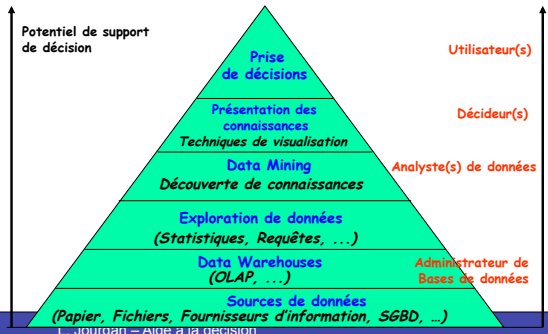
1

Fouille de données

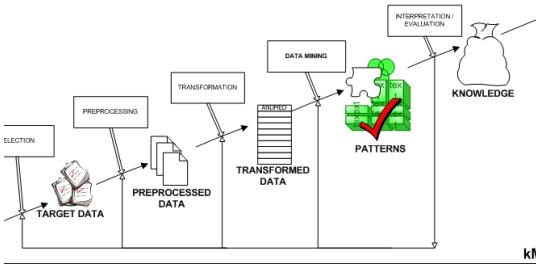
2

Data Mining et aide à la décision



3

Un processus complet



4

La fouille de données = Datamining

Le **datamining** est l'ensemble des

- Techniques et méthodes
- … Destinées à l'exploration et l'analyse
- … De (souvent) grandes bases de données
- … En vue de détecter dans ces données des règles, des associations, des tendances inconnues (non fixées a priori) des structures particulières restituant de façon concise l'esnentiel de l'information utile
- … Pour l'aide à la décision

Souvent on utilise le terme extraire de l'information de la donnée

Selon le MIT, le Datamining est l'une des 10 technologies émergentes qui changeront le monde au XXI siècle

5

Démarche méthodologique (1)

Comprendre l'application

- Connaissances *a priori*, objectifs, etc.

Sélectionner un échantillon de données

- Choisir une méthode d'échantillonnage

Nettoyage et transformation des données

- Supprimer le «bruit» : données superflues, marginales, données manquantes, etc.
- Effectuer une sélection d'attributs, réduire la dimension du problème, discrétisation des variables continues, etc.

Appliquer les techniques de fouille de données (DM)

- le cœur du KDD
- Choisir le bon modèle et le bon algorithme

6

Démarche méthodologique (2)

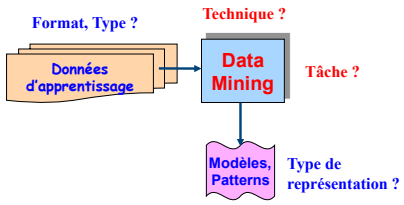
•Visualiser, évaluer et interpréter les modèles découverts

- Analyser la connaissance (intérêt, critères d'évaluation)
- Compréhensibilité souvent capitale
- Vérifier sa validité (sur le reste de la base de données)
- Réitérer le processus si nécessaire

•Gérer/déployer la connaissance découverte

- La mettre à la disposition des décideurs
- L'échanger avec d'autres applications (système expert, …)
- etc.

Paramètres d'un processus KDD



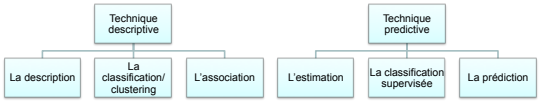
Les données : matière première

Valeurs des champs (p attributs ou variables) des enregistrements (n lignes ou cas) des tables de l'entrepôt (base de données, matrice n*p)

Types :

- Données discrètes : données binaires (sexe, …), données énumératives (couleur, …), énumératives ordonnées (réponses 1:très satisfait, 2:satisfait, …).
- Données continues : données entières ou réelles (âge, salaire, …)
- Dates
- Données textuelles
- Pages/liens web, Multimédia, …

2 types de techniques



Les 2 types de techniques

- Les techniques descriptives (recherche de patterns) :
  - Visent à mettre en évidence des informations présentes mais cachées par le volume de données
  - Réduisent, résument et synthétisent les données
  - Il n'y a pas de variables à expliquer
- Les techniques prédictives (modélisation) :
  - Visent à extrapoler de nouvelles informations à partir des informations présentes (c'est le cas du scoring)
  - Expliquent les données
  - Il y a une variable à expliquer

1 : la description (technique descriptive)

Principe :

La description consiste à mettre au jour

• Pour une variable donnée : la répartition de ses valeurs (tri, histogramme, moyenne, minimum, maximum, etc.).

