INTRODUCTION AU DATA MINING

6 séances de 3 heures

mai-juin 2006

EPF - 4ème année - Option Ingénierie d'Affaires et de Projets

Bertrand LIAUDET

INTRODUCTION

Qu'est-ce que le data mining

Fantasmes

- Un des développements technologiques les plus révolutionnaires des dix prochaines décennies.
- Une des dix technologies émergentes qui vont changer le monde.

Formules et métaphores

- Le data mining est un procédé qui permet de passer des données à la connaissance.
- Le data mining est un procédé qui permet de découvrir des « pépites » d'informations cachées dans les données.

Définition

Le data mining est un procédé d'analyse de grands volumes de données en vue d'une part de les rendre plus compréhensibles et d'autre part de découvrir des corrélations significatives (des règles et des tendances). Le data mining utilise des outils statistiques et informatiques.

Autre formulation :

Le data mining est un procédé d'exploration et d'analyse de grandes quantités de données dans le but de découvrir des phénomènes et des règles significatives.

Le data mining est un procédé de production de connaissance. En terme de logique philosophique traditionnelle[1], le data mining consiste à produire des jugements (toutes les personnes sont x, la moyenne des x des personnes vaut tant, etc. : c'est l'étape de description, l'étape concernant les données) et des raisonnements (toutes les personnes sont « a » alors elles seront « b » selon ce mode de corrélation : c'est l'étape prédiction, l'étape de modélisation).

Finalités

Le data mining est un outil qui permet de produire de la connaissance, soit dans un but informatif, soit dans un but d'aide à la décision : pour l'informatique décisionnelle.

Public

---

[1] **Thiry Philippe**, *Notions de logique*, De Boeck 1996, pp75-87 (182 p., environ 100 francs)

Chazal Gérard, *Eléments de logique formelle*, Hermes, 1996, pp. 17 à 29 (224 p., environ 150 francs). G. Chazal enseigne la logique à des philosophes et à des informaticiens. Ce livre est une initiation à la logique bien adaptée aux ingénieurs.

- L'analyste (qui produit des rapports pour le décideur)
- Le décideur (pour l'aide à la décision)
- Le gestionnaire (pour connaître ses données)
- Le publiciste (pour savoir comment informer)
- Le marketing

Méthodes

- Du bon sens : pour une part, il s'agit d'analyser les données avec bon sens.
- Des algorithmes de calculs statistiques. Pour une autre part, il s'agit d'appliquer des algorithmes de calculs à des données.

Pourquoi la naissance du data mining ?

- Augmentation des capacités de stockage des données (disques durs de giga octets).
- Augmentation des capacités de traitements des données (facilité d'accès aux données : il n'y a plus de bandes magnétiques ; accélération des traitements).
- Maturation des principes des bases de données (maturation des bases de données relationnelles).
- Croissance explosive de la collecte des données (scanners de supermarché, internet, etc.)
- Plus grande disponibilité des données grâce aux réseaux (intranet et internet).
- Développement de logiciels de data mining

Intérêt du data mining

Les entreprises sont inondées de données (scanners des supermarchés, internet, bases de données, etc.). Ces données languissent dans des entrepôts de données (ou référentiels, ou *data warehouse*).

Nous sommes noyés dans l'information, mais assoiffés de connaissance (trop d'information tue l'information).

Il n'y a pas assez d'humains entraînés et formés pour transformer ces données en connaissances.

Le data mining permet, en partie, de combler ce manque de ressources humaines.

Le data mining renforce la capacité des entreprises à découvrir des phénomènes et des tendances lucratives dans les bases de données existantes.

Le data mining permet d'augmenter le retour sur investissement des systèmes d'information.

Applications du data mining

Exemple de la gestion de la relation client : Customer Relationship Management (CRM)

Principe : amélioration de la rentabilité par l'amélioration de la connaissance du client.

Matière première : les données sur le client.

CRM analytique : collecte et analyse des données.

CRM opérationnel : communication entreprise vers client et client vers entreprise (pour de nouvelles collectes de données).

Difficulté : tirer partie de la masse de données. Ne pas se noyer dedans.

Objectif : on ne veut plus seulement savoir : « combien de clients ont acheté tel produit pendant telle période ?», mais on veut savoir « quel est leur profil ? », « quels autres produits les intéresseront ? », « quand seront-ils de nouveau intéressés ? ».

