Introduction générale

Contexte

La progression exponentielle des outils informatiques (loi de Moore, 1965) a conduit à une augmentation conjointe des puissances de calcul et des volumes de stockage, sans pour autant en augmenter le prix de revient. Cette progression a favorisé la démocratisation de l'outil informatique tout en lui permettant de traiter des données de plus en plus complexes. L'usage quotidien de l'informatique, au départ réservé essentiellement aux entreprises, s'est banalisé au contact de la sphère privée. Aujourd'hui, grâce à la progression en parallèle des moyens de communication, la numérisation des données est devenue systématique aussi bien pour les particuliers que les professionnels. Il se pose alors le problème de la gestion et de l'exploitation des données.

Extraction des connaissances et fouille de données

L'augmentation du volume de stockage des données a motivé la recherche de connaissances dans ces données. L'objectif est de trouver des connaissances qui résument l'information que les données contiennent afin qu'un utilisateur puisse y accéder facilement. Pour y parvenir, Fayyad et al. (1996) proposent le processus d'Extraction de Connaissances à partir des Données (ECD, ou *Knowledge Discovery from Databases, KDD*, en anglais) dont le but ultime est la production de connaissances nouvelles, validées et qui soient interprétables par un utilisateur. La figure 1 schématise les différentes étapes de ce processus. Chaque étape est importante et conditionne la qualité finale des connaissances extraites.

Le processus défini par Fayyad et al. (1996) est suffisamment générique pour être applicable à différents contextes. Son application à un cas particulier requiert avant tout la définition des besoins utilisateurs. En fonction de ces besoins, les trois premières étapes du processus (sélection, prétraitement, transformation) préparent le sous-ensemble de données concerné tout en garantissant la cohérence de ces données. L'étape de sélection des données a pour objectif de définir, parmi les données disponibles, celles qui présentent un intérêt pour l'extraction des connaissances visées. Cette étape conditionne fortement le reste du processus. Au cours de l'étape suivante (prétraitement), les données sont nettoyées pour as-

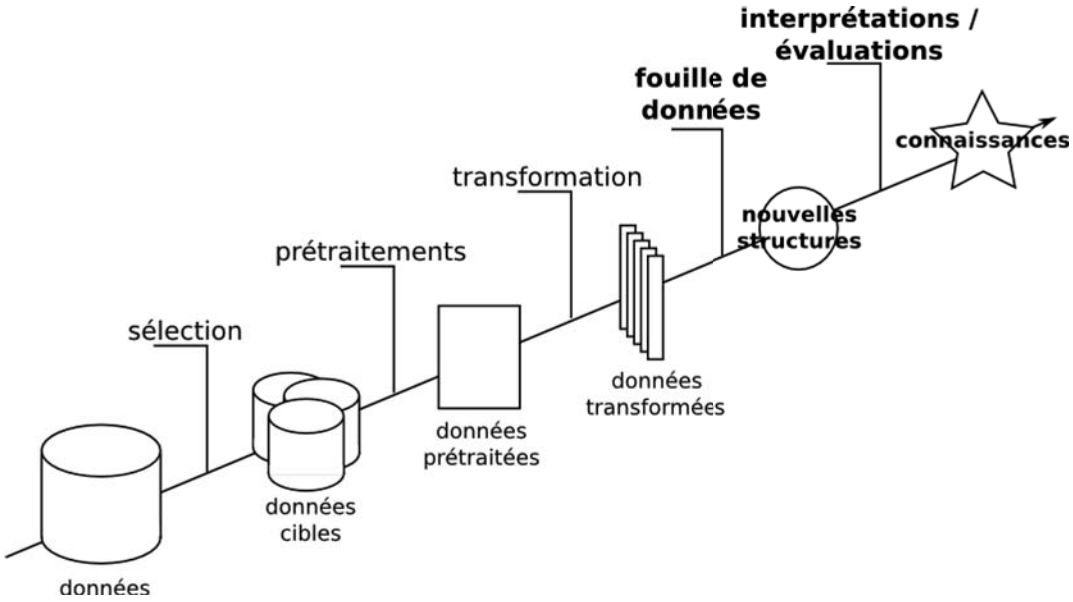


Fig. 1: Les différentes étapes du processus d'ECD selon Fayyad et al. (1996).

