouveau champ d'analyse situé à l'intersection de la statistique et des technologies de l'information (bases de données, intelligence artificielle, apprentissage, etc.) dont le but est de découvrir des structures dans de vastes ensembles de données. Son caractère nouveau est lié à l'émergence de nouvelles technologies et aussi à leur appropriation par la société. En effet, le numérique a peu à peu intégré notre quotidien, et en l'utilisant nous générions une quantité de données qui augmente sans cesse. Néanmoins, il n'en a pas toujours été ainsi et les méthodes utilisées par le passé pour traiter des données n'avaient pas été envisagées dans ce nouveau contexte. Autrefois, avant les années 1970, la donnée était rare, on allait la récolter de façon à pouvoir répondre à des questions précises, définies en amont. Et puis à partir des années 1970, l'ordinateur a commencé à envahir la société, avec lui on commence à accumuler davantage de données, il n'est alors plus possible d'en extraire ``à la main'' la structure cachée. C'est l'époque de l'*analyse des données* où apparaissent des méthodes permettant justement de mettre en évidence cette structure cachée. A partir des années 1990, on passe à une autre échelle avec la démocratisation du web. Les données s'accumulent encore avec une qualité moindre et des données de types nouveaux et différents (images, sons, etc). La donnée n'est plus récoltée pour répondre à une question, elle est à disposition et on l'analyse a posteriori pour répondre à une problématique. Les méthodes d'analyse employées jusqu'à présent doivent alors être repensées. Finalement, à partir de 2010, on passe à l'ère des données massives (le *Big data*). La quantité de données générées explose, les jeux de données sont de l'ordre du Téraoctet ou plus encore. Il n'est plus possible de stocker les données sur un seul ordinateur, il faut à nouveau repenser les méthodes pour extraire de l'information de de toute cette quantité de données.

L'objectif de ce document d'introduction est de présenter d'une part comment le data-mining est apparu et comment il s'applique aujourd'hui dans différents secteurs d'activité, et d'autre part, de préciser les différentes étapes de sa mise en oeuvre et les méthodes sur lesquelles il repose. Ce document renverra parfois via des url, vers des sections prochaines du cours. Ces liens seront effectifs quand les sections correspondantes seront accessibles sur la plateforme pédagogique.

2 Historique et définitions

Le data-mining s'inscrit dans un processus plus large, celui de l'extraction des connaissances à partir d'un ensemble de données. Si aujourd'hui il s'est démocratisé avec de nombreuses applications dans la société, il résulte d'une évolution progressive et qui se poursuit encore aujourd'hui.

2.1 De la donnée à la connaissance

2.1.1 Données

Les données étaient classiquement des tableaux croisant des individus et des variables, mais au jour d'aujourd'hui, il peut tout aussi bien s'agir de textes, d'images, de vidéos, de sons, ou de tout autre support traitable par un ordinateur. Elles sont générées en quantité importante par les entreprises (publiques ou privées) mais aussi par les particuliers. On distingue :

- les données *opérationnelles* (ou transactionnelles): elles varient de plus en plus régulièrement avec le temps (par exemple, un ticket de caisse). Elles sont générées à la suite d'une action, sans optique d'analyse a priori. Elles feront l'objet d'une analyse a posteriori, dite *analyse secondaire* ;
- les données *non-opérationnelles* (ou de référence): elles ne varient pas ou peu, et ne dépendent pas des données transactionnelles (par exemple, l'adresse du client);
- les *méta-données*: ce sont des données portant sur la donnée elle-même (par exemple, le numéro de la caisse où a été généré le ticket).

2.1.2 Connaissances

Des associations et relations entre toutes ces données émergent des informations. Par exemple, l'analyse des données de transactions dans un magasin permet d'identifier les produits les plus fréquemment achetés ensemble. Cette information permet alors de mieux appréhender les comportements du client et améliore ainsi les *connaissances* de l'entreprise vis-à-vis de son client.

2.1.3 Découverte de connaissances dans les bases de données

On appelle KDD (Knowledge Discovery in Data bases) la découverte de connaissances dans les bases de données. Celui-ci désigne l'ensemble d'un processus d'analyse, depuis la collecte des données jusqu'à la présentation des résultats. Au sein de ce processus, le *data-mining* correspond à la phase d'*extraction des connaissances*. Plus précisément, un processus de KDD (*cf*Figure 2.1) se résume à :

- Sélectionner les données a priori pertinentes à l'objet de l'étude à partir des sources disponibles, ces dernières pouvant être variées et centralisées dans des entrepôts de données (*Data Warehouse*). Les architecture de bases de données pour le data-mining seront abordées spécifiquement ici (). Il s'agit de sélectionner à la fois des variables et des individus.
- Pré-traiter les données. Les données sont généralement incomplètes, redondantes, dupliquées ou sont entachées d'erreur. Lors de cette étape, il s'agit de corriger ces erreurs et anticiper la gestion des données manquantes. Les différents outils de pré-traitement seront développés ici () .
- Transformer les données. Il peut être nécessaire de discréteriser certaines variables ou fusionner certaines modalités.
- Extraire l'information contenue au sein des données. Choisir les méthodes d'analyse adaptées aux données et répondant aux problématiques de l'étude, puis les appliquer afin d'extraire l'information.
- Interpréter et évaluer les résultats des méthodes de fouille. Il s'agit de donner un sens aux résultats, émettre des hypothèses, valider celles-ci par des procédures adéquates d'inférence statistique ou empirique et de les confronter à l'expertise. Cette étape appelle généralement à reprendre le processus en amont afin de répondre à de nouvelles questions ou à gagner en précision par rapport aux résultats obtenus. En ce sens, le procédé de KDD n'est pas linéaire.

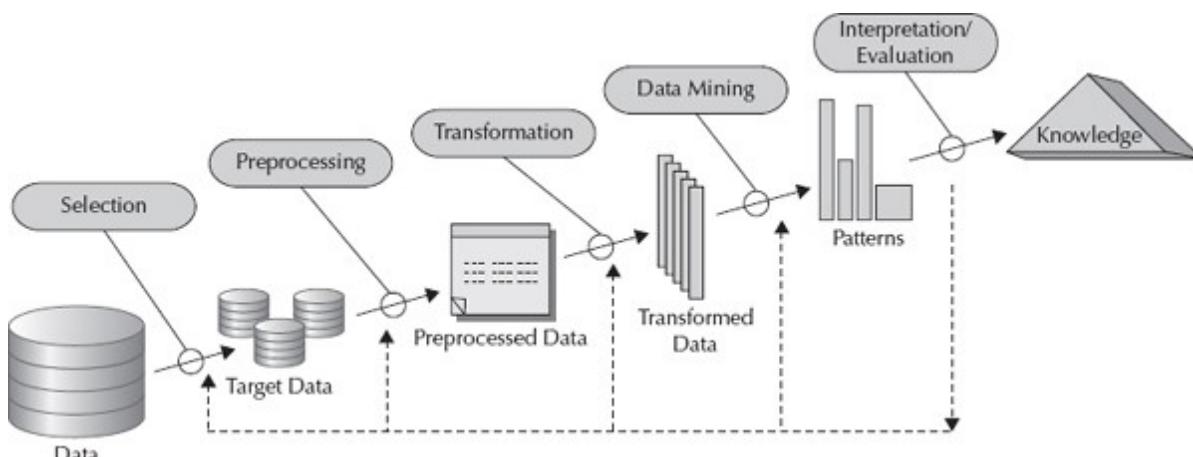


Figure 2.1: schéma des étapes du KDD, source <https://www.linkedin.com/pulse/knowledge-discovery-data-kdd-process-mohammad-valadkhani> (<https://www.linkedin.com/pulse/knowledge-discovery-data-kdd-process-mohammad-valadkhani>)

