Fouille de données

Christel Vrain

{Christel.Vrain@lifo.univ-orleans.fr}.

Contraintes et Apprentissage

LIFO (FRE 2490)

Université d'Orléans

Plan de l'exposé

1. Présentation de la fouille de données

2. Les travaux menés au LIFO

   (a) Programmation Logique Inductive et Classification
   (b) Recherche de règles d'association
   (c) Recherche de règles de caractérisation
   (d) Découverte de classes
   (e) Fouille de textes

Extraction de Connaissances dans les Bases de Données (ECD)
Knowledge Discovery in Databases (KDD)

ECD

extraction non triviale de connaissances implicites, inconnues au préalable, intéressantes, utiles à partir d'informations stockées dans des bases de données

ECD

extraction non triviale de connaissances implicites, inconnues au préalable, intéressantes, utiles à partir d'informations stockées dans des bases de données

- Quels types de connaissances ?
  - lois numériques
  - règles d'association
  - règles de classification
  - ...

  ← dépend de la tâche à résoudre

ECD

extraction non triviale de connaissances implicites, inconnues au préalable, intéressantes, utiles à partir d'informations stockées dans des bases de données

- Quels types de connaissances ?
- Champ de recherche multidisciplinaire
  - Bases de Données (BD)
  - Statistiques
  - Apprentissage Statistique / Symbolique
  - Visualisation

Pourquoi cette émergence ?

un univers de données en plein expansion, mais peu de connaissances

- de nombreuses bases de données, de plus en plus volumineuses
- connexion des BDs par les réseaux
- Internet

$\rightarrow$ besoin “vital” d’outils d’aide à la décision pour interpréter les données

- développement des techniques d’apprentissage automatique

Différentes tâches

Apprentissage supervisé

- Description de classes ou concepts
  - caractérisation
  - discrimination

→ apprentissage à partir d'exemples positifs

Différentes tâches

Apprentissage supervisé

- Description de classes ou concepts
- Classification

→ apprentissage à partir d'exemples positifs et négatifs

Différentes tâches

Apprentissage supervisé

- Description de classes ou concepts
- Classification
- Régression

Différentes tâches

Apprentissage supervisé

- Description de classes ou concepts
- Classification
- Régression
- Prédiction

Différentes tâches

Apprentissage supervisé

- Description de classes ou concepts
- Classification
- Régression
- Prédiction

Apprentissage non supervisé

- Analyse d'associations

Différentes tâches

Apprentissage supervisé

- Description de classes ou concepts
- Classification
- Régression
- Prédiction

Apprentissage non supervisé

- Analyse d'associations
- Découverte de classes

Différentes tâches

Apprentissage supervisé

- Description de classes ou concepts
- Classification
- Régression
- Prédiction

Apprentissage non supervisé

- Analyse d'associations
- Découverte de classes
- Organisation en hiérarchies

Différentes tâches

Apprentissage supervisé

- Description de classes ou concepts
- Classification
- Régression
- Prédiction

Apprentissage non supervisé

- Analyse d'associations
- Découverte de classes
- Organisation en hiérarchies
- Recherche des anomalies

Des exemples d'applications

- Analyse du panier de la ménagère
- Fidélisation des clients
- Publicité ciblée
- Détection des fraudes

Le processus d'ECD

- Prétraitement des données
  - Nettoyage
  - Sélection des attributs pertinents
  - Construction d'attributs synthétisés
  - Aplatissement de la base de données / *Prise en compte du caractère relationnel* de la base
- Fouille de données
  - Apprentissage Automatique
  - Apprentissage Statistique
  - Analyse de Données
- Interprétation et Validation des résultats

Fondements formels : Programmation Logique Inductive

Programmation Logique Inductive

- À la frontière entre l'Apprentissage Symbolique Automatique et la Programmation Logique
  - représentation des connaissances sous une forme *compréhensible* pour l'expert

