

Fouille de Données et Treillis de Galois

Engelbert MEPHU NGUIFO
CRIL – IUT de Lens
mephu@cril.univ-artois.fr

20/01/2004 - EMN

Tutoriel EGC'04 - Clermont-Ferrand

Contributeurs

- Patrick Njiwoua
- Huaiguo Fu
- Huaiyu Fu
- Sadok Ben Yahia

20/01/2004 - EMN

Tutoriel EGC'04 - Clermont-Ferrand

Bibliographie

- *Data Mining: concepts and techniques*,
 - Han & Kamber, chez Morgan Kaufmann Pubs., 2001
- *Mastering Data Mining*,
 - Berry & Linoff, chez Wiley Computer Publishing, 2000
- *Advances in Knowledge Discovery and Data Mining*,
 - Fayyad & al., chez AAAI/MIT Press, 1996

- *Apprentissage Artificiel: concepts et algorithmes*
 - Cornuéjols & Miclet, chez Eyrolles, 2002
- *Machine Learning*
 - Mitchell, chez McGraw-Hill, 1997

- *Formal Concept Analysis: Mathematical Foundations*
 - Ganter & Wille, chez Springer Verlag, 1999

(voir aussi Annonce Tutoriel sur site conférence)

Agenda

- **Introduction**
- Algorithme de génération des concepts
- Règles d'association et TG
- Classification Supervisée et TG
- Conclusions

Introduction – FD - Définition

- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information (knowledge) or patterns from data in large databases or other information repositories

[Fayyad et al., 1996]

'Knowledge Discovery in Databases' (KDD) or 'Data Mining' (DM)

- Processus interactif et itératif d'analyse d'un **grand ensemble de données brutes** afin d'en extraire des connaissances exploitables par l'utilisateur-analyste qui y joue un rôle central

[Kodratoff, Napoli, Zighed, dans Bulletin AFIA 2001 sur ECBD]

ECBD ou encore 'Fouille de données'

Introduction – FD - Discipline

- Plusieurs découvertes scientifiques concerne l'ECBD
 - Loi de Kepler,
 - Lois de Newton,
 - Table périodique des éléments chimiques,
 - ...,
- Disciplines dédiées à l'analyse de données
 - Statistiques
 - Apprentissage Automatique
- Pourquoi la FD ?

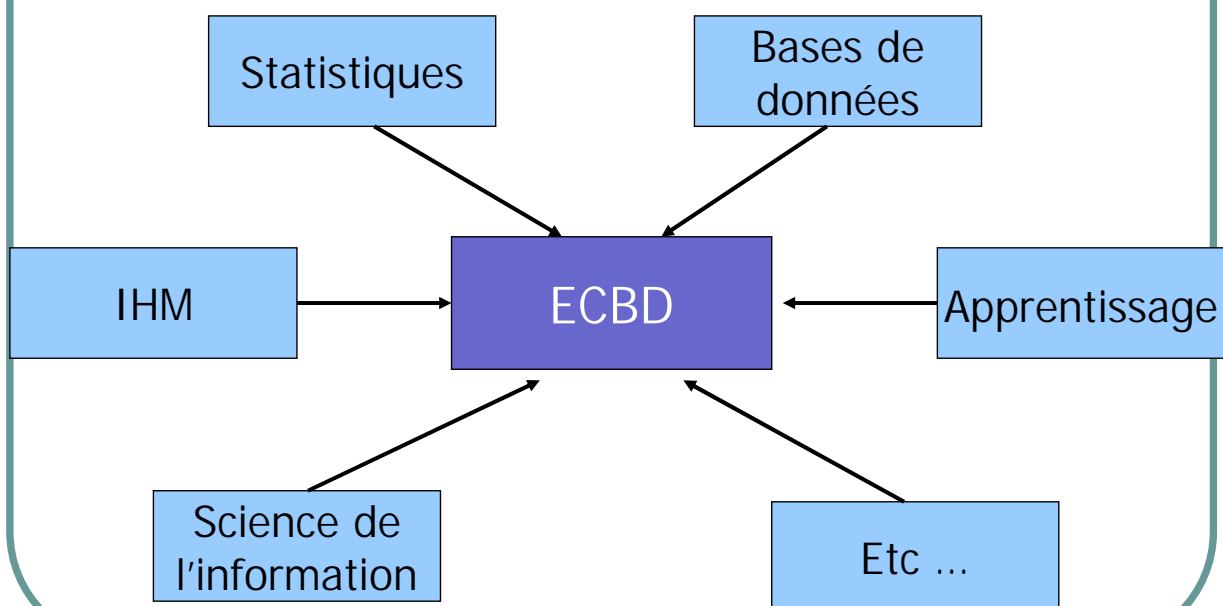
Introduction – FD - Discipline

- Pourquoi ECBD? Quelles sont les différences?
 - **Données de taille volumineuse** - du giga au tera octets
 - Ordinateur rapide- réponse instantanée, analyse interactive
 - Analyse multidimensionnelle, puissante et approfondie
 - Langage de haut niveau, “déclaratif” – Facilité d’usage et Contrôlable
 - Automatisée or semi- automatisée —fonctions de fouille de données cachées ou intégrées dans plusieurs systèmes

20/01/2004 - EMN

Tutoriel EGC'04 - Clermont-Ferrand

Introduction – FD - Discipline



20/01/2004 - EMN

Tutoriel EGC'04 - Clermont-Ferrand

Introduction – FD - Etapes

- Nettoyage et Intégration de bases de données
 - Suppression données inconsistantes ou combinaison de données de différentes sources pour constituer un entrepôt
- Prétraitement de données
 - Selection ou transformation de données de l'entrepôt pour les rendre exploitables
- Fouille de données
 - Utilisation de méthodes intelligentes pour extraire des motifs.
 - Tâches: Caractérisation, discrimination, **association**, **classification**, prédiction, analyse de données évolutives
- Evaluation et Présentation
 - Identifier les motifs intéressants, les visualiser, et interagir

Introduction – TG - Préliminaires

- ou Treillis de Concepts
- En anglais: Concept or Galois Lattices
- Travaux :
 - Birkhoff : 1940, 1973
 - Barbut & Monjardet: 1970
 - Wille: 1982 → FCA (ou AFC ou ACF)
 - Chein, Norris, Ganter, Bordat, ...
 - Diday, Duquenne, ...