Autres exemples

- Secteur bancaire : mesure des scores d'appétence et de risque pour mieux cibler les propositions commerciales et éviter les surendettements (scoring). L'objectif des banques est de réduire le risque des prêts bancaires.
- Sujets dominants dans le secteur de la téléphonie : le score d'attrition (usure, *churn* en anglais), c'est-à-dire le changement d'opérateur ; le text mining pour analyser les lettres de réclamation ; le score des impayés.
- Web mining et E-commerce : 50% des clients d'un constructeur de machine achètent ses machines à travers le web. Mais seulement 0,5% des visiteurs du site deviennent clients. L'idée est de stocker les séquences de click des visiteurs et d'analyser les caractéristiques des acheteurs pour adapter le contenu du site.
- Fidélisation de la clientèle : pour le public d'un théâtre, on peut segmenter les clients par des critères d'ancienneté, de durée et de fréquence de fréquentation (forme de la consommation), mais aussi faire une typologie par genre de spectacle (contenu de la consommation).
- Une société de cosmétique de luxe détecte automatiquement ses meilleurs clients dès les premières transactions dans sa base de données pour les traiter avec le plus d'égards possibles.
- Un opérateur de télévision par abonnement détecte les clients les plus sensibles à des offres de chaînes complémentaires à partir des appels téléphoniques des clients.

Présentation schématique des relations entre statistiques et data mining

Présentation

On peut avoir une opposition un peu caricaturale :

- D'un côté, les statisticiens ignorent ou méprisent le data mining en considérant que ce n'est pas de la statistique.
- D'un autre côté, les « data miners » marginalisent la statistique.

La situation est paradoxale :

- D'un côté, le data mining utilise des techniques statistiques, mais certains préfèrent l'ignorer.
- D'un autre côté, les instituts de statistique possèdent des gisements de données considérables mais les exploitent peu avec les techniques du data mining.

Distinction entre statistiques et data mining

Une recherche statistique se décompose en quatre étapes :

- la définition et la collecte des données
- la présentation des données en tableaux
- l'analyse des données
- la comparaison des résultats avec des lois statistiques connues (regarder si les résultats obtenus correspondent à des lois générales).

C'est à partir de ces quatre étapes de la recherche statistique qu'on va aborder la distinction entre statistiques et data mining.

La quantité de données :

On a tendance à considérer que le data mining traite plus de données que les statistiques. C'est une idée partiellement juste. Dans l'absolu, les deux techniques peuvent en traiter autant. Dans la pratique, le data mining en traite plus. Cela vient des deux distinctions suivantes : l'origine des données et l'aprioricité des traitements.

L'origine des données :

Les statisticiens peuvent travailler sur des populations entières, mais le plus souvent ils travaillent sur des échantillons. Leur travail consistera à construire la représentativité de ces échantillons. Souvent, le statisticien est donc amené à collecter des données qui n'existent pas encore.

Le data miner travaille plutôt sur des données qui existent déjà (issues de bases de données ou de *data warehouse*). De ce fait (et aussi du fait des techniques de modélisation qu'il utilise) il a tendance à travailler sur la population entière et pas sur des échantillons.

L'analyse des données :

Le data mining utilise des techniques issues de l'intelligence artificielle. Certaines techniques du data mining n'appartiennent qu'au data mining et ne font pas partie de la panoplie des techniques de l'analyse de données. C'est le cas des techniques de réseaux de neurones et d'arbres de décision.

Il tend à travailler avec moins d'*a priori* que les statistiques traditionnelles (moins d'hypothèses de départ). D'où la tendance à en faire un produit donnant des résultats miraculeux !

Le data mining recherche parfois plus la compréhensibilité des modèles que leur précision (d'où le mépris des statisticiens !)

Les modèles du data mining sont en général plus localisés que ceux des statisticiens.

Le data mining ne s'intéressent pas, contrairement aux statisticiens, aux lois générales de la statistique : c'est un domaine directement appliqué.

Ce que fait le data mining

Techniques descriptives et prédictives

Le data mining met en œuvre un ensemble de technique issues des statistiques, de l'analyse de données et de l'informatique pour explorer les données.

Il permet d'accomplir les six types d'analyse suivants :

Description, Classification, Association, Estimation, Segmentation et Prévision.