    $\Rightarrow$ formalisme à base de règles

    $$\hat{a}ge = moyen \wedge ens\_sup \rightarrow utilise\_internet$$
  - représentation en logique du 1er ordre

    $$grand\_père(X, Y) \leftarrow père(X, Z), père(Z, Y)$$

Programmation Logique Inductive

- À la frontière entre l'Apprentissage Symbolique Automatique et la Programmation Logique
  - représentation des connaissances sous une forme *compréhensible* pour l'expert

    $\Rightarrow$ formalisme à base de règles

    $$\hat{a}ge = moyen \wedge ens\_sup \rightarrow utilise\_internet$$
  - représentation en logique du 1er ordre

    $$grand\_père(X, Y) \leftarrow père(X, Z), père(Z, Y)$$

- Induction de programmes logiques / Apprentissage de connaissances "expertes"
  - traitement de données numériques et bruitées v
  - volume de données

Tâche de classification - Exemple 1

Bases de données - Exemples positifs et négatifs

$Personne$			$Père$		$positif$		$négatif$	
Id	Nom	$Prénom$	Id_1	Id_2	Id_1	Id_2	Id_1	Id_2
1	$Dupont$	$Jean$	1	2	1	3	3	1
2	$Dupont$	$Paul$	2	3	1	4	1	2
3	$Dupont$	$Marie$	2	4				
4	$Dupont$	$Antoine$						

Connaissances apprises

$$grand\_père(X, Y) \leftarrow père(X, Z), père(Z, Y)$$

Tâche de classification - Exemple 2

Connaissances du domaine	Concept cible	
	positifs	négatifs
$z\acute{e}ro(0)$, $succ(0,1)$ $succ(1,2)$, $succ(2,3)$	$pair(0)$, $pair(2)$ $impair(1)$, $impair(3)$	$le\ reste$

$\Rightarrow$ trouver un programme logique $\mathcal{P}$ définissant $pair$ et $impair$

$\Rightarrow$ programme correct

$$pair(X) \leftarrow succ(X,Y), impair(Y)$$
$$impair(X) \leftarrow \neg pair(X)$$

□

Principe

recherche d'une "bonne" hypothèse dans l'espace des hypothèses

$\longrightarrow$ relation de généralité

- Stratégies de recherche
  - exploration déterministe : critères statistiques
    $\longrightarrow$ ICN, MULT_ICN *[Martin, Vrain]*
  - exploration stochastique
    $\longrightarrow$ programmation génétique *[Martin, Moal, Vrain]*
    $\longrightarrow$ algorithme génétique $\longrightarrow$ GRIL *[Braud, Vrain]*
- Biais syntaxiques
  $\longrightarrow$ modélisation des biais par une grammaire d'arbres *[Moal]*

Limitations et extensions

- Limitation à des programmes Datalog
  - traduction des symboles de fonction en symboles de prédicats

    fonction $successeur$ *représentée par* $succ(X, Y)$
  - spécification d'un ensemble fini de constantes,
    $\Rightarrow$ Les domaines numériques doivent être bornées.

    Si $D = \{0, 1, 2, 3\}$*, que vaut* $succ(3, ?)$
  - Les expressions ne sont pas évaluées.

    $$pair(s(s(X)) \leftarrow pair(X)$$

$\Rightarrow$ représentation inadéquate des données numériques

Limitations et extensions

- Limitation à des programmes Datalog
- Extensions

  $\rightarrow$ Programmation Logique avec Contraintes *[Martin, Vrain]*

  - un langage pour exprimer les contraintes
  - un domaine de calcul

$$taxe(P, Y) \leftarrow Y = 0.186 * X, prix(P, X)$$

Limitations et extensions

- Limitation à des programmes Datalog
- Extensions

  → Programmation Logique avec Contraintes *[Martin, Vrain]*

  → Bases de Données Contraintes *[Turmeaux, Vrain]*

Limitations et extensions

- Limitation à des programmes Datalog
- Extensions
  - → Programmation Logique avec Contraintes *[Martin, Vrain]*
  - → Bases de Données Contraintes *[Turmeaux, Vrain]*
  - → Applications à la Fouille de Données
    - ← capacité à traiter plusieurs relations
      - méta-informations disponibles : types, contraintes d'intégrité, ...
      - taille importante des données : place mémoire, coût du test de couverture, ...