2.2 L'émergence du data-mining

L'émergence du data-mining est étroitement liée à l'augmentation du volume de données dont disposent les

entreprises. Il suscite beaucoup d'intérêt, car une réelle demande émane de secteurs très variés pour des applications toutes aussi différentes. Par ailleurs, l'amélioration de l'outil informatique, en termes de capacité de stockage, de vitesse de calcul, et des progrès méthodologiques en intelligence artificielle, statistique, base de données, ont permis de répondre en partie à ces demandes.

2.2.1 secteurs concernés

2.2.1.1 secteur bancaire

Le secteur bancaire est un des secteurs où le data-mining joue un rôle des plus importants. En effet, les banques disposent de dossiers sur leur différents clients, d'information sur le fonctionnement de leurs comptes ainsi que sur les produits bancaires auxquels ils ont souscrit. A partir de ces données, elles cherchent notamment à déterminer la probabilité qu'un client soit un bon payeur, ou qu'il soit intéressé par un service donné. C'est ce que l'on appelle le *scoring*.

2.2.1.2 grande distribution

Un autre secteur ou le data-mining est omniprésent et le secteur de la grande distribution. Ici aussi la quantité de données à disposition est importante, ne serait-ce que par l'intermédiaire des tickets de caisse et des cartes de fidélité. Les objectifs pour l'entreprise vont, par exemple, consister à identifier les règles d'association entre les achats de produits ou les séquences d'achats afin d'améliorer le rayonnage.

2.2.1.3 téléphonie

Le domaine de la téléphonie mobile fait face à une saturation du marché. Chaque année, plus d'un million d'utilisateurs changent d'opérateur. La problématique ici est plus d'évaluer la probabilité qu'un client parte chez un concurrent et lui proposer des offres en conséquence pour le conserver. On utilisera à nouveau des méthodes de scoring, mais on pourra aussi s'appuyer sur le contenu des lettres de réclamation, on parle de *text mining*.

2.2.1.4 secteur médical

Le secteur médical est très prolifique en matière de données : séquençage génomique, analyse d'examens, résultats d'imagerie, référencement de tous les actes effectués par le personnel médical. Dans ce domaine, le data-mining pourra par exemple être utilisé pour rechercher des associations entre gènes et pathologies, ou pour identifier des anomalies dans des résultats d'imagerie (on parle d'*image mining*) afin d'aider au diagnostic.

2.2.1.5 industrie automobile

Le data-mining a également investi le domaine de l'industrie automobile : à partir d'enquêtes de satisfaction, les entreprises peuvent adapter leurs produits aux demandes formulées par les consommateurs. Aussi, la classification des accidents en fonction des caractères susceptibles d'avoir provoqué celui-ci (caractéristiques du véhicule, du conducteur, état de la route, circulation, heure, météo) permet d'identifier des causes communes à certains groupes d'accidents, ce qui pourra ensuite être utilisé pour aiguiller sur les modifications à apporter sur le véhicule en vue d'améliorer la sécurité des automobilistes.

2.2.1.6 politique

Les appareils politiques ont également bien compris l'intérêt du data-mining, l'élection de Barack Obama en 2012 en est l'une des illustrations : à partir notamment des données web, on a pu identifier les attentes de certaines catégories d'électeurs susceptibles de voter démocrates. Cela a permis d'inciter les personnes à se mobiliser via, par exemple, l'envoi d'emails ciblés. On considère aujourd'hui que cette stratégie a été une des

clés du succès de l'élection.

2.2.1.7 autres

La liste n'est pas exhaustive, le data-mining répond à beaucoup d'autres problématiques que ce soit dans l'industrie agroalimentaire, dans le domaine de la sécurité, la biologie,...

2.2.2 les facteurs de l'émergence

Les raisons de l'émergence du data-mining dans la société s'explique par des facteurs à la fois conjoncturels et techniques.

En effet, si le data-mining est présent dans de nombreux secteurs d'activité c'est tout d'abord parce qu'il répond à des besoins nouveaux. La mondialisation favorisant la concurrence à grande échelle, la rentabilité des entreprises passe par l'optimisation des ressources dont elle dispose, notamment l'exploitation de ses données, mais aussi par une automatisation des tâches pour diminuer les coûts de production. Par exemple, la qualité des produits sur une chaîne peut être automatiquement et rapidement évaluée à partir d'image du produit. La mondialisation n'est néanmoins pas la seule raison à l'émergence du data-mining, des contraintes réglementaires ont aussi poussé les entreprises à y recourir. Les banques, par exemple, ont une obligation législative de solidité financière (Tufféry (2007), pp 403-404). Cela les oblige notamment à estimer précisément les risques encourus avant de prêter une somme d'argent de façon à pouvoir ajuster le crédit en fonction. Les entreprises pharmaceutiques ont quant à elles le devoir d'effectuer des tests avant de commercialiser des médicaments, et d'étudier, une fois la commercialisation effectuée, des éventuels effets à long terme qui n'auraient pas pu être étudiés avant commercialisation. Enfin, la consommation étant le moteur des entreprises, celles-ci ont le besoin d'identifier les nouvelles attentes des consommateurs afin d'y répondre et d'assurer leur pérennité.

Mais le data-mining n'aurait pas pu apporter de réponses à ces besoins sans l'émergence de moyens techniques, que ce soit d'un point de vue informatique ou méthodologique. En effet, la matière première du data-mining est la donnée qu'il faut pouvoir stocker et utiliser à des fins d'analyse. Les capacités de stockage de la donnée, et l'amélioration de la puissance de calcul des machines pour analyser ces données ont favorisé grandement l'émergence du data-mining. Les entreprises possèdent alors des données à profusion, sans que celles-ci soient réellement exploitées. Le coût de stockage est en effet faible et l'automatisation de la saisie (code barre, click web, cartes de crédit,...) permet d'accumuler la donnée. Par ailleurs, des données en libre accès (*open data*) sont aussi à disposition des entreprises et favorisent potentiellement l'enrichissement des connaissances. Néanmoins, les bases de données ne sont généralement pas optimisées pour l'analyse mais simplement pour le stockage. Elles permettent surtout d'accéder à des informations brutes contenues dans la base, mais pas d'aider à la décision. Afin d'analyser les données, des *entrepôts* de données (Data Warehouse) se sont développés. Ceux-ci permettent de centraliser toutes les données d'une entreprise et d'y accéder efficacement en vue d'être fouillées sous forme de *Data Marts* ou *base de données métiers*. Elles y sont organisées par thèmes (ventes, comptabilité, marketing,...), présentées sous un format uniifié à l'utilisateur, elles ne sont pas modifiables et sont historisées afin de disposer d'information évolutive (par exemple : évolution du nombre de produits achetés par un client au cours du temps). Les données accumulées sont très particulières, d'une part elles sont très volumineuses, mais elles contiennent aussi des valeurs manquantes, des valeurs aberrantes et sont de nature variées et nouvelles (image, vidéo, son).

