Outils Statistiques du Data Mining

Pr Roch Giorgi

 roch.giorgi@univ-amu.fr

SESSTIM, Faculté de Médecine, Aix-Marseille Université, Marseille, France
http://sesstim.univ-amu.fr

Introduction (1)

- Data Mining
  - ✓ Prospection ou fouille de données
- Objectif
  - ✓ Valorisation d'une grande base de données
  - ✓ Valorisation d'un entrepôt de données (data warehouse)
  - ✓ Pour la recherche d'informations pertinentes pour l'aide à la décision
- Contexte
  - ✓ Médical : facteurs étiologiques, pronostiques, médico-économiques, génomiques, …
  - ✓ Industriel, astrophysique, …

Introduction (2)

- Techniques utilisées pour représenter, analyser plus simplement les relations entre
  - ✓ 1 variable à expliquer par rapport à plusieurs variables explicatives
  - ✓ Des variables entre elles
- Existence de logiciels « clé en main » (www.kdnuggets.com)
- Trouvent toujours une réponse à la question : « *Comment trouver un diamant dans un tas de charbon sans se salir les mains ?* »
- Nécessité de connaître la démarche d'analyse, les principales méthodes, leurs bases méthodologiques

Méthodes Statistiques Utilisées (1)

- Statistique descriptive unifactorielle
  - ✓ Fonction du type de variable (qualitative, quantitative)
  - ✓ Statistiques de position (moyenne, médiane, …) ou de dispersion (variance, étendue, …)
  - ✓ Représentations graphiques (histogramme, box-plot, …)
- Statistique descriptive multifactorielle
  - ✓ Entre 2 variables (fonction du type)
    - Nuage de point, covariance, corrélation, …
    - Box-plot parrallèles, rapport de corrélation, …
    - Tableau de contingence, chi-deux, …
  - ✓ Entre plusieurs variables
    - Matrices des covariances, des corrélations
    - Tableaux de nuages, …

Méthodes Statistiques Utilisées (2)

- Modélisation multifactorielle
  - ✓ Contexte de l'estimation de paramètres associés aux variables étudiées
  - ✓ Régression logistique
    - Variable à expliquer binaire
    - Variables explicatives quantitatives ou qualitatives
  - ✓ Régression multiple
    - Variable à expliquer quantitative
    - Variables explicatives quantitatives ou qualitatives

Méthodes Statistiques Utilisées (3)

- Analyse factorielle
  - ✓ Principe de réduction d'un ensemble variable formant un espace de dimension N à un espace de dimension réduit
  - ✓ Analyse en composantes principales (ACP)
    - Pour des variables quantitatives ou ordinales
    - Pour réduire le nombre de variables en tenant compte de la variance totale
  - ✓ Analyse factorielle (AF)
    - Pour des variables quantitatives ou ordinales
    - Pour expliquer la variance commune entre les variables
  - ✓ Analyse des correspondances (AC)
    - Pour des variables qualitatives

Méthodes Statistiques Utilisées (4)

- Méthodes de classification
  - ✓ Basée sur la recherche d'une partition des individus en classes homogènes
  - ✓ Classification hiérarchique
    - Pour des variables quantitatives ou qualitatives

Méthodes Statistiques Utilisées (…)

- Intelligence artificielle
  - ✓ Réseaux neuronaux
  - ✓ Reconnaissance de formes

Les Différentes Étapes

1. Compréhension du domaine d'application
2. Création du sous-ensemble cible de données
3. Nettoyage des données (erreurs, données manquantes, valeurs atypiques)
4. Transformation des données (normalisation, linéarisation, découpage en classes, compression)
5. Explicitation de l'objectif et de la stratégie d'analyse
6. Choix des méthodes
7. Test, en précisant les critères
8. Exploitation
9. Diffusion

DATA MINING

Objectifs des Méthodes Statistiques

- Exploration
  - ✓ Statistique descriptive unifactorielle, multifactorielle
- Modélisation
  - ✓ Régression logistique, régression multiple, ACP, AF, AC
- Classification
  - ✓ Classification hiérarchique
- Recherche de formes
  - ✓ Réseaux neuronaux

Data Mining et Données

- Souvent préalables à l'étude
- Peuvent avoir été recueillies à d'autres fins
- Volume souvent important (nombre de variables et d'observations)
  - ✓ Traitement exhaustif : problème algorithmique possible
  - ✓ Traitement après sondage : perte de l'information pertinente possible si elle concerne des groupes de faible effectif
- La taille des données influe sur le choix des méthodes
  - ✓ Exemple : le nombre de paramètres que l'on peut estimer en régression augmente avec la taille de l'échantillon

Choix d'une Méthode

- Il n'y a pas de meilleure méthode
- Ont une certaine robustesse par rapport à leur propriétés intrinsèques et leurs hypothèses de base

- Essayer une méthode de chaque grande famille
- Comparer les résultats obtenus entre eux

Choix d'un Modèle

- Choix des variables à analyser
- Choix des éventuelles interactions à tester
- Critères spécifiques aux méthodes (nombre de composantes, nombre de feuilles d'un arbre de décision, …)
- Objectifs
  - ✓ Minimiser les erreurs de classement, de prévision
  - ✓ Sélection d'un modèle parcimonieux : compromis entre l'ajustement aux données et la variance des estimations des paramètres pour améliorer la qualité des prédictions

Démarche de Choix d'un Modèle

- Échantillonnage aléatoire de l'échantillon global (N) en 3 parties
- Phase d'apprentissage (n_1/N)
  - ✓ Estimation du modèle
- Phase de validation (n_2/N)
  - ✓ Optimisation du modèle
- Phase de test (n_3/N)
  - ✓ Test de l'adéquation du modèle aux données

Automatisation

- Logiciels de fouille de données
- Ergonomie simplifie le travail
- Automatisation toujours tentante mais …
- Intervention analytique humaine indispensable pour
  - ✓ Vérifier intégrité et cohérence des données
  - ✓ Traiter les données manquantes
  - ✓ Transformer les données
  - ✓ Choisir des critères d'estimation propres aux méthodes
  - ✓ Choisir le modèle final (un modèle sera toujours trouvé mais est-il adapté, interprétable, …)

Sources

- Besse P, Le Gall C, Raimbault N, Sarpy S. Data Mining et Statistique. *Journal de la Société Française de Statistique* 2001;142:5-36.
- http://www.lsp.ups-tlse.fr/Besse/enseignement.html
- www.kdnuggets.com