surer leur cohérence (suppression des données abérrantes, contrôle des informations issues de l'utilisateur, éventuellement normalisation des données). L'étape de transformation est la dernière étape avant la fouille. Elle recherche des projections des données qui puissent améliorer (en terme de qualité et/ou de temps de traitement) le processus de fouille de données. Dans un grand nombre de cas, à l'issue de la phase de transformation, les données sont reformulées sous la forme d'une matrice. Les lignes de cette matrice correspondent aux *exemples* décrits en colonne par leurs *attributs*. Ainsi, si l'on s'intéresse aux clients d'une banque, on cherchera à obtenir une matrice dont chaque ligne correspond à un client particulier et dont les attributs sont le nom, le solde du compte, le revenu, les prêts, etc... Les exemples sont des observations particulières du système à étudier, à partir desquels le procédé d'ECD va généraliser, suivant alors un raisonnement par induction. Ce n'est qu'à l'issue de ces étapes préparatoires qu'on procède à la fouille de données (*data mining*, en anglais). Cette étape procède à la découverte de structures inconnues dans les données. Il existe plusieurs familles de méthodes de fouille qui se distinguent les unes des autres suivant le type de données considérées :

Apprentissage non-supervisé. Les méthodes d'apprentissage non-supervisé cherchent à décrire et à structurer les données. Parmi elles, on distingue les méthodes de classification des méthodes d'association.

Les méthodes de classification cherchent à regrouper les exemples en classes en fonction de leur degré de ressemblance sur les attributs. Chaque classe fournit alors un résumé des données qu'elle contient. Dans le cas de la banque, par exemple, une classification des

clients permet d'obtenir des profils types de clients. Cette nouvelle connaissance pourra être exploitée par un service de marketing afin de proposer des offres promotionnelles ciblées.

Les méthodes d'association s'intéressent aux relations entre les états des attributs sous la forme de règles de type « si *prémisse* alors probablement *conclusion* ». Elles ont été développées par Agrawal et al. (1993) dans le contexte de bases de données de transactions clients où chaque exemple représente un ticket de caisse avec comme attributs les articles disponibles à l'achat. À partir de ces données, Agrawal et al. (1993) recherchent des règles de type « si un client achète du pain et du fromage, alors probablement il achète du vin ». Un ensemble de règles issu d'une telle base constitue une connaissance dont l'exploitation permet d'optimiser l'organisation des rayons de marchandises (nous exploitons les règles d'association dans le chapitre 4).

Apprentissage supervisé. En contexte supervisé, chaque exemple présent dans les données est déjà associé à une étiquette de classe. Les classes étant connues, le problème est alors de prédire pour un nouvel exemple sans étiquette, celle qui lui correspond. Prenons encore le cas d'une banque qui connaît pour chacun des clients à qui elle a déjà accordé un prêt, s'il l'a remboursé ou non (ici la classe d'intérêt). L'apprentissage supervisé consiste alors à prédire, pour un nouveau client qui demande un prêt, s'il va le rembourser ou non. Dans ce cas, la connaissance découverte est un modèle d'aide à la décision.

Apprentissage semi-supervisé. Le contexte de l'apprentissage semi-supervisé a émergé plus récemment, en partie suite à l'apparition de grands volumes de données. Dans une problématique de prédiction et face à des données comportant un grand nombre d'exemples, il devient parfois difficile pour un expert d'étiqueter chaque exemple d'apprentissage, en amont d'un processus supervisé. Cela conduit à deux sous-ensembles d'exemples, l'un étiqueté et l'autre non. Dans ce cadre, l'apport de l'apprentissage semi-supervisé par rapport aux méthodes usuelles d'apprentissage supervisé consiste à tenir compte des exemples non-étiquetés (souvent plus nombreux), en plus des exemples étiquetés, pour trouver le meilleur modèle de prédiction (voir section 7 du chapitre 3 pour une présentation détaillée).

Il existe encore d'autres familles de méthodes, comme celles d'indexation (dont l'objectif est de permettre un accès rapide à l'information, voir Weber et al., 1998, pour un exemple de cette problématique) ou encore celles de détection de fraudes (qui recherchent des comportements anormaux, voir par exemple Fawcett et Provost, 1997). Enfin, quelle que soit la méthode utilisée, le processus d'ECD termine toujours sur une étape de validation de la connaissance extraite, afin que son utilisateur puisse en apprécier la fiabilité.