Introduction – TG - Définition

- Structure mathématique
- Notions de base:
 - Contexte
 - Correspondance de Galois
 - Concept
 - Ordre

Introduction – TG - Définition

- Contexte = triplet (O, A, I) tel que:
 - O : ensemble fini d'exemples
 - A : ensemble fini d'attributs
 - I : relation binaire entre O et A , $(I \subseteq O \times A)$
- 2 exemples :

$O \setminus A$	a	b	c
1	1	1	1
2	1	1	
3		1	1

	a	b	c	d	e	f	g	h
1	1	1	1	1	1	1	1	
2	1	1	1	1	1	1		1
3	1	1	1	1	1		1	1
4	1	1	1	1		1		
5	1	1		1	1		1	
6	1	1	1		1			
7	1		1			1		

Introduction – TG - Définition

- Correspondance de Galois:
 - Soient $O_i \subseteq O$ et $A_i \subseteq A$, on définit f et g comme suit:
 - $f : P(O) \rightarrow P(A)$ $f(O_i) = \{ a \in A / (o,a) \in I, \forall o \in O_i \}$ **intension**
 - $g : P(A) \rightarrow P(O)$ $g(A_i) = \{ o \in O / (o,a) \in I, \forall a \in A_i \}$ **extension**
 - f et g sont 2 applications monotones décroissantes

(f,g) = correspondance de Galois entre $P(O)$ et $P(A)$.

- Soient $h = g \circ f$ et $h' = f \circ g$, elles sont:
 - isotones (monotones croissantes): $O_1 \subseteq O_2 \Rightarrow h(O_1) \subseteq h(O_2)$
 - extensives $O_1 \subseteq h(O_1)$
 - idempotentes $h(O_1) = h \circ h(O_1)$

h (resp. h') est une fermeture dans $P(O)$ (resp. $P(A)$)

Introduction – TG - Définition

- Concept (fermé, rectangle):
 - Soient $O_i \subseteq O$ et $A_i \subseteq A$, (O_i, A_i) est un **concept** ssi
 - O_i est l'extension de A_i **et**
 - A_i est l'intension O_i
 - c-à-d: $O_i = g(A_i)$ **et** $A_i = f(O_i)$

Soit $L = \{ (O_i, A_i) \in P(O) \times P(A) / O_i = g(A_i) \text{ et } A_i = f(O_i) \}$ ens. concepts

- Relation d'ordre (\leq) sur L :
 - Sous-concept / Sur-concept (spécialisation / généralisation)
 - $(O_1, A_1) \leq (O_2, A_2)$ si et seulement si $O_1 \subseteq O_2$ (ou $A_1 \supseteq A_2$)
- Treillis de Galois
 - $T = (L, \leq)$, ens. des concepts muni de la relation d'ordre

Introduction – TG - Exemple

- Correspondance de Galois: Exemple

- $O1 = \{6, 7\} \Rightarrow f(O1) = \{a, c\}$
- $A1 = \{a, c\} \Rightarrow g(A1) = \{1, 2, 3, 4, 6, 7\}$

	a	b	c	d	e	f	g	h
1	1	1	1	1	1	1	1	
2	1	1	1	1	1	1		1
3	1	1	1	1	1		1	1
4	1	1	1	1		1		
5	1	1		1	1		1	
6	1	1	1		1			
7	1		1			1		

Remarque:

$$h(O1) = g \cdot f(O1) = g(A1) \neq O1$$

Introduction – TG - Exemple

- Concept: Exemple

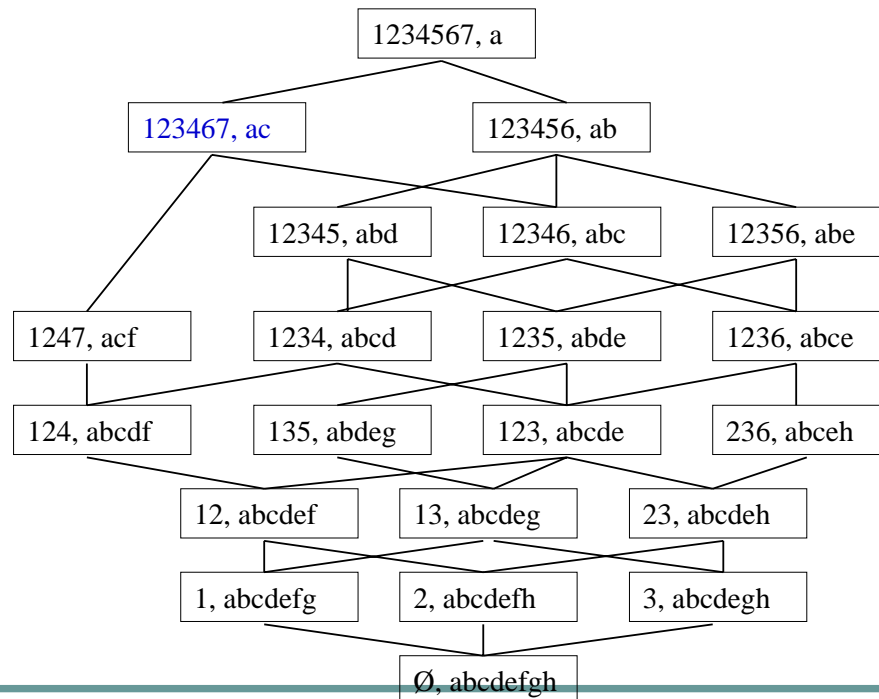
- $O1 = \{6, 7\} \Rightarrow f(O1) = \{a, c\}$
- $A1 = \{a, c\} \Rightarrow g(A1) = \{1, 2, 3, 4, 6, 7\}$