Techniques descriptives		**Techniques prédictives**		
Particulière	**Générale**	**Présent**		**Futur**
		Numérique	**Catégorie**	
1 : Description	2 : Classification 3 : Association	4 : Estimation	5 : Segmentation	6 : Prévision

Les techniques descriptives (archétype : la classification)

- Elles visent à mettre en évidence des informations présentes mais cachées par le volume des données.
- Elles réduisent, résument, synthétisent les données.
- Elles ne présentent pas de variable cible à prédire.

Les techniques prédictives (archétype : le scoring)

- Elles sont plus délicates à mettre en œuvre que les techniques descriptives. De plus, elles demandent plus d'historique.

- Elles visent à extrapoler de nouvelles informations à partir des informations présentes (c'est le cas du scoring).
- Elles présentent une variable cible à prédire.

Rappels de logique - vocabulaire - notions de donnée, de variable, de type, de modèle

Globalement, on travaille sur des tableaux de données.

- Le tableau, c'est « ce dont on parle », c'est-à-dire le « concept » dont on parle. Par exemple, un tableau de clients, de malades, etc.

  Rappelons qu'un concept (ou **notion**, ou **idée**) est une représentation mentale générale et abstraite d'un objet. Le concept est le résultat de l'opération de l'esprit qui fait qu'on place tel objet dans telle catégorie et non dans telle autre.
- Chaque ligne du tableau est un élément du tableau, c'est-à-dire un objet concret correspondant au concept abstrait dont on parle.

  En data mining, un objet concret est appelé : « donnée ».
- Chaque colonne du tableau est un attribut du « concept ». On parle aussi de « propriété ». La colonne a un nom général qui est le nom de l'attribut pour le concept. Pour un objet concret, la colonne a une valeur particulière qui est la valeur particulière de l'attribut pour l'objet concret.

  En data mining (et en statistique), les attributs des objets sont appelés : « variables ».
- **Un sous-ensemble de valeurs pour un ou plusieurs attributs donnés peut être appelé : « type », « classe », « catégorie » ou encore « segment ».** Par exemple, « grand » et « petit » sont deux types de l'attribut « taille ».
- Quand on fait de la prévision, on travaille sur une variable particulière appelée : « variable cible » et sur un ensemble d'autres variables utiles pour la prédiction appelées : « prédicteurs ».

  Le principe général de la prédiction sera : si le ou les prédicteurs valent tant, alors la variable cible vaut tant.
- Les statisticiens construisent des **modèles**. Un modèle est un résumé global des relations entre variables permettant de comprendre des phénomènes (description, jugement) et d'émettre des prévisions (prédiction, raisonnement). Dans l'absolu, tous les modèles sont faux. Cependant, certains sont utiles.

<u>L'analyse des données consiste à analyser :</u>

- Pour une variable donnée : la répartition de ses valeurs (tri, histogramme, moyenne, etc.).
- Pour deux ou trois variables données : le lien entre les répartitions des valeurs des variables.
- Pour n variables : des sous-ensembles disjoints de données.

1 : la description

Principe :

La description consiste à mettre à jour des phénomènes encore appelés des tendances, c'est-à-dire des liens entre deux ou trois variables.

Intérêt :

- Favoriser la connaissance et la compréhension des données.

Méthode :

- Méthodes graphiques pour la clarté : analyse exploratoire des données.

Exemples :

2 : la classification

Principe :

La classification (ou clustering) consiste à créer des classes (c'est-à-dire des sous-ensembles) de données similaires entre elles et différentes des données d'une autre classe (autrement dit, l'intersection des classes entre elles doit toujours être vide).

On dit aussi « segmenter » l'ensemble entier des données.

La classification définit donc les grands types de regroupement et de distinction : on parle de métatypologie (type de type).

Elle permet une vision générale de l'ensemble (de la clientèle, par exemple).

C'est une logique de « group by » multiples.

Intérêt :

- Favoriser, grâce à la métatypologie, la compréhension et la prédiction.
- Fixer des segments qui serviront d'ensemble de départ pour des analyses approfondies.
- Réduire les dimensions, c'est-à-dire le nombre d'attributs, quand il y en a trop au départ.

Méthodes :

- Classification hiérarchique des k plus proches voisins.
- Réseaux de Kohonen.

Exemples :

- Métatypologie d'une clientèle en fonction de l'âge, les revenus, le caractère urbain ou rural, la taille des villes, etc.
- Pour un audit comptable, classer un comportement financier en catégorie normale et suspecte.