Données volumineuses

Face à l'avènement du tout-numérique, le processus d'ECD est confronté à une augmentation conséquente du volume de données à analyser (Donoho, 2000). Ce changement d'échelle affecte les deux dimensions de la matrice attributs/valeurs qui sert à la description des données en vue de leur analyse par les algorithmes de fouille. Nous parlons ici de données volumineuses pour caractériser de telles données qui présentent à la fois un grand nombre d'exemples et d'attributs.

Concernant les attributs, leur nombre augmente consécutivement à deux phénomènes. Le premier concerne la nature des objets d'intérêt, qui passe de données symboliques et/ou numériques à des données dont la sémantique est difficilement accessible par la transformation en matrice attributs/valeurs (comme le texte, l'image ou encore les puces à ADN). Le deuxième phénomène concerne l'apparition de données complexes. En suivant la définition donnée par Darmont (2006), cette complexité provient essentiellement d'une caractéristique d'hétérogénéité qui apparaît soit au niveau de leur nature (données symboliques, numériques, textuelles, images, sons, multimédia), de leur structure (données structurées, semi-structurées ou non-structurées), de leur source (données en provenance de différents serveurs : bases de données, web, LDAP[1]), de leurs modalités (avec une modalité par point de vue sur un objet) et/ou de leurs versions (aspect temporel des données). Pour les algorithmes de fouille de données, cette complexité se traduit entre autres par une augmentation du nombre d'attributs afin de rendre compte de leur hétérogénéité.

Parallèlement à cette évolution des données vers davantage de complexité, le nombre d'exemples à analyser augmente également, en grande partie grâce à la numérisation et au stockage quasi-systématique de l'information.

Les données volumineuses ont un impact important sur les étapes de fouille de données et de validation du processus d'ECD. En effet, les algorithmes de fouille de données ne sont pas toujours adaptés aux données volumineuses, que ce soit au niveau du nombre d'attributs ou d'exemples. Le nombre d'attributs, d'abord, détermine la dimensionnalité de l'espace d'apprentissage. Or, des espaces de grandes dimensions entraînent des problèmes théoriques pour de nombreux algorithmes de fouille. Ces algorithmes sont également sensibles au nombre d'exemples puisque celui-ci a un impact direct sur le temps nécessaire à l'apprentissage. Ainsi, de nombreux algorithmes ne sont pas adaptés aux données volumineuses tout simplement parce que leur complexité algorithmique ne permet pas le passage à l'échelle.

[1] pour *Lightweight Directory Access Protocol*. Ce type de serveur peut servir à la gestion d'annuaire de personnes.

Contributions de la thèse

Dans le cadre de cette thèse, nous nous sommes plus spécifiquement intéressés aux problèmes rencontrés par l'apprentissage supervisé de données volumineuses, au cours des étapes de fouille de données (nous présentons nos contributions sur ce point aux chapitres 2 et 3) et de validation (nos contributions dans ce domaine sont présentées au chapitre 4).

Parmi les différentes méthodes d'apprentissage supervisé, nous nous sommes focalisés sur celles qui fondent leur prédiction sur une représentation des données, suivant ainsi les propositions de plusieurs travaux de recherche menés au sein du laboratoire ERIC (Seban et Zighed, 1996; Muhlenbach, 2002; Scuturici et al., 2005; Clech, 2004; Hacid, 2008). Néanmoins, la notion de représentation en apprentissage supervisé n'est pas une notion clairement établie. Aussi, commençons-nous cette thèse en proposant une définition de cette notion, suivie d'un état de l'art des stratégies d'apprentissage fondées sur la représentation. Nous terminons le premier chapitre en recensant les principaux problèmes rencontrés face à l'apprentissage de données volumineuses.

Dans le second chapitre, nous envisageons le recours aux cartes auto-organisatrices (Kohonen, 1982) pour obtenir rapidement une représentation des données en apprentissage supervisé. Après un état de l'art sur ces cartes en apprentissage non-supervisé, nous détaillons différentes stratégies pour les utiliser en prédiction. Nous proposons alors une adaptation de ces stratégies qui satisfasse à notre définition de représentation. Nous montrons également qu'il est possible de valider statistiquement cet apprentissage avant de s'en servir pour la prédiction. Enfin, nous expérimentons les limites des cartes auto-organisatrices face à l'apprentissage supervisé de données en grandes dimensions.