- *Remarque:* $h(O1) = g \cdot f(O1) = g(A1) \neq O1$

- $(\{6, 7\}, \{a, c\}) \notin L$
- $(\{1, 2, 3, 4, 6, 7\}, \{a, c\}) \in L$

	a	b	c	d	e	f	g	h
1	1	1	1	1	1	1	1	
2	1	1	1	1	1	1		1
3	1	1	1	1	1		1	1
4	1	1	1	1		1		
5	1	1		1	1		1	
6	1	1	1		1			
7	1		1			1		

Introduction – TG - Définition



20/01/2004 - EMN

Tutoriel EGC'04 - Clermont-Ferrand

Agenda

- Introduction
- **Algorithme de génération de concepts**
- Règles d'association et TG
- Classification Supervisée et TG
- Conclusions

20/01/2004 - EMN

Tutoriel EGC'04 - Clermont-Ferrand

Algorithmes TG

- 1969 ... à 2003 → problème ouvert
- Points clés
 - Génération des concepts
 - Génération du treillis
 - Incrémentalité
 - Partitionnement de données
 - Contexte vs Transposé Contexte
- Ordre de complexité ?
 - Fonction de la taille de données, mais aussi du **contenu du Contexte**

Algorithmes TG

- Non Incrémental
 - Chein, 1969 ;
 - Ganter, 1984 ;
 - Bordat, 1986 ;
 -
 - Nourine et Raynaud, 1999
 -
- Incrémental
 - Norris, 1978 ;
 - Godin et al., 1991 ;
 - Oosthuisen, 1991 (pseudo-TG) ;
 - Carpineto et Romano, 1993;
 -

Algorithmes TG – Non Incrém.

- Principe:
 - $T_k = \text{fonction } (C, o_1, o_2, \dots, o_k)$
 - Partir toujours du contexte formel pour générer les concepts, et éventuellement le TG
- Construction du graphe de Hasse
 - Pendant la génération des concepts
 - ou Après

Algorithmes TG – Non Incrém.

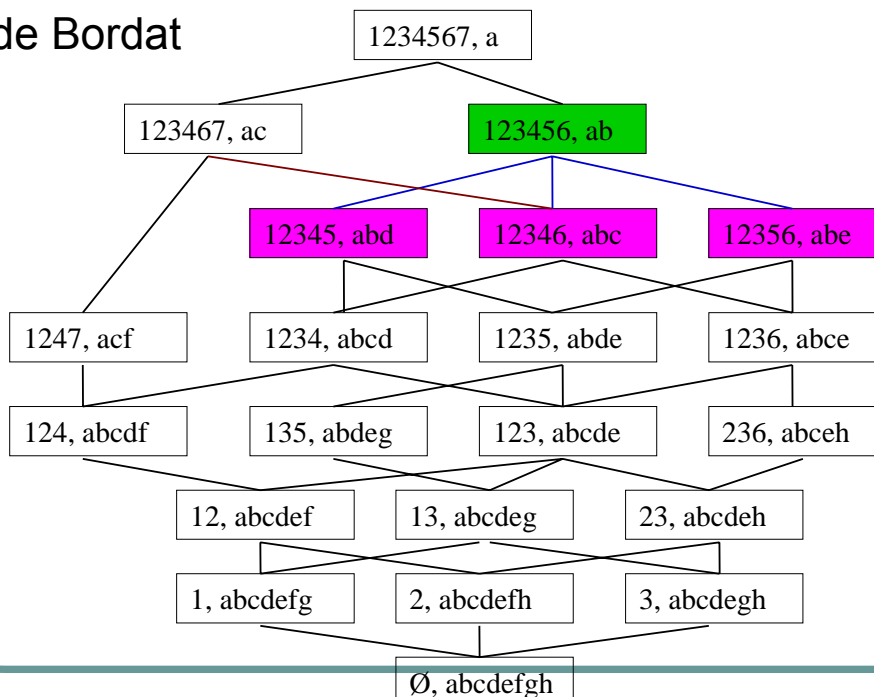
- Algorithme Bordat -- 1986 Math Info Sci Hum.
 - Approche de génération structurée (descendante, par niveau)
 - Approche par spécialisation/généralisation
 - Principe s'appuyant sur la relation de couverture de la rel. d'ordre \leq
 - Couverture d'un concept (O_i, A_i) , notée $\underline{(O_i, A_i)}$
Ens des (O_j, A_j) tq - $(O_j, A_j) \leq (O_i, A_i)$ et
- $\exists (O_k, A_k)$ tq $(O_j, A_j) \leq (O_k, A_k) \leq (O_i, A_i)$

Algorithmes TG – Non Incrém.

- Algorithme de Bordat :
 - L = (O, f(O))
 - Pour chaque concept (O_i,A_i) de L
 - Rechercher couverture C = (O_i,A_i)
 - Pour chaque (O_j,A_j) ∈ C
 - Si (O_j,A_j) ∉ L alors ajouter (O_j,A_j) à L
 - Sinon rajouter un arc seulement
 - Fin Pour
 - Fin Pour
- **Inconvénient:** Concept engendré autant de fois qu'il a de sur-concepts
- **Avantage:** Enumération des arêtes du graphe de Hasse du treillis

Algorithmes TG – Non Incrém.