3 : l'association

Principe :

L'association consiste à **trouver quelles variables vont ensemble**. C'est une sorte de classification générale : on croise toutes les valeurs possibles des variables qui nous intéressent. On va alors déterminer le poids statistique de chaque combinaison : c'est-à-dire déterminer quelles variables vont ensemble.

Les règles d'association sont de la forme : si antécédent, alors conséquence.

On appelle aussi ce type d'analyse une « analyse d'affinité ».

Intérêt :

Mieux connaître les comportements.

Méthode :

- Algorithme *a priori.*
- Algorithme du GRI (induction de règles généralisée).

Exemples :

- Analyse du panier de la ménagère..
- Étudier la part des abonnés à une compagnie de téléphone portable qui répond positivement à une offre de nouveaux services.

4 : l'estimation[2]

Principe :

L'estimation consiste à définir le lien entre un ensemble de prédicteurs et une variable cible. Ce lien est défini à partir de données « complètes », c'est-à-dire dont les valeurs sont connues tant pour les prédicteurs que pour la variable cible. Ensuite, on peut déduire une variable cible inconnue de la connaissance des prédicteurs.

À la différence de la segmentation qui travaille sur une variable cible catégorielle, l'estimation travaille sur une variable cible numérique.

Intérêt :

- Permettre l'estimation de valeurs inconnues.

Méthodes :

- Analyse statistique classique : régression linéaire simple, corrélation, régression multiple, intervalle de confiance, estimation de points.
- Réseaux de neurones

Exemples :

- Estimer la pression sanguine systolique à partir de l'âge, le sexe, le poids et le niveau de sodium dans le sang.
- Estimer les résultats dans les études supérieures en fonction de critères sociaux.

5 : la segmentation

Principe :

La segmentation est une estimation qui travaille sur une variable cible catégorielle.

On parle de segmentation car chaque valeur possible pour la variable cible va définir un segment (ou type, ou classe, ou catégorie) de données.

Intérêt :

- Permettre l'estimation de valeurs inconnues.

Méthodes :

- Graphiques et nuages de points.
- Méthode des k plus proches voisins.

[2] Reprise du 1er cours.

- Arbres de décision.
- Réseau de neurones.

Exemples :

- Segmentation par tranche de revenus : élevé, moyen et faible (3 segments).
- Déterminer si un mode de remboursement présente un bon ou un mauvais niveau de risque crédit (deux segments).

6 : la prévision

Principe :

La prévision est similaire à l'estimation et à la segmentation mise à part que pour la prévision, les résultats portent sur le futur.

Intérêt :

- Permettre l'estimation de valeurs inconnues.

Méthodes :

- Celles de l'estimation ou de la segmentation.

Exemples :

- Prévoir le prix d'action à trois mois dans le futur.
- Prévoir le temps qu'il va faire.
- Prévoir le gagnant du championnat de football, par rapport à une comparaison des résultats des équipes.

Historique des techniques de statistiques et de data mining

1875	**Régression linéaire de Francis Galton.**
1896	**Formule du coefficient de corrélation de Karl Pearson[3].**
1900	Distribution du X^2 de Karl Pearson.
1936	Analyse discriminante de Fischer et Mahalanobis
1941	Analyse factorielle des correspondances de Guttman
1943	Réseaux de neurones de Mac Culloch et Pitts
1944	Régression logistique de Joseph Berkson
1958	Perceptron de Rosenblatt
1962	Analyse des correspondances de J.-P. Benzécri
1962	Régression logistique de J. Cornfield
1964	Arbre de décision AID de J.-P. Sonquist et J.-A. Morgan
1965	Méthode des centres mobiles de E. W. Forgy
1967	Méthode des k-means de Mac Queen
1972	Modèle linéaire généralisé de Nelder et Wedderburn
1975	Algorithme génétique de Holland
1977	Méthode de classement DISQUAL de Gilbert Saporta
1980	Arbre de décision CHAID de KASS
1983	Régression PLS de Herman et Svante Wold
1984	Arbre CART de Breichman, Friedman, Olshen, Stone
1986	Perceptron multicouches de Rumelhart et Mac Clelland
1989	Réseaux de T. Kohonen (cartes auto-adaptatives)
1990	**Apparition du concept de Data Mining**
1993	Arbre C4.5 de J. Ross Quinlan
1996	Bagging (Breiman) et boosting (Freund-Shapire)
1998	Support vector machine de Vladimir Vapnik
2001	Régression logistique PLS de Tenenhaus

[3] Karl Pearson, (1857-1936), mathématicien et philosophe britannique qui a mis au point les principales techniques statistiques modernes et les a appliquées aux questions de l'hérédité.