Ainsi, le développement de méthodes d'analyse nouvelles, qu'elles soient statistiques, algorithmiques, d'intelligence artificielle, ou de construction des bases de données, a également favorisé l'émergence du data-mining. Finalement, ces améliorations techniques, à la fois computationnelles et méthodologiques, ont pu être diffusées via le développement de logiciels.

2.3 Une évolution au cours du temps

Les technologies et la société évoluant, les données à traiter ont progressivement changées, et des problématiques méthodologiques nouvelles sont apparues.

2.3.1 Des données de plus en plus volumineuses

La quantité de données générées est de plus importante, du fait principalement du développement des technologies du numérique (cf figure 2.2). Aujourd’hui, cette quantité est située aux alentours de 20 zettabytes (1 zettabit = 10^{21} bytes) et continue d’augmenter. La quantité de données à analyser devient colossale, on estime que la NSA ou le FBI dispose de plusieurs yottabytes (1 yottabyte = 10^{24} bytes) de données sur les individus du globe.

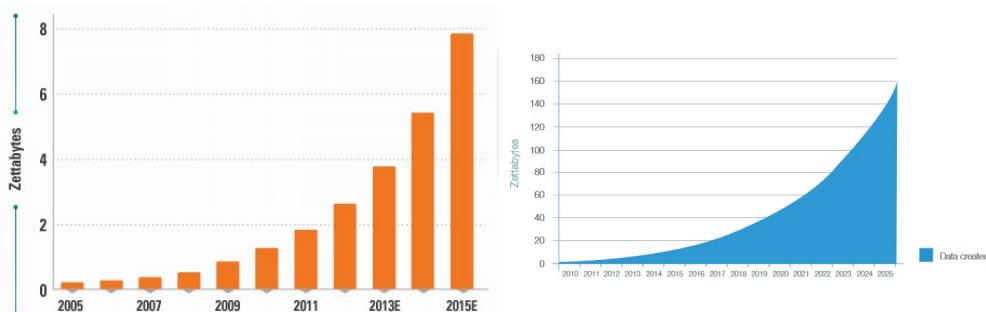


Figure 2.2: Evolution de la quantité de données numériques générée. A gauche, entre 2005 et 2015 (source KPCB, IDC) ; à droite entre 2010 et 2025 (source IDC's Data Age 2025 study, sponsored by Seagate, April 2017)

2.3.2 Les âges de la statistique

2.3.2.1 avant 1970 : la statistique classique

Avant les années 1970, la donnée est assez rare et chère. En effet, l’ordinateur n’est pas encore démocratisé et les capacités de stockage restent limitées. Les données sont récoltées pour répondre à une question précise définie en amont. Les individus statistiques sont choisis avec soin selon des plans d’échantillonnage et n’excèdent généralement pas la centaine. Les variables sont également bien choisies pour répondre à la question posée et n’excèdent pas la dizaine. Ainsi les jeux de données représentent une centaine d’octets (1 octet = 1 byte) et sont de bonne qualité (peu de données manquantes, peu d’hétérogénéité par exemple).

2.3.2.2 années 1970-1990 : analyse des données

Après 1970, l’ordinateur se démocratise : moins cher et plus performant, il permet de stocker et d’analyser des jeux de données de l’ordre du Mo ou Go représentant quelques milliers d’individus et plusieurs dizaines de variables. Il n’est plus possible d’explorer le jeu de données de façon exhaustive, d’où la nécessité de trouver des moyens synthétiques de le visualiser. Les méthodes d’analyse des données (Analyse en composantes principales, analyse des correspondances multiples, etc) ont alors connu des avancées importantes.

2.3.2.3 années 1990-2010 : data-mining

Les années 1990 correspondent à la naissance du web tel qu’on le connaît aujourd’hui. La production de la donnée devient plus facile. Par ailleurs, les capacités des ordinateurs continuant à évoluer, on s’intéresse alors à des jeux de données comportant des millions d’individus et des centaines, voire des milliers de variables, représentant des données de l’ordre du Go ou To. Cette démocratisation fait aussi venir la

statistique dans le monde de l'entreprise. Le paradigme statistique change : les tailles des échantillons deviennent trop grandes pour appliquer des tests car leur puissance est trop importante (*i.e.* que les hypothèses nulles sont systématiquement rejetées). Cette taille impose aussi des méthodes avec des coûts algorithmiques faibles pour pouvoir être appliquées. Aussi, les données sont désormais créées sans qu'un objectif d'analyse ne soit fixé à l'avance, on parle d'analyse secondaire. Les échantillons ne sont plus représentatifs, les données deviennent *déséquilibrées*, d'où la nécessité de corriger les analyses pour en tenir compte. Par exemple, une des applications du data-mining à EDF est de prédire le mode de chauffage principal des clients, mais les clients les mieux connus d'EDF sont ceux qui se chauffent à l'électrique. Il y a donc sur-représentation de ce type de chauffage dans l'échantillon (Hébrail and Lechevallier (2003)). Les données sont également d'origine *multimédia* (images, textes,...), posant la question non seulement de l'analyse de chacun des types de données (image mining, text mining, ...), mais aussi de leur analyse simultanée qui demande de gérer l'*hétérogénéité* de ces données. Enfin, ces données deviennent de qualité moindre (incomplètes ou incohérentes), impliquant de développer de nouvelles méthodes d'analyse adaptées. Nous aborderons certaines d'entre elles dans ce cours ()

2.3.2.4 depuis 2010 : machine-learning Big data

Depuis dix ans, une nouvelle ère a débutée, celle du *Big data*. La quantité de données à disposition a explosé, les jeux à analyser peuvent être de l'ordre du To (10^{12} bytes), Po (10^{15} bytes) ou plus encore. Par exemple, les données dont dispose Facebook sont de l'ordre du To. Ces données sont parfois obtenues en temps réel, de sources différentes et deviennent ainsi *massives*, on parle de *Big data*. Le Big data est devenu un phénomène de société car nombreuses sont les entreprises qui y voient un intérêt industriel majeur.