Algo. de Bordat



Algorithmes TG – Non Incrém.

- Autres algorithmes
 - Chein, 1969 ;
 - Ganter, 1984 ; NextClosure
 - .
 - Zabezhalo et al.(1987),
 - Kuznetsov(1993),
 - Lindig(1999);
 - Nourine et Raynaud, 1999, IPL
 - Stumme et al., 2000
 - Nourine et Raynaud, 2002, JETAI
 - Fu et Mephu, 2003, ScalingNextClosure, ICTAI
 -

Algorithmes TG – Incrémental

- Principe:
 - $T_k = \text{fonction}(T_{k-1}, o_k)$
 - A l'étape k, générer le TG à partir du TG engendré à l'étape k-1, et du nouvel objet o_k .
- Construction du graphe de hasse du TG en même temps que la génération des concepts

Algorithmes TG – Incrémental

- Algo de Godin:
 1. Produire toutes les intersections entre la description du nouvel objet, $f(ok+1)$, d'une part, et les intensions existantes, d'autre part, nécessaire pour "fermer" la famille de Moore des intensions du treillis initial à laquelle un nouveau fermé, $f(ok+1)$, est ajouté ;
 2. Par cette intersection avec $f(ok+1)$, se produisent deux types d'ensembles d'attributs Y :
 1. Y déjà fermé ou intension existante dans le treillis initial,
 2. Y non-fermé – ceux-ci sont les nouvelles intensions ;
 3. Dans les deux cas, il y a un concept maximal dans le treillis initial qui engendre une intersection Y particulière :
 1. pour une intension existante, logiquement, c'est le concept sous-jacent, c'est-à-dire $(Y, g(Y))$, qui est le maximal (ce type de concepts sont dits modifiés),
 2. pour une nouvelle intension, le concept maximal est $(f(g(Y)), g(Y))$ où f et g viennent du contexte initial (ces concepts sont les générateurs).

Algorithmes TG – Incrémental

- Autres algorithmes
 - Norris, 1978 ; (seul L est généré)
 - Oosthuisen, 1991 (pseudo TG);
 - .
 - Carpineto et Romano, 1993;
 - Dowling(1993)
 - Vatchev et al, 2000

Algorithmes TG – Etude

- Complexité théorique exponentielle
 - meilleure : Nourine & Raynaud, [IPL 1999]
- Complexité est fonction du Contexte
 - ⇒ complexité sur des cas pratiques ?
 - [Godin, 89], [Kuznetsov & Obiedkov, 2001]
- Plusieurs études comparatives
 - Guénoche, 1978 ; (revue des 4 1ers algo)
 - Godin et al., 1995 (4 algorithmes)
 - Kuznetsov & Obiedkov, JETAI 2002 (une dizaine d'algo.) ;
 - Fu et Mephu Nguifo, EGC'2004 (4 1ers algo);

Algorithmes TG – Etude

- Godin et al, 1995
 - Algo: Chein, Ganter, Bordat, Godin
 - Données aléatoires
 - Petite taille
 - Incrémental vs Non incrémental
 - Avantage à Godin (incrémental), et parfois à Chein

Algorithmes TG – Etude

- Kuznetsov & Obiedkov, JETAI 2002
 - Algo: Bordat, Chein, Ganter, CBO, Lindig, Nourine, Dowling, Godin, Valtchev, Norris
 - Données aléatoires + 1 ens tiré du UCI
 - Taille moyenne

 - Etude théorique et expérimentale
 - Etude des algo implementés en fonction des contextes

 - Avantages à :
 - Godin sur de contexte de petite taille et épars
 - Norris, CBO, Ganter sur des contextes denses
 - Bordat sur des contextes de densité moyenne

 - Bordat pour générer le TG
 - Norris pour génération des concepts

Algorithmes TG – Etude

- Fu et Mephu Nguifo, EGC'2004
 - Limitée aux 4 premiers algo (Chein, Norris, Ganter, Bordat)
 - Données UCI repository

 - Etude efficacité des algo implémentés sur données couramment utilisées en FD
 - Analyse de la transposée du contexte

 - Avantage à Ganter / Norris et Chein
 - Génération TG vs Génération Concepts

Algorithmes TG – FD

- Iceberg Concept Lattices
 - Stumme et al. 2000,
- Incrémentalité et Partitionnement
 - Valtchev et al. 2001, ICCS
- Non Incrémentalité et Partitionnement
 - Fu et Mephu Nguifo, 2003, ICTAI → EGC'04
- Algorithmes fonction du problème
 - Recherche d'association (travaux Lakhal, Zaki, Han)
 - Classification supervisée (Xie et al 2002)

Agenda

- Introduction
- Algorithme de génération des concepts
- **Règles d'Association et TG**
- Classification Supervisée et TG
- Conclusions

Règle d'Association (RA)

- Objectif:
 - Recherche de relations d'association ou de corrélation intéressantes parmi un grand ensemble de données.
- Applications:
 - Analyse du panier d'un client en grande distribution
 - Quel groupe ou ensemble de produits sont fréquemment achetés ensemble par un client lors d'un passage au magasin ?
 - ⇒ Disposition de produits à l'étalage
 - Exemple : Lait et Pain
 - Lorsqu'un client achète du lait, achète-t-il aussi du pain ? Si oui avec quelle fréquence? → 2 Mesures : Support, Confiance

RA - Définition

- Item- Attribut ex: un produit
- Ensemble d'items - Ensemble d'items fréquents
- Transaction : ensemble d'items, ex: un panier
- Soient A et B deux sous ensembles d'items,
 - une **règle d'association** est une implication de la forme $A \Rightarrow B$ avec $A \cap B = \emptyset$.
- Deux mesures :
 - Support : pourcentage de transactions qui contiennent à la fois A et B support $(A \Rightarrow B) = P(A \cup B)$.
 - Confiance : pourcentage de transactions contenant A qui contiennent aussi B confiance $(A \Rightarrow B) = P(B / A)$.