• Pour deux ou trois variables données : des liens entre les répartitions des valeurs des variables. Ces liens s'appellent des « tendances ».

Intérêt :

• Favoriser la connaissance et la compréhension des données.

Méthode :

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

Exemples :

• Répartition des votes par âge (lien entre les variables « vote » et « âge »).

2 : la classification/clustering (technique descriptive)

Principe :

La classification (ou clustering ou segmentation) consiste à créer des classes/groupe (c'est-à-dire des sous-ensembles) de données similaires entre elles et différentes des données d'une autre Classe. Elle permet une vision générale de l'ensemble (de la clientèle, par exemple).

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

• Classification des K moyennes

• Réseaux de Kohonen.

• Règles d'association.

Exemples :

• Métatypologie d'une clientèle en fonction de l'âge, les revenus, le caractère urbain ou rural, la taille des villes, etc.

13

3 : l'association (technique descriptive)

Principe :

L'association consiste à trouver quelles valeurs des variables vont ensemble. Par exemple, telle valeur d'une variable va avec telle valeur d'une autre variable.

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

L'association ne fixe pas de variable cible. Toute les variables peuvent à la fois être prédicteurs et variable cible.

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

Intérêt : Mieux connaître les comportements.

Méthodes : Algorithme a priori.

Exemples :

- Analyse du panier de la ménagère (si j'achète des fraises, alors j'achète des cerises).
- Étudier quelle configuration contractuelle d'un abonné d'une compagnie de téléphone portable semble être favorable à un changement d'opérateur.

14

4 : l'estimation (technique prédictive)

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 classification supervisée 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 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.

15

5 : la classification supervisée
(technique prédictive)

Principe :

C'est une estimation qui travaille sur une variable cible catégorielle.

Intérêt : Permettre l'estimation de valeurs inconnues.

Méthodes :

- Graphiques et nuages de points.
- Méthode des k plus proches voisins.
- Arbres de décision.
- Réseau de neurones.

Exemples :

- Segmentation par tranche de revenus : élevé, moyen et faible (3 segments). On cherche les caractéristiques qui conduisent à ces segments.
- Déterminer si un mode de remboursement présente un bon ou un mauvais niveau de risque crédit (deux segments).

16

6 : la prévision (technique prédictive)

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.

17

Intérêt du Datamining

On ne veut pas simplement confirmer des intuitions a priori par des requêtes dans les BD mais détecter sans a priori les combinaisons de critères les plus discriminantes

- Ex: dans le domaine commercial, on ne veut pas savoir « Combien de clients on acheté tel produit pendant telle période »

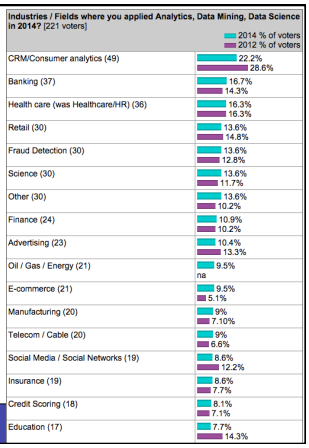
MAIS

- « Quel est leur profil ? »
- « Quels autres produits les intéresseront ? »
- « Quand seront ils intéressés »

Les profils de clientèle à découvrir sont en général des profils complexes en opposition à des profils devinables par statistiques descriptives

Utilisation du datamining

http://www.kdnuggets.com/polls/2014/industries-applied-analytics-data-mining-data-science.html



Datamining et CRM
(gestion de la relation client)

19

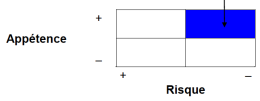
- Mieux connaître le client
  - Pour mieux le servir
  - Pour augmenter sa satisfaction
  - Pour augmenter sa fidélité
    - Il est plus couteux d'acquérir un client que de le conserver
- La connaissance du client est encore plus utile dans le secteur tertiaire
  - Les produits se ressemblent entre établissements
  - Le prix n'est pas toujours déterminant
  - Ce sont surtout le service et la relation avec le client qui font la différence