La définition des données massives n'est pas claire, mais on les caractérise souvent par les "3V" (Volume, Vélocité, Variété) voire les "5V" (Valeur, Véracité), qualifiant les différents éléments à prendre en compte pour les analyser. Le *volume* correspond à la difficulté de stockage de la donnée, en particulier il devient impossible de la stocker en mémoire vive, voire même sur un seul ordinateur. La *vélocité* fait référence à la vitesse à laquelle les données sont produites, partagées, et mises à jour. La *vélocité* nécessite une analyse en temps réel et pose le problème du stockage de la donnée. C'est par exemple le cas des moteurs de recherche qui doivent proposer des résultats actualisés alors que la quantité d'information sur internet est sans cesse grandissante. La *variété* correspond à l'hétérogénéité des données, celles-ci pouvant être de nature diverse (textes, images, son, vidéos) et provenir de sources différentes (données personnelles, open data, etc). Par ailleurs, celles-ci ne sont pas nécessairement *structurées* comme peut l'être une base de données, il devient alors difficile d'associer ces données entre-elles. Par exemple un texte sans mots clés peut être considéré comme une donnée *non structurée*. Il sera *semi structuré* en lui associant des mots-clés (résumant les sujets abordés dans celui-ci par exemple). La *véracité* fait référence au fait que les données n'ont pas été collectées par la personne analysant les données, et qu'en conséquence il est difficile d'en connaître la qualité. Certaines données peuvent être simplement déclaratives, ou provenir du web et avoir été générées par des robots, ou simplement être de fausses informations diffusées sur la toile (notamment via les réseaux sociaux). Enfin, la *valeur* est relative au fait que toutes les données ne peuvent pas être utilisées pour pouvoir répondre à la question d'intérêt. Il faut alors identifier les données actionnables, *i.e.* celles qui permettront potentiellement de prendre des décisions rapidement, sans intervention humaine.

L'analyse des données massives nécessite des méthodes et d'outils informatiques spécifiques. En particulier, les données sont réparties sur différents clusters ou sur un cloud. Les méthodes d'analyse doivent prendre en compte cette dispersion des données. Elles doivent aussi être parallélisables, ou suffisamment peu complexes algorithmiquement de façon à pouvoir être mises en oeuvre sur de tels jeux.

Le Big data fait à nouveau évolué le paradigme statistique, avec un besoin moindre en termes d'interprétation, mais accru en termes de prédition. Dans cette optique, les méthodes de machine learning

sont très appréciées. En particulier, l'analyse des données massives s'appuie sur les développements en intelligence artificielle. Par exemple, les voitures autonomes doivent analyser en temps réel les données présentent dans leur environnement pour prendre des décisions rapidement. Toute cette technologie peut être réinvestie dans l'analyse des données massives, et réciproquement.

3 Effectuer une étude de data-mining

3.1 Le data-mining étape par étape

3.1.1 Comprendre et analyser les objectifs de l'étude

La première étape consiste à définir les objectifs, la population étudiée et l'unité statistique (par exemple les clients fidèles d'une société, les fumeurs d'une région, les familles monoparentales françaises,...), le critère à prédire, et l'application attendue. Cette étape nécessite de réunir à la fois la maîtrise d'ouvrage (marketing, experts, utilisateurs) et la maîtrise d'oeuvre (statisticiens, informaticiens). Elle peut déjà définir le choix des outils qui seront utilisés. Par exemple, on n'utilisera pas des modèles complexes si l'interprétation des résultats est essentielle.

3.1.2 Création d'une base de données

La création de la base commence par l'inventaire des données existantes. Celles-ci doivent être accessibles, légalement exploitables, fiables, mises à jour, suffisamment précises. Elles peuvent provenir de l'entreprise, être libres, être achetées à l'extérieur, etc. Si besoin, l'entreprise pourra les récolter elle-même à partir d'échantillons de clients (en leur proposant des cadeaux s'ils remplissent un questionnaire). Il sera aussi nécessaire de disposer d'un historique du phénomène à prédire afin de valider les modèles employés. Par exemple, si on s'intéresse à prédire le comportement des clients au mailing (savoir si le client ouvre le mail ou non), on pourra utiliser des campagnes commerciales antérieures et comparer le comportement des clients observés à celui prédit par le modèle.

Toutes ces données sont stockées dans un *entrepôt de données* (*Data warehouse*) afin d'être exploitées efficacement (cf Figure 3.1). L'entrepôt fournit des données normalisées, homogénéisées de façon à être compréhensibles par tous. En effet, les données avant normalisation peuvent être définies au niveau du produit (une ligne par produit) plutôt qu'au niveau de l'individu. Il faut alors remettre en forme ces données pour avoir tous les produits achetés par un même individu sur une seule ligne. Les données de l'entrepôt ont la propriété d'être *non-volatiles*, c'est-à-dire qu'elles ne sont jamais écrasées par de nouvelles données plus récentes. Au contraire, elles sont affectées d'une date ce qui permet de disposer de séries temporelles (par exemple, nombre de produits achetés par un client chaque jour de l'année).