RA - Exemple

- Transactions : ensemble $O = \{1, 2, 3, 4, 5\}$
- Items : ensemble $A = \{a, b, c, d, e\}$

Valeurs booléennes, Dimension simple, Abstraction simple

$$\text{Support}(a \Rightarrow b) = 2/5$$

$$\text{Confiance}(a \Rightarrow b) = 2/3$$

$$\text{Support}(b \Rightarrow c) = 3/5$$

$$\text{Confiance}(b \Rightarrow c) = 3/4$$

	a	b	c	d	e
1	1		1	1	
2		1	1		1
3	1	1	1		1
4		1			1
5	1	1	1		1

RA - Approche

- Démarche:
 1. Rechercher tous les ensembles d'items **fréquents**, c-à-d dont le support est supérieur à un seuil minimum
 2. Générer les règles d'association **fortes** à partir des ensembles d'items fréquents, c-à-d dont le seuil minimum du support et le seuil minimum de confiance sont satisfaits
- Etape 2 est la plus facile
- Performance du processus de génération des règles d'association repose sur la 1ère étape.
- Algorithme : Apriori [[Agrawal](#), [Mannila](#), [Srikant](#), [Toivonen](#) et [Verkamo](#), 1994, 1994, 1996]

RA - Typologie

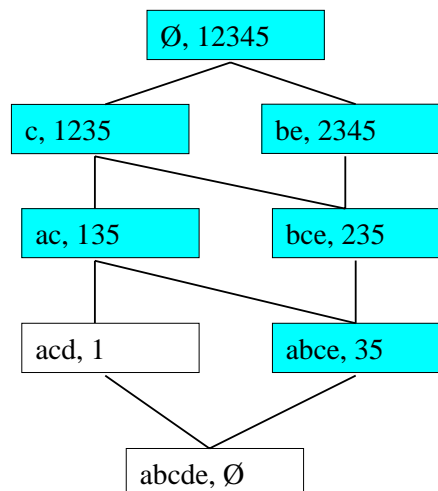
- Plusieurs types basés sur:
 - Types de valeur
 - Booléennes, Quantitatives
 - Dimensions des données
 - Simple, Multiple ex: tenir compte de +sieurs propriétés
 - Niveaux d'abstraction
 - Simple, Multiple ex: prise en compte d'une hiérarchie
 - Autres extensions:
 - Ensembles d'items maximum (ou "Maxpatterns")
 - Ensembles **fermés** d'items (ou "frequent closed itemsets") - **TG**
 - Contraintes sur les règles d'associations
 - Méta-règles pour guider la génération de règles d'association

20/01/2004 - EMN

Tutoriel EGC'04 - Clermont-Ferrand

RA - Exemple

- Frequent Closed Itemset:
- Seuil support = 2
- $|L| = 6$
- Support ($a \rightarrow c$) = 3/5
- Support ($ac \rightarrow b$) = 2/5
Extension(fermé(abc))
- Confiance ($a \rightarrow c$) = 1
- Confiance ($ac \rightarrow b$) = 2/3
Support(abc)/Support(ac)



20/01/2004 - EMN

Tutoriel EGC'04 - Clermont-Ferrand

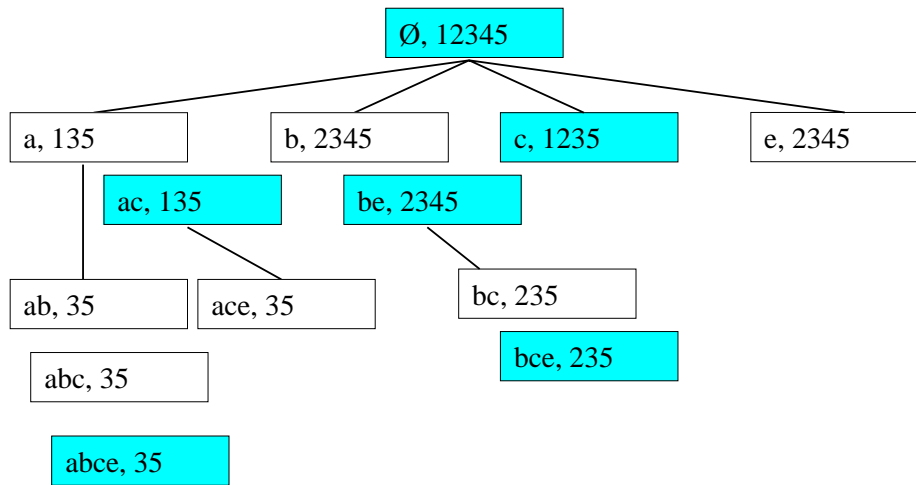
RA / TG - Algorithmes

- CLOSE (Pasquier et al. 98)
- ACLOSE (Pasquier et al. 99)
- CLOSET (Han et al. 00)
- PASCAL (Bastide et al. 00)
- FARD (Ben Yahia et al. 01)
- CHARM (Zaki et al. 02)
- TITANIC (Stumme et al. 00 et 02)
- GALICIA (Valtchev et al. 02)
- CLOSET+ (Han et al. 03)
- ...