Quelques idées fausses sur le data-mining

Concernant le caractère automatique du data mining

- Le data mining nécessite peu ou pas de supervision humaine.
- Il existe des logiciels de data mining que nous pouvons faire tourner automatiquement sur nos bases de données pour trouver des réponses à nos questions.
- Les logiciels de data mining produisent de la connaissance.
- Les logiciels de data mining nettoient les bases de données erronées automatiquement.
- Les logiciels de data mining sont intuitifs et faciles à utiliser.

Concernant les résultats attendus du data mining

- Le data mining va répondre à mes questions.
- Le data mining s'amortit rapidement.

Concernant la technique du data mining

- Un algorithme de data mining est d'autant plus efficace qu'il a plus de données en entrée.

  ⇒ L'important n'est pas la quantité, mais la qualité des données.
- Aucun *a priori* n'est nécessaire.

  ⇒ Ce n'est pas vrai pour les techniques prédictives qui nécessitent au moins l'*a priori* de la variable cible.

  ⇒ C'est en partie vrai pour les techniques descriptives et la classification particulièrement.
- Il ne faut jamais échantillonner

  ⇒ Certaines techniques prédictives (arbres de décision, réseaux de neurones) imposent un échantillonnage.
- Il ne faut toujours échantillonner

  ⇒ Un bon échantillonnage n'est pas facile à faire. Certaines techniques permettent de l'éviter.

Concernant les compétences requises pour faire du data mining

- Avec le data mining, on peut se passer des spécialistes du métier

  ⇒ On a très souvent besoin d'un spécialiste du métier dans la phase de compréhension des données : pour la préparation et le nettoyage particulièrement. Bien sûr, c'est aussi le spécialiste métier qui peut définir les objectifs, et le spécialiste métier qui peut orienter les résultats et dire s'ils présentent un intérêt ou s'ils sont triviaux (attention à ne pas croire qu'on a découvert l'eau chaude !)
- Avec le data mining, on peut se passer des statisticiens

⇒ Dans une étude de data mining, la partie la plus longue consiste à préparer les données. Ce travail est un travail de statisticien (même s'il est assez simple). Ce n'est qu'une fois réalisé qu'on peut appliquer les algorithmes des modèles.

Conclusions

Lancer une étude de data mining, c'est comme partir chercher de l'or dans une mine ou dans une rivière :

1) Ca nécessite du travail !
2) On n'est pas sûr de trouver quoi que ce soit !
3) Il vaut mieux être entouré de personnes compétentes

Dernière remarque : la corrélation statistique n'est pas la causalité physique !

La corrélation statistique ne nous dit jamais rien sur le plan individuel, mais seulement sur le plan statistique. Même si on a le profil de quelqu'un qui ne va pas payer ses dettes, on les paiera peut-être quand même.

La corrélation statistique n'est pas la causalité physique. Même si on sait qu'une campagne de promotion augmente en moyenne les ventes de x%, il se peut que la nouvelle campagne de promotion ne produise aucun résultat (par contre, quand je lâche mon crayon, il tombe toujours par terre !).

Le processus standardisé CRISP- DM

Le besoin d'un contrôle humain dans le data mining

Une discipline et pas un produit

À l'origine, le data mining était vue comme un procédé automatique ou semi automatique.

Aujourd'hui, on est revenu de cette illusion. Le data mining n'est pas un produit qui peut être acheté, mais bien une discipline qui doit être maîtrisée.

Avant d'appliquer automatiquement des algorithmes de calculs sur les données, il faut passer par une phase d'exploration et d'analyse qui ne saurait être automatisée : elle fait intervenir le bon sens et la connaissance du contexte (culture générale).

Quand on veut produire de la connaissance, le problème ne se limite pas à répondre à des questions. Il faut d'abord poser les questions. C'est cette première étape qui, pour l'essentiel, fait que le data mining est une discipline et pas un simple produit.

Comment faire du mauvais data mining ?