Construire une Infrastructure d'Information Intelligente pour l'Entreprise

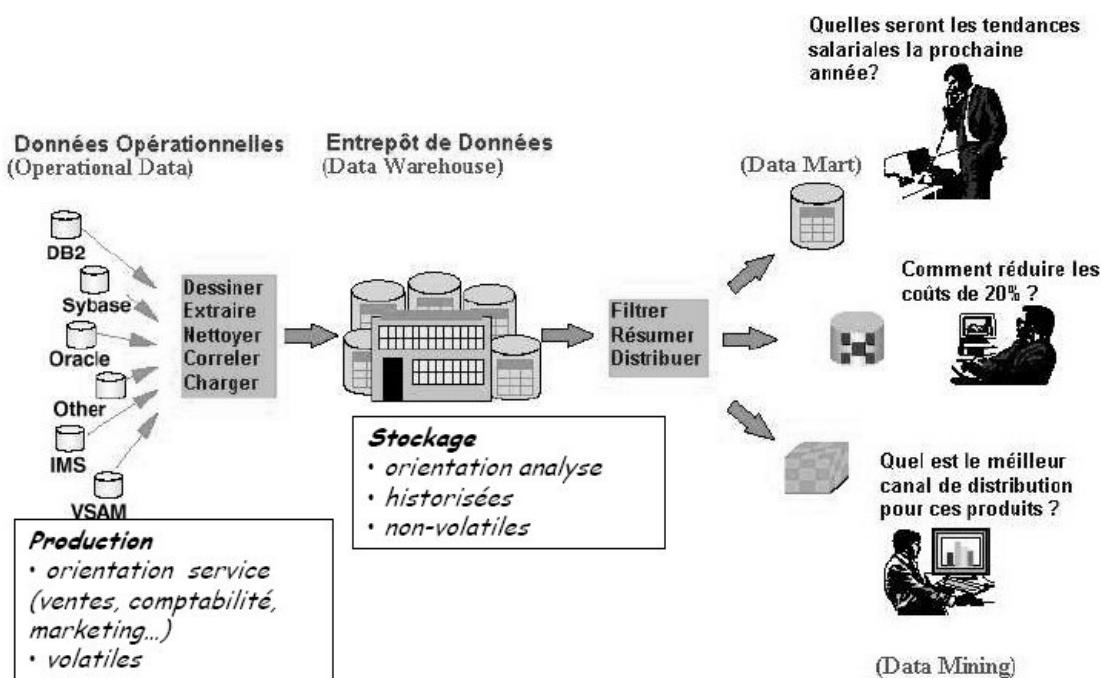


Figure 3.1: Schéma d'une vue d'ensemble d'une architecture d'entrepôt de données

3.1.3 Prétraitement et nettoyage des données

Les données collectées ne sont pas encore exploitables. En effet, il est très fréquent que celles-ci soient incomplètes, contiennent des valeurs erronées (e.g. taille négative). Il convient donc de fiabiliser les données incorrectes. Si peu d'individus ont une anomalie sur une même variable, on pourra soit écarter les individus problématiques, soit remplacer ces valeurs par d'autres plus vraisemblables en utilisant des méthodes dites d'imputation. Si en revanche le nombre de données incorrectes est trop grand, on pourra dans un premier temps recoder la variable sous la forme d'une indicatrice du caractère incorrect de la donnée (0 pour correct, 1 pour incorrect) et voir si cette indicatrice ressort comme importante dans l'analyse. Si ce n'est pas le cas, on pourra décider dans un second temps de ne pas en tenir compte.

D'autre part, certaines variables peuvent être dans des formats peu pertinents. Par exemple, on pourra souhaiter modifier les unités de certaines, recoder des modalités par des noms plus synthétiques, ou recoder des dates par des durées. On pourra aussi ajouter des indicateurs couramment utilisés dans le domaine de l'étude, c'est souvent le cas de certains scores (par exemple l'IGS2 qui, dans le domaine médical, rend compte de l'état de gravité d'un patient et se calcule à partir d'autres mesures comme la pression artérielle, la fréquence cardiaque, etc).

3.1.4 Réduction des données

Il y a plusieurs intérêts à réduire la dimension des données, que ce soit en considérant un sous-ensemble des variables, des individus ou en diminuant le nombre de modalités des variables qualitatives.

Généralement, le nombre de paramètres des méthodes de data-mining employées est fonction du nombre de variables. Il convient alors de réduire leur nombre pour pouvoir estimer ``convenablement'' les paramètres de la méthode (au sens du compromis biais-variance qui sera rappelé ici ()). Par ailleurs, des méthodes très classiques ne peuvent pas être appliquées en présence de fortes colinéarités entre les variables explicatives

(régression logistique, analyse discriminante, régression linéaire). Il apparaît donc nécessaire d'ignorer certaines variables trop corrélées aux autres, ou de fusionner des variables corrélées entre elles. Les méthodes d'analyse factorielle, telles que l'analyse en composantes principales (ACP) seront des outils utiles pour repérer des corrélations trop fortes ou pour effectuer ces fusions. Notons que la colinéarité peut aussi se gérer via des méthodes telles que la régularisation que nous aborderons dans la partie dédiée aux méthodes avancées ()).

Par ailleurs, de nombreuses méthodes sont peu pertinentes en présence de modalités rares, ou d'un grand nombre de modalités. Il est donc nécessaire de fusionner ces modalités entre elles. Par exemple pour une variable dont les modalités sont ordonnées (e.g. "pas d'accord", "plutôt pas d'accord", "plutôt d'accord" et "d'accord"), on pourra rassembler deux modalités successives si l'une d'entre elle est de faible effectif (e.g. fusionner "pas d'accord" et "plutôt pas d'accord" en "avis négatif"). On pourra aussi écarter des variables quand elles sont une version discrétisée d'une variable quantitative déjà disponible (soit en écartant la variable originale, soit en écartant la variable discrétisée).

Enfin, les coûts algorithmiques de certaines méthodes peuvent nécessiter de considérer des échantillons plus petits. Dans ce cas, on pourra échantillonner les données, soit via un tirage aléatoire simple, soit selon un plan plus complexe assurant une représentativité.

3.1.5 Segmenter la population

La présence de groupes d'individus hétérogènes doit être prise en compte par les méthodes d'analyse qui seront appliquées sur les données. Soit on utilisera des méthodes qui modélisent cette hétérogénéité (e.g. modèles de mélanges), soit on appliquera un modèle spécifique à chaque segment d'individus homogènes.

Différentes façons de segmenter sont envisageables. Une première façon est d'effectuer une segmentation non supervisée, i.e. sans tenir compte de la variable que l'on cherchera à expliquer par la suite. Par exemple, on peut classer les clients d'une banque en fonction de leurs CSP, sans tenir compte de la probabilité qu'un individu rembourse son crédit. Ces classifications s'appuient généralement sur une expertise du domaine, mais peuvent aussi être effectuées selon des approches purement statistiques.

Par opposition, on peut aussi utiliser des méthodes supervisées. Par exemple, un arbre de décision à un ou deux niveaux pourra être utilisé dans ce but : il s'agit d'identifier une variable permettant de séparer les individus en deux groupes de façon à ce que la variable à prédire soit la plus homogène possible au sein de chaque groupe. Par exemple, en segmentant les conducteurs selon leur région, on peut bien séparer ceux qui risquent d'avoir plus d'accidents de voiture (les régions urbaines) que les autres (régions rurales). On peut ensuite segmenter ces deux groupes indépendamment selon une autre variable, ce qui amène à une segmentation en 4 classes.

Quel que soit la méthode utilisée, celle-ci doit être simple (facile à justifier), avec un nombre raisonnable de classes.

3.1.6 Choix et validation du modèle

Le choix du modèle d'analyse est le problème central, bien qu'il ne soit pas nécessairement le plus long. Les familles de méthodes envisageables sont en partie dictées par la nature des variables considérées (cf Section 3.2.2), mais aussi par les objectifs de l'étude (besoin de modèles facile à interpréter par exemple). Néanmoins, au sein de ses familles, plusieurs choix restent possibles. Par exemple, pour un modèle de régression linéaire, il est nécessaire de choisir les variables explicatives. Il existe généralement des moyens classiques pour sélectionner le meilleur modèle pour une famille donnée. En présence de plusieurs familles on pourra comparer directement les capacités de prédiction des modèles en utilisant des données tests, i.e. des

données n'ayant pas servies à la construction de ces modèles. Le choix et la validation de modèle seront traités en détails ici () .

3.1.7 Itérations du processus

Ce processus étant séquentiel, les choix aux différents niveaux ont des répercussions sur les résultats finaux. Il sera généralement nécessaire d'itérer ce processus depuis le prétraitement jusqu'à ce que les résultats en sortie soient satisfaisants. Par exemple, on peut finalement s'apercevoir sur le modèle final que certains individus ont des profils atypiques qui n'avaient pas été détectés auparavant, on pourra alors souhaiter recommencer l'analyse en les écartant ou en s'en accommodant via l'utilisation d'une autre méthode d'analyse robuste.