RA / TG – Algorithmes - CHARM

- CHARM (Zaki et al. 02), SIAM conf ICDM
- Idées :
 - Explorer simultanément l'espace de recherche des itemsets fermés et des tidsets fermés (nouvelle structure: IT-Tree)
 - Utiliser une méthode de recherche hybride et efficace permettant de ne pas explorer des noeuds inutiles
 - Possibilité une représentation de données verticale appelée diffsets pour améliorer l'efficacité des calculs

RA – Algorithmes - CHARM



20/01/2004 - EMN

Tutoriel EGC'04 - Clermont-Ferrand

RA / TG – Algorithmes - Etude

- Comparaison expérimentale
 - CLOSE vs APRIORI
 - A CLOSE vs CLOSE et APRIORI
 - CLOSET vs A CLOSE et CHARM
 - FARD vs APRIORI
 - CHARM vs CLOSE, PASCAL, CLOSET
- Etude synthétique
 - CLOSE, A CLOSE, CLOSET, FARD, CHARM, TITANIC (Ben Yahia et Mephu Nguifo, 2004)

20/01/2004 - EMN

Tutoriel EGC'04 - Clermont-Ferrand

Agenda

- Introduction
- Algorithme de génération des concepts
- Règles d'association et TG
- **Classification Supervisée et TG**
- Conclusions

Classification Supervisée (CS)

- "Classification" en anglais
- Définition : Classification supervisée
 - Processus à deux phases:
 - Apprentissage : construire un modèle (ou classifieur) qui décrit un ensemble prédéterminé de classes de données,
 - Classement : utiliser le classifieur pour affecter une classe à un nouvel objet
- Applications:
 - Attribution de crédit bancaire, Diagnostic Médical, Marketing Sélectif, Reconnaissance de gènes en Biologie, Prédiction de sites archéologiques, Prédiction du Ballon d'Or Européen (Football),

Classification S - Techniques

- Induction d'arbres de décision,
- Induction de règles de décision,
- Réseaux de neurones,
- Réseaux bayésiens,
- Algorithmes génétiques,
- Apprentissage à partir de d'instances, k-PPV,
- **Induction à partir des treillis,**
- ...

CS - Apprentissage

- **Problème d'apprentissage (supervisée):**

Données :

- f : fonction caractéristique de l'ensemble d'apprentissage; **inconnue**
- O : ensemble d'apprentissage de taille fini, $n \in \mathbb{N}$, suite de couples (x_i, y_i) - exemple ou tuple ou objet ou instance ou observation
- (x_i, y_i) $1 \leq i \leq n$, exemple d'apprentissage tel que $y_i = f(x_i)$
- y_i indique la classe des exemples, nombre fini, valeur symbolique
- A : ensemble d'attributs (propriété ou descripteur), $m \in \mathbb{N}$
- $x_i = (x_{i1}, \dots, x_{im})$, tel que x_{ij} = valeur de x_i pour l'attribut j .

But :

- Construire un modèle (classifieur) f^{\wedge} qui **approxime** au mieux la fonction f à partir d'un ensemble d'exemples sélectionnés de manière aléatoire dans O

CS - Classement

- **Problème de classement :**

Données :

- f^{\wedge} : classifieur; **modèle appris**
- x_k : exemple

But :

- Déterminer $y^{\wedge}_k = f^{\wedge}(x_k)$, classe d'un nouvel exemple x_k :

Question :

- Comment apprécier la différence entre f et f^{\wedge} ?
 - Réponse: calcul du taux de précision ou du taux d'erreur

CS - Validation

- **Taux de précision du classifieur :**

- Pourcentage des exemples de l'ensemble test qui sont correctement classés par le modèle
- Taux d'erreur = $1 - \text{Taux de précision}$

- **Ensemble d'exemples dont on connaît les classes, découpé en 2 (technique du "holdout") :**

- Un ensemble utilisé dans la phase d'apprentissage
- Un ensemble de test utilisé dans la phase de classement

- **Plusieurs autres techniques de découpage, issues des statistiques : (voir [Dietterich, 1997], pour comparaison)**

- Validation croisée, Resubstitution, "Leave-one-out"

CS - Validation

- Critères de comparaison de classifieurs :
 - Taux de précision : capacité à prédire correctement
 - Temps de calcul : temps nécessaire pour apprendre et tester f^{\wedge}
 - Robustesse : précision en présence de bruit
 - **Volume de données** : efficacité en présence de données de grande taille
 - Compréhensibilité : Niveau de compréhension et de finesse
- Problèmes
 - Critères 1 et 2 "mesurables"
 - Critère 4 important pour l'ECBD
 - Critères 3 et 5 "laissés à l'appréciation" de l'utilisateur-analyste

CS - Exemple

- *Ballon d'or 2001*
 - O+ = {Platini, Weah}
 - O- = {Desailly}
 - O? = {Anelka}

 - A =
{JouerNordFrance,
JouerEnItalie,
JouerEquipeFrance}

O/A	a	b	c	Classe
1:Platini	1	1	1	oui
2:Weah	1	1		oui
3:Desailly		1	1	non
4:Anelka	1		1	?

Classification Supervisée - TG

- Pourquoi les treillis de Galois ?
 - Complexité exponentielle !
 - Cadre pour la classification supervisée et non supervisée
 - Concept \approx Extension + Intension
 - Exploration d'une alternative aux arbres de décision
 - Structure redondante \rightarrow duplication supprimée
 - Espace de recherche exhaustif et concis
 - Représentation géométrique intuitive – organisation hiérarchique
 - Propriétés de symétrie et d'invariance
 - Règles de la forme Si-Alors
 - Précision des méthodes existantes

CS – TG - Systèmes

- CHARADE [Ganascia, 87,]
- GRAND [Oosthuisen, 88]
- LEGAL [Liquière & Mephu, 90]
- GALOIS [Carpineto & Romano, 93]
- RULEARNER [Sahami, 95]
- GLUE, IGLUE/CIBLE [Njiwoua & Mephu, ...]
- Flexible-LEGAL [Zegaoui & Mephu, 99]
- CLNN & CLNB [Xie, Hsu & al., 02]