Limitations et extensions

- Limitation à des programmes Datalog
- Extensions
  - → Programmation Logique avec Contraintes *[Martin, Vrain]*
  - → Bases de Données Contraintes *[Turmeaux, Vrain]*
  - → Applications à la Fouille de Données
  - Domaines visés :
    - Bases de Données Relationnelles
    - Systèmes d'Information Géographiques

Travaux Actuels

Personnes

- Maître de conférences
  - Sylvie Billot
  - Matthieu Exbrayat
  - Lionel Martin
  - Frédéric Moal

- ATER
  - Agnès Braud
  - Ansaf Salleb

- Doctorants
  - Andrei Letchnenko
  - Guillaume Cleuziou
  - Teddy Turmeaux

Travaux actuels

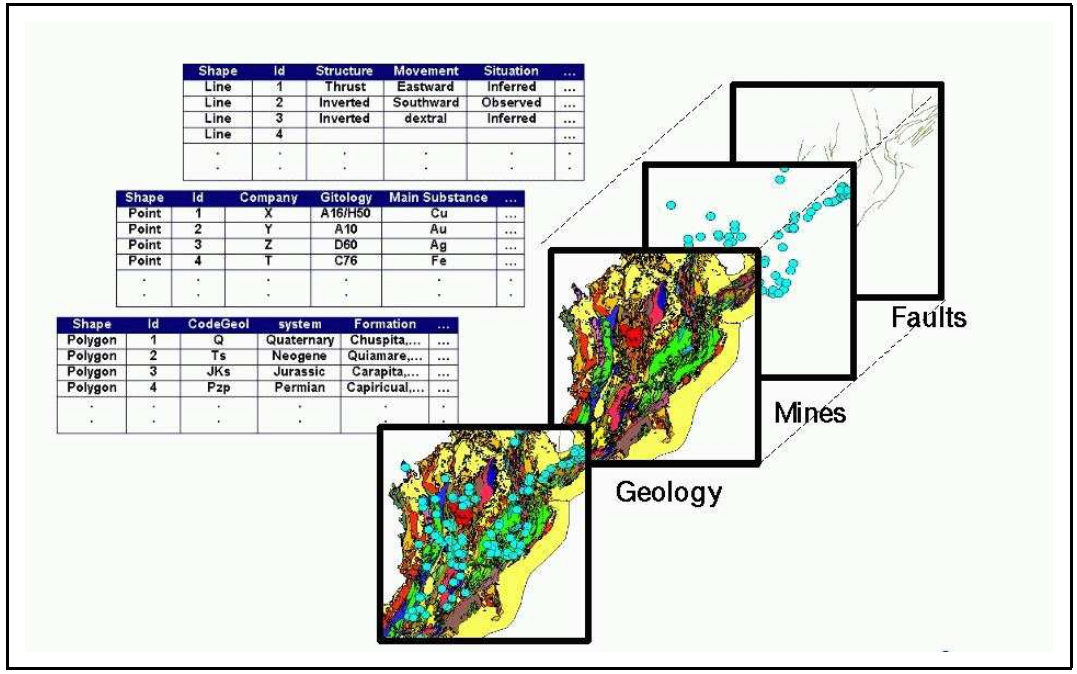
→ apprentissage et BD relationnelle :
complexité *[Exbrayat, Braud, Turmeaux]*
→ approche à base de distances *[Martin, Moal]*

- Apprentissage supervisé
  - Règles de classification
  - Règles de caractérisation *[Turmeaux, Salleb, Vrain]*
- Apprentissage non supervisé
  - Règles d'association *[Salleb, Vrain]*
  - Regroupement (Clustering) *[Cleuziou, Martin, Vrain]*
    → recherche de classes non disjointes
  - BD transactionnelles et chargement en mémoire
    [Maazouzi, Salleb, Vrain]

Applications

- collaboration avec le BRGM *[D. Cassard]*
- données athérosclérose (PKDD challenge)
- apprentissage de solveurs *[Letchnenko & al.]*
- applications aux textes *[Billot, Cleuziou, Martin, Vrain]*