3.1.8 Déploiement des modèles

Une fois le modèle établi, il faut implémenter informatiquement la méthode afin de la rendre accessible auprès des utilisateurs potentiels qui s'appuieront dessus pour prendre des décisions. Dans un premier temps, les sorties des modèles peuvent être mises à disposition dans des tableurs installés sur chaque poste de travail. Ainsi, si l'outil de data-mining a été développé dans le but d'identifier de potentiels clients, les employés de la société peuvent contacter ces individus de façon privilégiée. Dans un second temps, il faudra rendre disponible des données actualisées auprès des utilisateurs de façon automatisée. Enfin, il conviendra de faire évoluer l'application en fonction des retours des utilisateurs avant de la développer dans tout le réseau de l'entreprise.

3.2 Les outils du data-mining

Le data-mining vise à identifier des structures au sein de données. On distingue généralement deux types de structures : *les modèles* et *les patterns*. Les modèles visent à expliquer ou prédire une variable (appelée variable réponse, définie a priori) à partir d'autres variables (appelées variables explicatives). On s'intéresse donc à des associations communes à l'ensemble des individus. Par exemple, on peut souhaiter comprendre ce qui déclenche l'acte d'achat d'un client. Les patterns eux sont des associations qui concernent juste quelques individus. On ne les connaît pas a priori. Par exemple, un pattern peut être que les clients qui achètent du caviar achètent aussi de la vodka.

En fonction du type de structures recherchées, on aura recours à des méthodes différentes : les approches *supervisées* chercheront à établir un lien entre une variable réponse et d'autres variables, elles seront plus adaptées à la recherche de modèles. Au contraire, les approches *non-supervisées* chercheront à mettre en évidence des associations fortes entre certaines variables pour certains individus, sans qu'aucune variable ne joue un rôle privilégié a priori, elles seront donc plus adaptées à la recherche de patterns.

Ces différentes méthodes seront décrites spécifiquement dans la suite de ce cours.

3.2.1 Méthodes non-supervisées

Parmi les méthodes non-supervisées, on retrouve essentiellement les méthodes d'analyse factorielle, les méthodes de classification et les méthodes de recherche de règles d'association.

Les méthodes d'analyse factorielle sont des approches géométriques qui consistent à résumer les données sous la forme de graphiques de dimension deux. La visualisation de ces graphiques synthétiques permet alors d'identifier les ressemblances entre les individus, les relations entre les variables et de résumer des groupes d'individus qui se ressemblent par quelques variables caractéristiques. On distingue différentes méthodes

d'analyse factorielle en fonction de la nature et du nombre de variables considérés : analyse en composantes principales pour les données quantitatives, analyse factorielle des correspondances pour deux variables qualitatives, analyse des correspondances multiple pour plusieurs variables qualitatives pour les plus populaires.

Les méthodes de classification consistent à identifier des groupes d'individus homogènes. On peut distinguer deux grandes familles de méthodes : celles basées sur des modèles et celles basées sur des distances. Les méthodes basées sur des modèles consistent à faire l'hypothèse que chaque individu appartient à une classe et que dans chacune de ces classes, la distribution des variables est spécifique. Ainsi, une fois les différentes distributions identifiées, on pourra établir la classe la plus probable pour chacun des individus parmi les k (k défini à l'avance) classes. Les méthodes de classification basées sur les distances sont quant à elle des approches géométriques. On distingue les méthodes de partitionnement et les méthodes hiérarchiques. Les méthodes de partitionnement consistent à définir une partition de l'ensemble des individus en k (défini à l'avance) groupes telle que la dispersion des individus (au sens de la distance préalablement définie) au sein (resp. extérieur) des groupes soit la plus faible (resp. forte) possible. Elles regroupent les méthodes des centres mobiles et ses variantes (k-means, nuées dynamiques) et les cartes de Kohonen pour les méthodes les plus connues. Les méthodes hiérarchiques visent quant à elle à construire des classes imbriquées les unes dans les autres. On distingue les méthodes ascendantes, qui partent d'une partition très fine puis qui fusionnent les classes les plus voisines (au sens de la distance choisie) de proche en proche, des méthodes descendantes, qui partent, elles, d'une classe regroupant l'ensemble des individus et qui la partitionnent de façon itérative de façon à obtenir des classes plus fines.

Les méthodes de recherche de règles d'association, globalement moins répandues que les méthodes précédentes, sont très utilisées dans le domaine du marketing. Elles consistent à comparer, pour certains produits, les probabilités d'achats simultanés (e.g. achat de pain et achat de fromage) aux probabilités obtenues dans le cas où il n'y aurait pas de lien entre les achats des produits. Ceci permet d'identifier des associations importantes. La difficulté étant que quand le nombre de produits est grand, on génère de nombreuses règles et qu'il devient difficile d'extraire les règles pertinentes non triviales ou simplement déjà connues. Ce type de méthodes est aussi utilisé dans les systèmes de recommandation sur internet. Par exemple, ayant préalablement identifié des règles d'association entre des achats de livres via l'analyse des flux de clics (*webmining*), si une personne a acheté un premier polar sur internet, alors lors de sa reconnexion sur un site de vente de livre, on va pouvoir lui conseiller un autre livre du même genre apprécié par d'autres lecteurs amateurs de polars.

Les méthodes non-supervisées seront abordées plus précisément ici () .

3.2.2 Méthodes supervisées

Parmi les méthodes supervisées, on distingue généralement les cas où la variable à expliquer (ou à prédire) est quantitative (on parle de problème de régression) aux cas où cette variable est qualitative (on parle de problème de classement ou de classification supervisée). La raison étant que certaines méthodes supervisées ne sont dédiées qu'à un type de variable réponse particulier.

3.2.2.1 Réponse quantitative

La méthode la plus classique en présence d'une variable réponse quantitative est le modèle linéaire. Il peut être utilisé en présence d'une ou plusieurs variables explicatives. En fonction de la nature des variables explicatives, on parlera de régression linéaire (variables quantitatives), d'Anova (variables qualitatives), ou d'Ancova (variables quantitatives et qualitatives). Cette méthode est dite paramétrique car elle fixe, à travers le choix d'une fonction (appelée *fonction de lien*), la nature du lien entre les variables explicatives et la

variable à expliquer (ici lien linéaire). Elle est commode pour l'interprétation du lien entre variables explicatives et variables à expliquer. Remarquons qu'il existe aussi des méthodes de régression non-paramétriques, où la nature du lien est dictée par les données elles-mêmes (e.g. méthodes à noyau, polynômes locaux).

Une autre méthode très utilisée est l'arbre de régression. Cette méthode peut être appliquée quelle que soit la nature et le nombre de variables explicatives. La différence majeure par rapport au modèle linéaire est que cette méthode est non-paramétrique, le lien entre les variables à expliquer et explicative n'est pas défini a priori.