CS – TG – CLNN & CLNB

- Z. Xie, W. Hsu, Z. Liu, and M.L. Lee - JETAI'02, vol.14(2/3)
- Idée : Combinaison de méthodes, Usage de règles contextuelles
 - NBTtree (Kohavi, KDD'96) : Decision Tree and NB
 - LBR (Zheng & Webb, ML journal'00) : Lazy learning & NB
- Principe
 - Intégration d'un classifieur de base (NB ou NN) ds chq noeud du treillis
 - Usage de contraintes pour rechercher les motifs (noeuds) intéressants
 - Stratégie de vote pour classer un nouvel objet
- Résultat: Amélioration du classifieur de base

CS – TG – CLNN & CLNB

- Classifieur Bayésien Naïf
 - Simple, Efficacité en temps de calcul
 - Robuste au bruit et attributs non pertinents
 - Hypothèse: Indépendance conditionnelle des attributs
 - Calcul de la probabilité conditionnelle de la classe C_i étant donné un exemple o . $P(C_i|o) = P(C_i) \times P(o|C_i) / P(o)$
 - Prédiction de la classe ayant la plus grande probabilité
- Classifieur k -NN (k - Plus proches voisins)
 - Recherche des k plus proches voisins étant donné une similarité
 - Prédiction de la classe majoritaire parmi les k -voisins

CS – TG – CLNN & CLNB

- **Classifieur Contextuel Composé**
 - Règle contextuelle $r : H \rightarrow CLS$
 - H est un concept formel (O_i, A_i)
 - CLS classifieur de base induit sur l'extension de H
 - Si \wedge intension(H) alors CLS
 - Plusieurs règles contextuelles $r_1 : H_1 \rightarrow CLS_1$
 - Si intension(H_1) \subseteq f(o) alors r_1 est activé par o
 - CLS_1 est utilisé pour prédire la classe de o , notée $r_1(o)$
- **Vote majoritaire pour trouver la classe finale**

CS – TG – CLNN & CLNB

- **Contraintes pour réduire la recherche**
 - Support
 - $||\text{Ext}(H)|| \geq \alpha \times ||O||$
 - Si $H_2 \leq H_1$ alors $||\text{Ext}(H)|| \geq \sigma / (1 - \text{acc}(H_1 \rightarrow CLS_1))$
 - Précision
 - Si $H_2 \leq H_1$ alors $\text{acc}(r_2) > \text{acc}(r_1) + \frac{\delta}{\log(||\text{ext}(H_1)|| / ||\text{ext}(H_2)||)}$
 - Rejet
 - Si $\text{int}(H_2) \subset \text{int}(H_1)$ et $||\text{ext}(H_1)|| > \gamma \times ||\text{ext}(H_2)||$ alors supprimer r_1
- **Valeurs par défaut**
 - $\alpha = 0.05$ $\sigma = 3$ $\delta = 0$ $\gamma = 0.9$

CS – TG – CLNN & CLNB

- Stratégie de vote pour classement
 - Marquer tous les classifieurs contextuels activés par 0
 - Etant donné 2 règles contextuelles activées $r1$ et $r2$,
 - désactiver $r1$, si $\text{int}(H2) \subset \text{int}(H1)$
 - Désactiver $r1$, si $\exists r2$ statistiquement plus précis que $r1$
 - Utilisation du Chi-2
 - Vote majoritaire sur les règles actives
 - En cas d'égalité, prendre le classifieur avec la précision la plus élevée

CS – TG – CLNN & CLNB

- Expérimentations
 - Visual C++, sous Win98
 - 26 ensembles test - UCI ; VC d'ordre 10, paramètres par défaut
 - Attribut-valeur discrète, Discrétisation s'il y a lieu
 - Taux de précision / NBTree, CBA et C4.5Rules-V8
 - Même jeu de données en apprentissage et test
 - CLNN (17) "meilleur" que NN (2)
 - CLNB (15) "meilleur" que NB (7)

 - CLNB (17) vs NBTree (9)
 - CLNB (14) vs CBA (9)
 - CLNB (18) vs C4.5Rules-V8 (8)
 - CLNB a la meilleure moyenne des taux de précision

CS – TG – CLNN & CLNB

- Expérimentations
 - Temps de calcul
 - Les moins bons pour CLNB
 - 12,92s pour Vehicle data (18 att, 4 classes, 846 ex)
 - 10,98s pour Waveform (21 att, 3 classes, 5000 ex)
 - 7,88s pour Sonar (60 att, 2 classes, 229 ex)

CS – TG – Etude

- Comparaison expérimentale
 - Rulelearner vs C4.5
 - CIBLE vs LEGAL, C4.5, IB1
 - CLNN & CLNB vs C4.5,

 - GRAND, LEGAL, GALOIS, Rulelearner [Huaiyu Fu et al., EGC'04]
- Etude synthétique
 - Mephu Nguifo & Njiwoua, 2002, RR CRIL

Agenda

- Introduction
- Algorithme de génération des concepts
- Règles d'association et TG
- Classification Supervisée et TG
- **Conclusion**

Conclusion

- Treillis de concepts pour la FD ?
- Atouts
 - Structuration; Exhaustivité et Concision; Dualité
- Limites
 - Complexité de génération
- Quid ?
 - **Données de taille volumineuse** - du giga au tera octets
 - Ordinateur rapide - réponse instantanée, analyse interactive
 - Analyse multidimensionnelle, puissante et approfondie
 - Langage de haut niveau, "déclaratif" – Facilité d'usage et Contrôlable
 - Automatisée or semi-automatisée —fonctions de fouille de données cachées ou intégrées dans plusieurs systèmes