- En travaillant sans méthode
- En ne préparant pas correctement ses données.
- En appliquant des boîtes noires de calculs sans les comprendre.

Un mauvais data mining peut amener à des conclusions erronées et donc à des conséquences très coûteuses.

Comment faire du bon data mining ?

- En suivant une méthode
- En préparant les données correctement
- En comprenant le principe des modes opératoires (des algorithmes de calculs). En étant capable de savoir pourquoi on en choisit un plutôt qu'un autre. Une compréhension des modèles statistiques appliqués par le logiciel est donc nécessaire.

Présentation du CRISP-DM

Le data mining est un procédé : une méthode employée pour parvenir à un certain résultat.

C'est aussi un processus : une suite ordonnée d'opérations aboutissant à un résultat.

Le CRISP-DM (Cross Industry Standard Process for Data Mining) décrit le data mining processus itératif complet constitué de 4 étapes divisées en tout en 6 phases.

PROCESSUS du DATA MINING	
Étapes	**Phases**
Objectifs	1 : Compréhension du métier
Données	2 : Compréhension des données
Traitements	3 : Préparation des données
	4 : Modélisation
	5 : Évaluation de la modélisation
Résultats	6 : Déploiement des résultats

Les étapes se suivent dans l'ordre : il s'agit d'abord de déterminer les objectifs de l'étude, puis de comprendre et de préparer les données, puis de choisir les traitements à appliquer et enfin de produire les résultats.

Dans chaque étape, les phases se font aussi dans l'ordre.

À chaque phase, on peut toujours revenir en arrière (passer de la phase n à la phase n-1).

Le caractère itératif du data mining vient du fait que les résultats obtenus après le déploiement peuvent conduire à une nouvelle compréhension tant du métier que des données et, de là, à un nouveau processus.

À noter que la phase de préparation des données peut aussi précéder celle de la compréhension des données.

Description des phases du data mining

1. **La phase de compréhension du métier**

   Elle consiste à :

   - Énoncer clairement les objectifs globaux du projet et les contraintes de l'entreprise.
   - Traduire ces objectifs et ces contraintes en un problème de data mining.
   - Préparer une stratégie initiale pour atteindre ces objectifs.

2. **La phase de compréhension des données**

   Elle consiste à :

   - Recueillir les données.
   - Utiliser l'analyse exploratoire pour se familiariser avec les données, commencer à les comprendre et imaginer ce qu'on pourrait en tirer comme connaissance.
   - Évaluer la qualité des données.
   - Éventuellement, sélectionner des sous-ensembles intéressants.

3. **La phase de préparation des données**

   Elle consiste à :

- Préparer, à partir des données brutes, l'ensemble final des données qui va être utilisé pour toutes les phases suivantes.
- Sélectionner les cas et les variables à analyser.
- Réaliser si nécessaire les transformations de certaines données.
- Réaliser si nécessaire le nettoyage de certaines données.

Cette phase s'applique essentiellement avant chaque application d'un modèle. Elle s'applique aussi avant la compréhension des données.

4. La phase de modélisation

Elle consiste à :

- Sélectionner les techniques de modélisation appropriées (souvent plusieurs techniques peuvent être utilisées pour le même problème).
- Calibrer les paramètres des techniques de modélisation choisies pour optimiser les résultats.
- Éventuellement revoir la préparation des données pour l'adapter aux techniques utilisées.

5. La phase d'évaluation de la modélisation

Elle consiste à :

- Pour chaque technique de modélisation utilisée, évaluer la qualité (la pertinence, la signification) des résultats obtenus.
- Déterminer si les résultats obtenus atteignent les objectifs globaux identifiés pendant la phase de compréhension du métier.
- Décider si on passe ou pas à la phase de déploiement (autrement dit décider de la fin de l'étude).

6. La phase de déploiement des résultats obtenus

Elle consiste à :

- Produire un rapport.

Conclusion synthétique

Un processus d'étude de data mining consiste à

1. Poser le problème.
2. Comprendre et préparer les données.
3. Appliquer des modèles pour produire des corrélations entre les données et évaluer la pertinence de ces corrélations.
4. Faire un rapport de synthèse.

Les logiciels de data mining

Il existe de nombreux logiciels de statistiques et de data mining sur PC. Certains sont gratuits, d'autres sont payants. Certains sont mono-utilisateur. D'autres fonctionnent en architecture clients-serveur.