Parmi les autres méthodes non-paramétriques, on peut citer les réseaux neuronaux (qui seront présentés en détails ici ()), notamment le perceptron multicouches. Cette méthode peut aussi être appliquée dans le cas où la variable réponse est qualitative.

3.2.2.2 Réponse qualitative

Quand la variable réponse est qualitative, une méthode très couramment utilisée est la régression logistique. C'est une approche paramétrique. Comme le modèle linéaire, elle peut être utilisée quelle que soit la nature des variables explicatives (régression logistique et régression linéaire sont deux cas particuliers du modèle linéaire généralisé). Notons que la méthode de base est dédiée à une variable réponse binaire (à deux modalités). Les modèles logit ordonnés et multinomiaux sont les équivalents pour une variable à plusieurs modalités (ordonnées ou non respectivement).

L'analyse discriminante linéaire est une méthode très proche de la régression logistique, à la différence que celle-ci fait une hypothèse supplémentaire sur la distribution des variables explicatives. La différence repose essentiellement sur la technique d'estimation des paramètres de la méthode. On pourra la privilégier quand les classes à prédire sont bien séparées (cas où la régression logistique est mise en défaut).

D'un point de vue géométrique, ces deux méthodes consistent à rechercher un hyperplan qui sépare au mieux les classes. Quand la séparation est plus complexe, on pourra utiliser les ``support vector machines'' (SVM). Néanmoins, l'interprétation de la séparation par rapport aux variables explicatives devient difficilement interprétable.

Une autre méthode très usitée est l'arbre de classification. A l'image des arbres de régression, il s'agit d'une méthode supervisée qui peut être employée pour des variables explicatives de nature diverse. Rappelons que les réseaux neuronaux constituent aussi une méthode non-paramétrique à considérer quand la variable réponse est qualitative.

Nous aborderons plus précisément les méthodes supervisées ici ().

3.2.3 Méthodes avancées

Les méthodes précédemment évoquées restent des méthodes très classiques, non dénuées de défauts. En particulier, elles peuvent être instables, c'est-à-dire sensibles aux données qui ont servi à estimer les paramètres des modèles. Ainsi, on pourra avoir recours à des méthodes plus avancées, comme les *méthodes d'ensembles*. Ces méthodes consistent à construire plusieurs modèles puis à agréger ces modèles entre eux. Par exemple, les forêts aléatoires constituent une méthode d'ensemble pour des arbres de décision (de régression ou de classification). Cette approche améliore très significativement les capacités de prédiction des arbres (mais elles sont plus difficiles à utiliser en termes d'interprétation).

Une autre approche pour diminuer les problèmes d'instabilité sont les approches *ridges*, essentiellement utilisées dans les modèles linéaires généralisés. En effet, des problèmes d'instabilité apparaissent dans ces

modèles quand les liaisons entre les variables explicatives sont trop fortes, ou quand le nombre d'individus est trop petit devant le nombre de variables. Les approches ridge corrigeant cette instabilité.

Une autre source d'instabilité peut provenir des valeurs aberrantes. Celles-ci sont fréquentes dès que la quantité de données est grande. Des méthodes de statistiques dites *robustes* ont été développées pour faire face à cette difficulté. A nouveau dans le cadre du modèle linaire, ces méthodes vont consister à ajuster le modèle sur les individus les moins extrêmes, ou à choisir un critère de minimisation moins sensible aux valeurs extrêmes.

Citons enfin les méthodes *parcimonieuses* ("sparse" en anglais) qui sont appropriées quand toute l'information à disposition n'est pas pertinente pour le modèle choisi. Par exemple, en régression, quand le nombre de variables est grand (plusieurs centaines), il est légitime de penser que seule une partie des variables explicatives ont un lien avec la variable à expliquer. De la même façon, en ACP, on peut supposer que certaines variables sont très peu liées aux autres et qu'il n'est pas nécessaire de les considérer.

Par la suite, différentes méthodes avancées seront présentées en détails ici () .

3.2.4 Des logiciels dédiés

Nombreux sont les logiciels à disposition pour effectuer une étude de data-mining (cf Table 3.1). Ils diffèrent par l'offre des méthodes d'analyse (certains sont simplement mono-techniques), les capacités à gérer de gros fichiers, la capacité à automatiser les tâches courantes, leur prix, etc. Ces logiciels sont généralement disponibles en version micro, i.e. avec un mode de fonctionnement local, mais c'est davantage les versions client-serveur qui permettront de travailler sur de grandes bases de données.

Table 3.1: Différents logiciels de data-mining

Logiciel micro	Logiciels client-serveur
Logiciels multi-techniques	
Insight - S-PLUS	R/Rstudio serveur
R/Rstudio	SAS/STAT
Weka	SAS - Entrepise Miner
TANAGRA	SPSS
Orange	SPSS - Clementine
	SPAD
	Statsoft - Statistica Data Miner
	Insightful - Insightful Miner
	KXEN - SAP
	IBM- intelligent Miner

Logiciel micro	Logiciels client-serveur
	Oracle - Oracle Data Mining
	Microsoft - Analysis Services
	Spark
	Hadoop
Logiciels mono-techniques	
Salford Systems - CART	Isoft - Alice
Neuralware - Predict	SPSS - Answer Tree
Complex Systems - DATALAB	

Parmi eux, R, TANAGRA, Weka, Orange, Spark, Hadoop sont libres.

Selon l'étude Rexter Analytics 2011 Data Miner, R apparaît comme le logiciel qui compte le plus grand nombre d'utilisateurs parmi la communauté des ``data miners'', mais c'est le logiciel Statistica qui apparaît comme le plus fréquemment utilisé en première utilisation pour effectuer du data-mining, notamment dans le secteur de l'entreprise (*cf* Figure 3.2).