Conclusion - Applications

- Indexation documentaire :
 - Godin & al., 1986, *Information Sciences*
 - «Lattice Model of Browsable Data Spaces»
 - Carpineto & Romano, 1996, *Machine Learning*
 - «A lattice conceptual clustering system and its application to browsing retrieval
 - Cole, Eklund & Stumme, 2002, preprint WEB
 - « Document retrieval for email search and discovery using formal concept analysis »
 - ...
- BioInformatique :
 - Thèse Mephu, 1993, Univ. de Montpellier II
 - Duquenne & al., 2001, CLKDD proceedings
 - « Structuration of phenotypes/genotypes through Galois lattices and Implications »

Conclusion - Logiciels

- Logiciels
 - GLAD (Duquenne, 1996)
 - TOSCANA et ANACONDA (Wille et al., 1995 >)
 - CERNATO (Sté Navicon GmbH)
 - TkConcept (Lindig, 1996)
 - Concept Explorer ...
 - GALICIA (Valtchev et al., 2003)
- Sites
 - www.lattices.org
 - Fca list (karlsruhe)

Conclusion - Logiciels

● Outils commercialisés :

- <http://www.kdnuggets.com/software/suites.html>
- **Intelligent Miner** (<http://www.ibm.com>), XML & PMML support
- **Enterprise Miner** (SAS Institute, <http://www.sas.fr>),
- **MineSet** (Silicon Graphics Inc., <http://www.sgi.com>), arrêté en sept 2001
- **Clementine** (Integral Solutions Ltd, racheté par SPSS, <http://www.spss.com/SPSSBI/Clementine>),
- **DBMiner** (<http://www.dbminer.com>),
- **Teradata Warehouse Miner** (NCR solutions),
-

Conclusion - Logiciels

● Outils de recherche :

- **WEKA** (<http://www.cs.waikato.ac.nz/ml/weka>), Open source in Java, plusieurs plateformes, tâches: prétraitement, classification, regression, clustering, règles d'association, et visualisation.
- **YALE** (<http://yale.cs.uni-dortmund.de>), XML format
- **Weka-Parallel** (<http://www.mathcs.carleton.edu/weka>),
- **SIPINA** (Laboratoire ERIC, Lyon),
- ? **DBMiner** (<http://db.cs.sfu.ca>, version libre 90 jours),
- **IBM Intelligent Miner** (<http://www.ibm.com>),
-

Conclusion - Logiciels

- Bases de données Tests :
 - **UCI KDD Archive** (<http://kdd.ics.uci.edu>),
 - **UCI Machine Learning Repository** (<http://www.ics.uci.edu/~mlearn/MLRepository.html>),
 - **MLnet Datasets** (<http://www.mlnet.org>), European
 - ...
- **Kdnuggets** (...),

Conclusion - Sites

- <http://www.kdnuggets.com> : leading and most comprehensive web site on DM & KD
- <http://www.crisp-dm.org> : Cross-Industry Standard Process for Data Mining - standardization effort
- <http://www.dmg.org> : Data Mining Group (DMG), independent, vendor led group which develops data mining standards, such as PMML (Predictive Model Markup Language, XML-based language).

Conclusion - Perspectives

- Algorithmes
 - Etude contextuelle des algorithmes
 - Kuznetsov et Obiedkov, JETAI'02
 - Fu et Mephu, ICFCA'03
 - Partitionnement de données
 - Valtchev et al., ICCS'01
 - ...
 - Mémoire / Disque
 - Parallélisme

Conclusion - Perspectives

- Pertinence concepts générés / Nature Pb.
 - Technique d'approximation
 - Concepts flous / concepts flexibles
 - Treillis de Galois alpha
 - Iceberg Concept Lattices
- Langage de description
- Applications

Conclusion - Diffusion

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
 - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
 - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
 - Journal of Data Mining and Knowledge Discovery (1997)
- 1998 ACM SIGKDD, SIGKDD'1999-2003 conferences, and SIGKDD Explorations
- More conferences on data mining
 - PAKDD, PKDD, SIAM-Data Mining, (IEEE) ICDM, DaWaK, SPIE-DM, etc.
- **En France** : EGC janv 01 (Nantes), janv 02 (Montpellier), janv 03 (Lyon)

Conclusion - Diffusion

- 2001 ICCS workshop on Concept Lattices for KDD
 - Concept Lattices-based Theory, Methods and Tools for Knowledge Discovery in Databases, Stanford (CA), July 30, 2001. <http://CEUR-WS.org/Vol-42> (E. Mephu Nguifo, V. Duquenne and M. Liquière)
 - Special issue of (E. Mephu Nguifo, V. Duquenne and M. Liquière, Eds) :
 - JETAI - Journal of Experimental and Theoretical Artificial Intelligence – April-September 2002, vol. 14(2/3);
 - AAI – Applied Artificial Intelligence - March 2003, vol.17
- 2002 ECAI workshop on Formal Concept Analysis for KDD
 - Advances in Formal Concept Analysis for Knowledge Discovery in Databases, Lyon (France) July 22-23, 2002 (M. Liquière, B. Ganter, V. Duquenne, E. Mephu Nguifo, and G. Stumme)

Conclusion - Diffusion

- 2003 1st ICFCA – Formal Concept Analysis
 - International Conference on Formal Concept Analysis : State of art, Darmstadt (Allemagne), February 27-March 1st, 2003. <http://fzbw.de/icfca03> (R. Wille)
 - Special issue of Journal ... or Book (G. Stumme, B. Ganter and R. Wille);
- 2003 Atelier francophone sur TG pour IA
 - Usages des treillis de Galois pour l'IA, Laval (France) 4 Juillet 2003, Plate-Forme de l'AFIA (P. Valtchev, E. Mephu Nguifo et M. Liquière)
- 2004 2nd ICFCA – Formal Concept Analysis
 - Sidney (Australie), February 23-26, 2004.

Questions ?