Collaboration avec le BRGM



- rechercher des règles d'association entre les dépôts de minerais, les mines, les failles, les volcans . . .
- caractériser des dépôts de minerais □

Fouille de textes

- ACI BIOTIM : IRD, INRIA (ATOLL ET IMEDIA), CEDRIC, INRA

  $\Rightarrow$ méthodes génériques d'exploration de masses de données contenant textes et images pour acquérir la sur-couche sémantique commune

  $\Rightarrow$ développer des méthodes génériques d'interrogation pluri-modale des données.

- plate-forme CORAL - LIFO *[Billot, Clavier & al.]*

□

Conclusion

- Travaux formels : complétude, correction, complexité des algorithmes
- ECD
  - Tâches :
    - Classification
    - Caractérisation
    - Recherche de règles d'associations
    - Découverte de classes non disjointes
  - Types de bases de données :
    - relationnelles, géographiques
    - textes
    - transactionnelles
  - Applications : BRGM, IRD, . . .

Exemple de Bases de Données Contraintes

$Ville$			$Forêt$
Nom	$Surface$	$Population$	$Surface$
$Paris$	$3x - 4y \leq 2$ $x \leq 10$ $x - y \geq -3$ $x + 3y \leq 37$ $x \geq 6$ $3x + 4y \geq 46$	2000000	$2x \leq 19$ $-3x - 4y \leq 2$ $2x - 6y \geq -17$
$Rocquencourt$	$y \geq 0$ $x \geq 8$ $2x + 2y \leq 17$	3877	$2x - 6y \geq -17$ $2y \geq 11$ $2x - 4y \leq 5$
$Orsay$	$2x - 2y \geq -13$ $2x + 2y \leq 17$ $2x \geq 1$	15000	$x \geq 9$ $9x - y \leq 85$ $x - y \leq 10$

Opérations algébriques

- *jointure naturelle* de deux relations R_1 et R_2 :

  Exemple : $Ville \bowtie Forêt$

- *sélection* de R sur une contrainte c :

  Exemple : $\sigma_{Population \geq 20000}(Ville)$

- *projection* de R sur $\tilde{Z}$, ($\tilde{Z} \subseteq Var(R)$)

  Exemple : $\Pi_{X_O}(Ville)$

- *union* de deux relations : $R_1 \cup R_2 = \{t | t \in R_1 \ or \ t \in R_2\}$

□

Exemple d'apprentissage

Nom	$Classe$	$Définitions\ possibles$
$Orsay$	$(+)$	$\Pi_{Name}(Ville \bowtie Forêt)$
$Paris$	$(+)$	$\Pi_{Name}(\sigma_{Population>3877}(Ville)$
$Rocquencourt$	$(-)$	

□

Intérêts

- un cadre unifié pour représenter les domaines numériques et symboliques
- permet de représenter des relations avec des tuples infinis
- permet de représenter des ensembles infinis d'exemples

$\Rightarrow$ extension de l'ILP

- implémentation ouverte :
  définition de nouvelles stratégies (modèle, stochastiaue ...)
- Problèmes : données tests ?

Règles d'association

- Recherche des propriétés fréquentes

→ étape coûteuse en temps de calcul et en espace mémoire

→ élagage

- Recherche des règles (seuil de confiance)

→ génération de beaucoup de règles

Règles d'association

- Règles statistiques

$$Mine(x) \wedge Gitologie(x, A) \rightarrow Gitologie(x, A_1)(92, 12\%)$$

- Règles de contrôle

$$Mine(x) \wedge Gitologie(x, H12)$$
$$\rightarrow$$
$$Substance\_principale(x, Au)(89, 32\%)$$

- Nouvelles règles

$$Mine(x) \wedge Faille(z) \wedge Gitologie(x, C5) \wedge Proche\_de(x, z)$$
$$\rightarrow$$
$$Structure(x, Strike\_slip)(43, 75\%)$$