Parmi les gros logiciels, on peut citer :

- Clementine de SPSS. Clementine est la solution de data mining la plus vendue dans le monde. C'est celle qu'on utilisera en démonstration dans ce cours.
- Entreprise Miner de SAS.
- Statistica Data Miner de StatSoft
- Insightful Miner de Insightful

Parmi les logiciels gratuits, on peut citer :

- TANAGRA, logiciel de data mining gratuit pour l'enseignement et la recherche.
- ORANGE, logiciel libre d'apprentissage et de data mining.
- WEKA, logiciel libre d'apprentissage et de data mining.

En résumé

Le data mining, ou l'extraction des connaissances, est un ensemble de techniques permettant de dégager des généralités de grands ensembles de données, c'est-à-dire de décrire de manière utile et compréhensible les régularités émergeant de ces données.

Les années 90 ont vu les bases de données croître de manière exponentielle, pour atteindre des capacités se mesurant en terabits (10^{12} bits). L'émergence de ces entrepôts de données (ou *data warehouse* en anglais) a rendu impossible toute exploitation manuelle de ces données. Pour pouvoir procéder à l'analyse de ces données, il a fallu adapter d'anciennes techniques de statistiques ou d'intelligence artificielle, ou en inventer de nouvelles.

L'extraction de connaissances (ou knowledge discovery) se déroule en plusieurs étapes. La première consiste à sélectionner, dans un entrepôt de données, celles qui sont pertinentes pour le but recherché. Ces données sont ensuite transformées, réorganisées ou converties, afin de pouvoir être analysées : c'est la phase de nettoyage. L'analyse de ces données, également appelée fouille de données (ou data mining), peut alors être réalisée, à l'aide de différentes techniques d'intelligence artificielle, tels les arbres de décision et les réseaux de neurones, ou à l'aide de techniques statistiques, permettant de dégager des régularités à l'intérieur de ces données. Les résultats sont alors transformés pour permettre leur interprétation et leur visualisation en fonction du but recherché.

Les applications de l'extraction de connaissances sont multiples : ces techniques sont utilisées par les banques (détection de fraudes, sélection de clients susceptibles d'être intéressés par un produit particulier), les investisseurs (prédictions boursières), les scientifiques (identification et classification d'objets célestes), les services de police (recherche de transactions indiquant

le blanchiment d'argent), les gérants de supermarchés (analyse des corrélations existant entre différents produits).

Méthode, principe et pré-requis du cours

- Le cours va présenter le data mining en tant que processus.

  Il traitera donc essentiellement deux parties :

  La compréhension et la préparation des données.

  La présentation de certaines techniques utilisées dans la modélisation, c'est-à-dire l'analyse des données.
- Ce cours s'adresse à des ingénieurs : ni à des statisticiens, ni à des informaticiens.

  Il ne s'agit ni de maîtriser parfaitement les mathématiques des statisticiens, ni de comprendre parfaitement les algorithmes de calcul. Il s'agit de comprendre les données qu'on manipule pour en faire des connaissances qui permettent de prendre des décisions.
- Ce cours est adapté à un niveau « master ». Un cours d'introduction aux statistiques est souhaitable mais non nécessaire. Aucune expertise en programmation n'est nécessaire.

Bibliographie

- Des données à la connaissance, une introduction au data-mining. Daniel T-Larose. Vuibert, Paris, 2005. Traduction de An introduction to data-mining, New-York, 2005.

  Inclus: une version d'évalution de Clémentine (SPSS).
- Data mining et statistique décisionnelle. Stéphane Tuffery. Editions Technip, Août 2005.
- Introduction au Data Mining, Analyse intelligente des données. Michel Jambu. Eyrolles, 1999.

Sites internet

http://chirouble.univ-lyon2.fr/~ricco/data-mining : sur le data mining.

http://chirouble.univ-lyon2.fr/~ricco/data-mining/logiciels : liste de logiciels libres de data mining.

http://chirouble.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html : présentation du logiciel TANAGRA.

http://eric.univ-lyon2.fr/~ricco/tanagra/fr/contenu_tutoriaux_comparaison_logiciels.html: comparaison de logiciels libres.
http://www.kdnuggets.com/software/suites.html : présentation de logiciels libres.
http://data.mining.free.fr : site de Stéphane Tuffery.