Afin de rendre une prédiction fondée sur la représentation des cartes auto-organisatrices robuste face à des données en grandes dimensions, nous envisageons le recours aux méthodes ensemblistes dans le troisième chapitre. Ces méthodes agrègent le résultat de plusieurs classifieurs pour obtenir une prédiction. Après un état de l'art sur le sujet, nous proposons deux stratégies d'apprentissage d'un ensemble de cartes auto-organisatrices, l'une ayant recours à l'étiquette pour choisir le meilleur ensemble de cartes et l'autre non. Nous montrons alors que les ensembles de cartes améliorent la qualité de la prédiction obtenue par une seule carte pour se rapprocher de la qualité obtenue par des méthodes ensemblistes réputées pour leurs performances en prédiction, telles que les forêts aléatoires (Breiman, 2001) et le boosting (Freund et Schapire, 1996). Les ensembles de cartes conduisant à l'apprentissage d'une multitude de représentations, nous proposons également une stratégie pour résumer l'information de ces cartes en une seule représentation qui peut être présentée à l'utilisateur. À la suite de ces travaux sur les ensembles de cartes, nous envisageons deux champs d'application, celui des données bruitées et celui de l'apprentissage semi-supervisé. Dans ces 2 domaines, les ensembles de cartes se montrent plus robustes que les méthodes d'ap-

prentissage supervisé usuelles et, dans le cas de l'apprentissage semi-supervisé, compétitifs par rapport aux méthodes du domaine. À partir des résultats obtenus et afin d'illustrer les avantages d'un apprentissage fondé sur la représentation, nous profitons de la discussion pour caractériser les bases de données qui peuvent tirer parti de ce type d'apprentissage.

Alors que les trois premiers chapitres sont orientés vers l'étape de fouille de données, le chapitre 4 s'intéresse, quant à lui, à la validation des résultats dans le contexte de données volumineuses, en particulier à la problématique de contrôle du risque multiple. Ce problème émerge lorsque le processus de fouille produit un ensemble de connaissances qui doivent être validées individuellement. Il en résulte une multitude de tests, ce qui a pour effet d'accroître le risque de première espèce, c'est-à-dire ici celui de valider une connaissance à tort. Les travaux que nous avons menés dans ce domaine sont reliés au processus d'extraction de règles d'association. La taille de l'ensemble de règles produit est d'autant plus grand que les données sont volumineuses. Pour valider ces règles (ou tout autre ensemble de connaissances évalué suivant une mesure statistique), nous proposons une méthode non-paramétrique de contrôle du risque multiple fondée sur le *bootstrap*. Nous appliquons cette stratégie dans deux contextes. Le premier contexte est non-supervisé et correspond à la découverte de règles d'association significatives. Dans le second contexte, supervisé, nous nous intéressons à la découverte et à l'exploitation de règles d'association différentiellement exprimées suivant la variable de classe. Nous montrons l'intérêt de cette approche sur des données génomiques qui ont la particularité de présenter un très grand nombre d'attributs.

Notations

Dans la suite de cette thèse, les différents algorithmes de fouille sont appliqués sur une matrice de données X, de taille $n \times d$, avec x_i le i^e exemple et x_{ij} la valeur du j^e attribut de cet exemple i. Dans le cas de l'apprentissage supervisé, les exemples sont également décrits par une variable à expliquer, que nous appelons classe dans le présent document, notée Y, avec y_i sa valeur pour le i^e exemple ou encore l'étiquette du i^e exemple.

En fouille de données, les exemples peuvent être vus comme des points dans l'espace des attributs, encore appelé espace d'apprentissage. Les méthodes d'apprentissage fondées sur la représentation reposent alors sur le calcul des distances entre points. Nous notons $d(A, B)$ la distance euclidienne entre les points A et B dans cet espace d'apprentissage. Par raccourci, lorsque ces points sont des exemples d'apprentissage, nous utilisons la notation $d_{ij} = d(X_i, X_j)$ où i et j sont les indices des exemples dans la matrice X. Par ailleurs, afin de permettre la visualisation, les méthodes de représentation projettent les données dans un espace de petite dimension, appelé espace de projection ou espace de représentation. La distance euclidienne dans ce nouvel espace est alors notée δ.