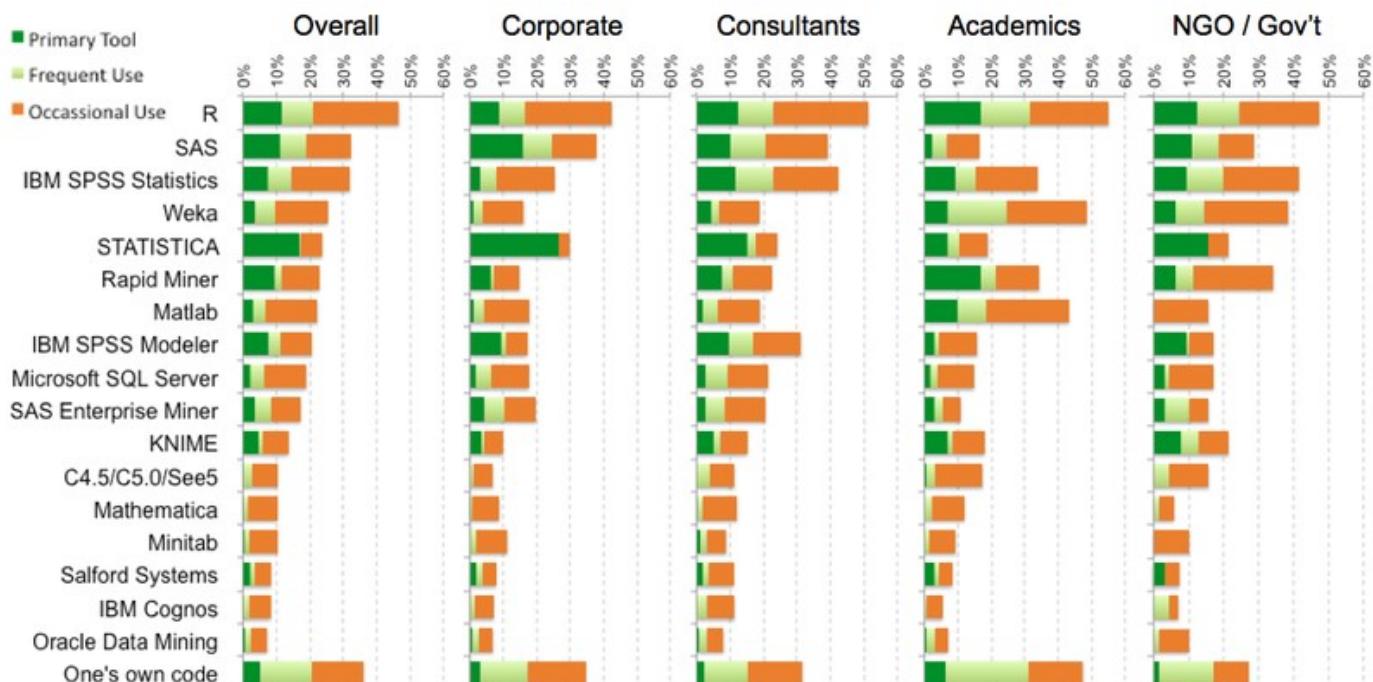


Figure 3.2: Logiciels de data-mining les plus utilisés selon les domaines (tous domaines confondus, domaine de l'entreprise, du consulting, du secteur académique, utilisation gouvernementale). source : Rexter Analytics 2011 Data Miner Survey

4 Conclusion

La fouille des données est en plein essor depuis une vingtaine d'années. Ses applications sont déjà

importantes dans de nombreux domaines d'activité, mais la fouille de données reste en perpétuelle évolution, les méthodes employées profitant des avancées technologiques et suivant les évolutions sociétales. Elles tentent alors de répondre perpétuellement à de nouvelles attentes.

La fouille de données a largement modifié la façon de traiter ces dernières. On s'intéresse désormais de plus en plus à des analyses secondaires. La prévision, en vue de prendre des décisions, passe alors au premier plan devant l'interprétation. Les méthodes de machine learning complètent alors la gamme des outils d'analyse à disposition. Les données traitées sont également d'un type nouveau (vidéos, images, textes, données symboliques, données multimédia) et ne se présentent plus nécessairement sous la forme d'un simple tableau *individus × variables*.

Toutefois, il faut reconnaître que beaucoup de méthodes employées en data-mining ont des origines anciennes. Jean-Paul Benzécri, père de l'analyse des données disait déjà dans les années 1970 que *l'analyse des données est un outil pour dégager de la gangue des données le pur diamant de la véridique nature*. Les réseaux de neurones, très populaires aujourd'hui, sont eux antérieurs aux années 1980. De ce point de vue, le data-mining est aussi la remise au goût du jour de certaines méthodes anciennes, et constitue en ce sens une sorte de mode, utilisée à profit par certaines entreprises pour vendre des logiciels (Friedman (1997)).

Doit-on toujours avoir pour objectif d'utiliser toutes les données à disposition, et donc analyser des données de plus en plus massives ? Pas nécessairement. Au fur et à mesure que la quantité de donnée croît, il devient de plus en plus difficile de maîtriser les données que l'on utilise. Celles-ci peuvent être de très mauvaise qualité, ou de valeurs aberrantes. Par ailleurs, la multiplicité des données rend inévitable l'identification des patterns fortuits, qui n'ont en réalité aucun intérêt, car purement liés au hasard. Aussi, même s'il était possible d'analyser des données de n'importe quelles tailles, le data-mining ne serait pas l'alternative à des approches statistiques plus classiques. En effet, il est par exemple difficile d'identifier des relations de cause à effet à partir des analyses secondaires faites en data-mining, seules des associations peuvent être clairement identifiées. Or, les problématiques de causalité sont très importantes dans notre société. L'interdiction de commercialiser certains produits repose essentiellement sur la causalité entre son utilisation et des conséquences néfastes que l'on peut lui attribuer. De la même façon, la prévision n'est pas toujours suffisante et l'interprétation sera parfois indispensable. Par exemple, on n'imagine pas un médecin annoncer son devenir à un patient sans lui fournir d'explication.

Ainsi, le data-mining, bien que devenu incontournable aujourd'hui, reste une approche complémentaire à d'autres approches plus traditionnelles.

Références

Adriaans, P., and D. Zantinge. 1996. *Data Mining*. Addison-Wesley.

Berry, M. J., and G. Linoff. 1997. *Data Mining Techniques: For Marketing, Sales, and Customer Support*. New York, NY, USA: John Wiley & Sons, Inc.

Fayyad, Usama M., Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. "Advances in Knowledge Discovery and Data Mining." In, edited by Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, 1–34. Menlo Park, CA, USA: American Association for Artificial Intelligence.

Friedman, Jerome H. 1997. "Data Mining and Statistics: What's the Connection?" In *Keynote Address, 29th Symposium on the Interface: Computing Science and Statistics*.

Hand, D.J. 1999. "Why Data Mining Is More Than Statistics Writ Large." In *Bulletin of the International Statistical Institute, 52nd Session*, 433–36.

- Hand, David J. 2000. "Methodological Issues in Data Mining." In *COMPSTAT: Proceedings in Computational Statistics 14th Symposium Held in Utrecht, the Netherlands, 2000*, edited by Jelke G. Bethlehem and Peter G. M. van der Heijden, 77–85. Heidelberg: Physica-Verlag HD.
- Hébrail, G., and Y. Lechevallier. 2003. "Analyse de Données." In, edited by G. Govaert, 323–56. Hermès.
- Lefébure, R., and G. Venturi. 1998. *Le Data Mining*. Eyrolles.
- "Machine Learning, Data Science, Data Mining, Big Data, Analytics, AI." n.d. <http://www.kdnuggets.com> (<http://www.kdnuggets.com>).
- Tufféry, S. 2002. *Data Mining et Scoring: Bases de Données et Gestion de La Relation Client*. InfoPro (Paris). Dunod.
- . 2007. *Data Mining et Statistique décisionnelle: L'intelligence Des Données*. Editions Technip.
- Zighed, D.A., and R. Rakotomalala. 2000. *Graphes d'induction: Apprentissage et Data Mining*. Hermes Science Publications.