Caractérisation

tâche descriptive de fouille de données

- ciblé sur un ensemble d'exemples positifs
- ne nécessite pas d'exemples négatifs

$\Rightarrow$ un *cadre général* pour la caractérisation d'un ensemble d'objets, *ensemble cible*, à partir

- des propriétés des objets cibles
- des propriétés des objets liés

Applications

- bases de données géographiques
- bases de données relationnelles

Règle caractéristique

- *Règle caractéristique* $\delta :: p$: conjonction d'un schéma caractéristique δ et d'une propriété *p*

  $\forall M \quad Profondeur\_Beniof f(M) \in [75..150]$
  $\forall M \exists G \quad Mine(M) \wedge Geologie(G) \wedge Age(G, tertiaire)$

- Utilisation d'agrégats : $Agrégat_{proche}(V, count) \geq 2$

- Recherche pour chaque type de schémas des propriétés p vérifiées par au moins ϵ, ϵ seuil donné

Elagage

- *Règle caractéristique intéressante*

  $\Rightarrow$ caractère contrastant de p entre $\mathcal{E}_{target}$ et $\mathcal{E} - \mathcal{E}_{target}$

- *Relation de généralité* entre règles

$$\forall M \forall_{3Km} F \succeq \forall M \forall_{5Km} F \succeq \forall M \forall_{10Km} F$$
$$\forall M \exists_{10Km} F \succeq \forall M \exists_{5Km} F \succeq \forall M \exists_{3Km} F$$

Propriété :

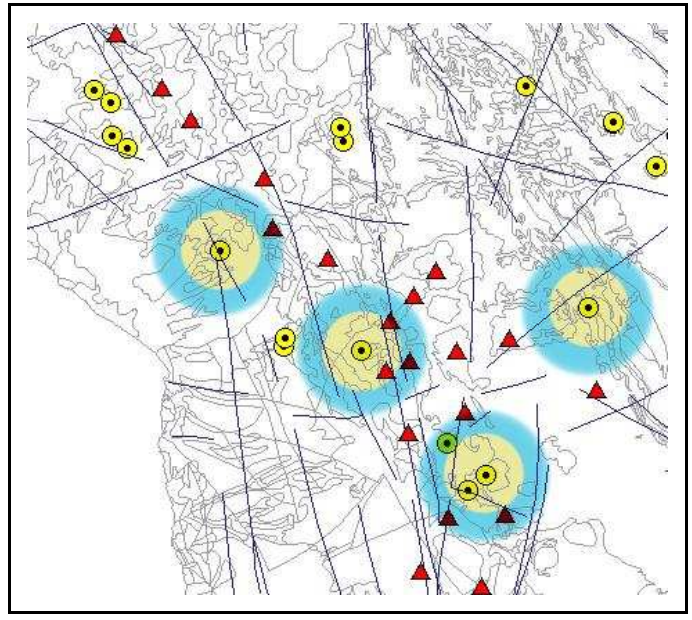
$$couverture(\delta_2, p, \mathcal{E}_{target}) \geq \epsilon$$
$$\Rightarrow$$
$$couverture(\delta_1, p, \mathcal{E}_{target}) \geq \epsilon$$

Application

Entrées :

- *mines*, *géologie*, *volcans*, *failles*, *séismes*
- relations de distance entre objets

But : caractériser les mines d'or



$\Rightarrow$ construction de *buffers croissants* autour des cibles

Résultat

Un exemple de règle obtenue couvrant près de 60% des mines d'or et rejetant la majeure partie des autres mines.

$$\begin{aligned}\forall M \exists_{10km} G :: \quad & Mine(M) \wedge Geologie(G) \wedge \\ & Substance(M, or) \wedge \\ & Profondeur\_Benioff(M) \in [75..150] \wedge \\ & Distance\_Benioff(M) \in [170..275] \wedge \\ & Pente(M) \in [8°..16°] \wedge \\ & Age(G, tertiaire) \wedge \\ & Lithologie(M, volcanique) \wedge \\ & Gitologie(M, épithermale) \wedge \\ & Morphologie(M, veines)\end{aligned}$$