Exemple du *credit scoring*

- Objectifs de la banque :
  - vendre plus
  - en maîtrisant les risques
  - en utilisant les bons canaux au bon moment
- Le crédit à la consommation
  - un produit standard
  - concurrence des sociétés spécialisées sur le lieu de vente (Cetelem…)
  - quand la banque a connaissance du projet du client, il est déjà trop tard

Conclusion :
il faut être pro-actif
détecter les besoins des clients et leur tendance à emprunter

> Faire des propositions commerciales aux bons clients, avant qu'ils n'en fassent la demande



Le data mining dans la banque

- Naissance du score de risque en 1941 (David Durand)
- Multiples techniques appliquées à la banque de détail et la banque d'entreprise
- Surtout la banque de particuliers :
  - montants unitaires modérés
  - grand nombre de dossiers
  - dossiers relativement standards
- Essor dû à :
  - développement des nouvelles technologies
  - nouvelles attentes de qualité de service des clients
  - concurrence des nouveaux entrants (assureurs, grande distribution) et des sociétés de crédit
  - pression mondiale pour une plus grande rentabilité
  - surtout : ratio de solvabilité Bâle 2

Le data mining dans l'assurance de risque

- Des produits obligatoires (automobile, habitation) :
  - soit prendre un client à un concurrent
  - soit faire monter en gamme un client que l'on détient déjà
- D'où les sujets dominants :
  - attrition
  - ventes croisées (*cross-selling*)
  - montées en gamme (*up-selling*)
- Besoin de décisionnel dû à :
  - concurrence des nouveaux entrants (bancassurance)
  - bases clients des assureurs traditionnels mal organisées :
    - compartimentées par agent général
    - ou structurées par contrat et non par client

Le data mining dans la téléphonie

- Deux événements :
  - ouverture du monopole de France Télécom
  - arrivée à saturation du marché de la téléphonie mobile
- D'où les sujets dominants dans la téléphonie :
  - score d'attrition (*churn* = changement d'opérateur)
  - optimisation des campagnes marketing
    - *text mining* (pour analyser les lettres de réclamation)
- Problème du *churn* :
  - coût d'acquisition moyen en téléphonie mobile : 250 euros
  - plus d'un million d'utilisateurs changent chaque d'année d'opérateur
  - la loi Chatel (juin 2008) facilite le changement d'opérateur en diminuant le coût pour ceux qui ont dépassé 12 mois chez l'opérateur
  - la portabilité du numéro facilite le churn

Le data mining dans le commerce

- Vente Par Correspondance
  - utilise depuis longtemps des scores d'appétence
  - pour optimiser ses ciblages et en réduire les coûts
  - des centaines de millions de documents envoyés par an
- e-commerce
  - personnalisation des pages du site web de l'entreprise, en fonction du profil de chaque internaute
  - optimisation de la navigation sur un site web
- Grande distribution
  - analyse du ticket de caisse
  - détermination des meilleures implantations (géomarketing)

Autres exemples

- De l'∞ petit (génomique) à l'∞ grand (astrophysique pour le classement en étoile ou galaxie)
- Du plus quotidien (reconnaissance de l'écriture manuscrite sur les enveloppes) au moins quotidien (aide au pilotage aéronautique)
- Du plus ouvert (e-commerce) au plus sécuritaire (détection de la fraude dans la téléphonie mobile ou les cartes bancaires)
- Du plus industriel (contrôle qualité pour la recherche des facteurs expliquant les défauts de la production) au plus théorique (sciences humaines, biologie…)
- Du plus alimentaire (agronomie et agroalimentaire) au plus divertissant (prévisions d'audience TV)

Exemples médicaux

- Mettre en évidence des facteurs de risque ou de rémission dans certaines maladies (infarctus et des cancers) – Choisir le traitement le plus approprié – Ne pas prodiguer des soins inutiles
- Déterminer des segments de patients susceptibles d'être soumis à des protocoles thérapeutiques déterminés, chaque segment regroupant tous les patients réagissant identiquement
- Décryptage du génome
- Tests de médicaments, de cosmétiques
  - Prédire les effets sur la peau humaine de nouveaux cosmétiques, en limitant le nombre de tests sur les animaux