Fouille de données visuelles

Définition informelle :

Découverte d'informations intéressantes dans un paquet de données, en anglais Data Mining. Il est fortement lié à l'apprentissage automatique (machine learning)

Données :

- Un tableau de données :
  - N lignes : les individus, les objets d'étude
  - P colonnes : les variables, les caractéristiques des objets
- Une base de données relationnelle :
  - Des tables similaire aux tableaux
  - Des liens entre les tables
- Un entrepôt de données (data warehouse)
  - Mise en commun de bases de données
  - Agrégation de valeurs : nombre de commandes par enseigne et par mois d'un produit

Pb : données complexes, hétérogènes, évolutives et volumineuses

Information intéressantes

- Liens entre variables
  - Corrélation
  - Dépendance non linéaire
  - Capacité de prédiction
- Liens entre individus
  - Interactions significatives
  - Groupes homogènes
- Liens entre évènements
  - Co-occurrence
  - Dépendance logico-temporelle

Découverte

- Exploration :
  - L'analyste fait tout, ex : moyenne, médiane, coefficient de corrélation
  - Rapports
  - Outils visuels, ex : histogrammes, diagramme à bâtons
- Semi-automatique :
  - L'analyste guide le processus, ex : segmentation d'un ensemble de clients
  - Algorithmes d'apprentissage : inférence à partir d'exemples de résultats voulus
- Automatique :
  - Intervention minimale de l'analyste : choix d'une méthode et analyse des résultats. Ex : reconnaissance d'empreintes digitales
  - Parfois proche de l'exploration
  - Souvent presque impossible mais souhaitable

Apprentissage automatique

Définition informelle

- Observations d'un phénomène
- Construction d'un modèle de ce phénomène
- Prévisions et analyse du phénomène grâce au modèle
- Ps : le tout automatiquement : presque sans intervention humaine

Ex : observations d'un phénomène->des données zi ↔Z

- lien avec ce qui précède :
    - situation simple :zi correspond à un individu (une ligne d'un tableau)
    - situations complexes :
        1. regroupement de plusieurs zi pour former un individu
        2. structure complexe pour Z
        3. plusieurs phénomènes
        4. un phénomène par individu

Deux grands paradigmes

1. cas non supervisé :
    - pas de structure interne à z
    - classification, règles d'association, etc.
2. cas supervisé :
    - z=(x,y)
    - modélisation du lien entre x et y
    - pour faire des prévisions : connaissant x , on prédit y

Apprentissage non supervisé

- positionnement :
    - représenter des objets dans le plan (un point par objet)
    - applications : visualisation globale d'un jeu de données, analyse visuelle (groupes, corrélation, etc.)
- classification (Clustering)
    - trouver dans un ensemble d'objets des groupes homogènes (classes) et bien distincts les uns des autres
    - s'appuie sur une mesure de similarité entre objets
    - applications : typologie de clients, regroupement de gènes, regroupement de pages web
- recherche de schémas fréquents
    - trouver des groupes d'objets fréquemment ensembles
    - trouver des séquences fréquentes d'actions
    - applications : recommandations, offres marketing,etc

Apprentissage supervisé

- discrimination/classement
  - y={1,….,q} : q classes d'objets
  - prévision : placer une nouvelle observation x dans une des q classes
  - applications : diagnostic médical(malade/sain), reconnaissance de caractères,etc.
- ranking/scoring :
  - apprendre un ordre sur un ensemble d'objets
  - prévision : donner des objets intéressants (grands au sens de l'ordre) ; dire si un objet est plus intéressant qu'un autre ; donner un score d'intérêt à un objet
  - y={0,1} : 1 pour intéressant, 0 pour inintéressant
  - autre choix possibles pour y(par ex. R ou tout ensemble ordonné)
  - applications : recherche d'informations (page rank de Google), suggestions (amazon..)
- régression
- sortie structurée :
  - y est un ensemble structuré complexe : ensemble de fonctions, chaînes de caractères, arbres, graphes, etc.
  - prévision : associer un objet de l'ensemble complexe à une nouvelle observation
  - application : inférence grammaticale (associer un arbre de syntaxe à texte), traduction automatique, etc.

Démarche

1. exploration manuelle des données :
   - moyens : visualisation et statistiques
   - buts : identifier des schémas simples (corrélation, dépendances, etc.) et formuler des hypothèses associées
2. puis exploration non supervisée :
   - moyens : clustering, schémas fréquents
   - buts : identifier des schémas plus complexes (classes, etc.) et formuler des hypothèses associées
3. puis modélisation :
   - moyens : méthodes supervisées
   - buts : valider les hypothèses, prévoir, classer…

vocabulaire : classification(clustering) ; classement(classification/ranking) ; discrimination(classifiction)

le traitement des données visuelles. Nous verrons, dans le détail, différentes applications de la transformée de Fourier (théorie de l'échantillonnage, débruitage, déconvolution, interpolation, zoom,...),
quelques problèmes classiques (recalage, indexation, segmentation,...).
Acquisition et restitution de données visuelles, Méthodes de base du traitement de données visuelles statiques, échantillonnage bidimensionnel, quantification, transformation de Fourier, filtrage et prétraitement, Restauration et

rehaussement, Réduction de redondance, compression, compactage, Extraction de contour, Segmentation, Reconnaissance d'objets, Indexation et recherche par le contenu