Treillis de Galois et Fouille de données

Engelbert MEPHU NGUIFO

CRIL – CNRS FRE 2499

IUT de Lens

Paris, 12 Juin 2003 -

GDR I3 Fouille de Données

Motivations

- **Treillis de Galois est-il adapté pour la Fouille de données ?**

Plan

- Rappel Fouille de données
- Treillis de Galois en fouille de données
  - Travaux sur les ItemSets Fréquents
  - … Classification supervisée
- Conclusion
  - Limites des approches
  - Pistes à Explorer
- Miscellaneous

Rappel Fouille de données

- Steps of KDD:

  From **Databases and Flat Files**,

  - Cleaning & Integration of DB
  - Selection & Transformation of DW
  - Data Mining of selected DW
  - Evaluation & Presentation of Patterns

  To **Knowledge**

Fouille de données : rappel

- Tasks of Data Mining:
  - Classification
  - Clustering
  - Prediction
  - Estimation
  - Affinity Grouping or Association Rules
  - Description and Visualization

Fouille de données : rappel

- KDD, an interdisciplinary approach
  - Databases
  - Statistics / Data Analysis
  - Machine Learning
  - …

- Pourquoi l'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

Fouille de données : rappel

- **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'01]

ECBD ou encore 'Fouille de données'

- **Concept Lattices as conceptual structures, can be used to address a variety of problems in these research areas ?**

Treillis de Galois : rappel

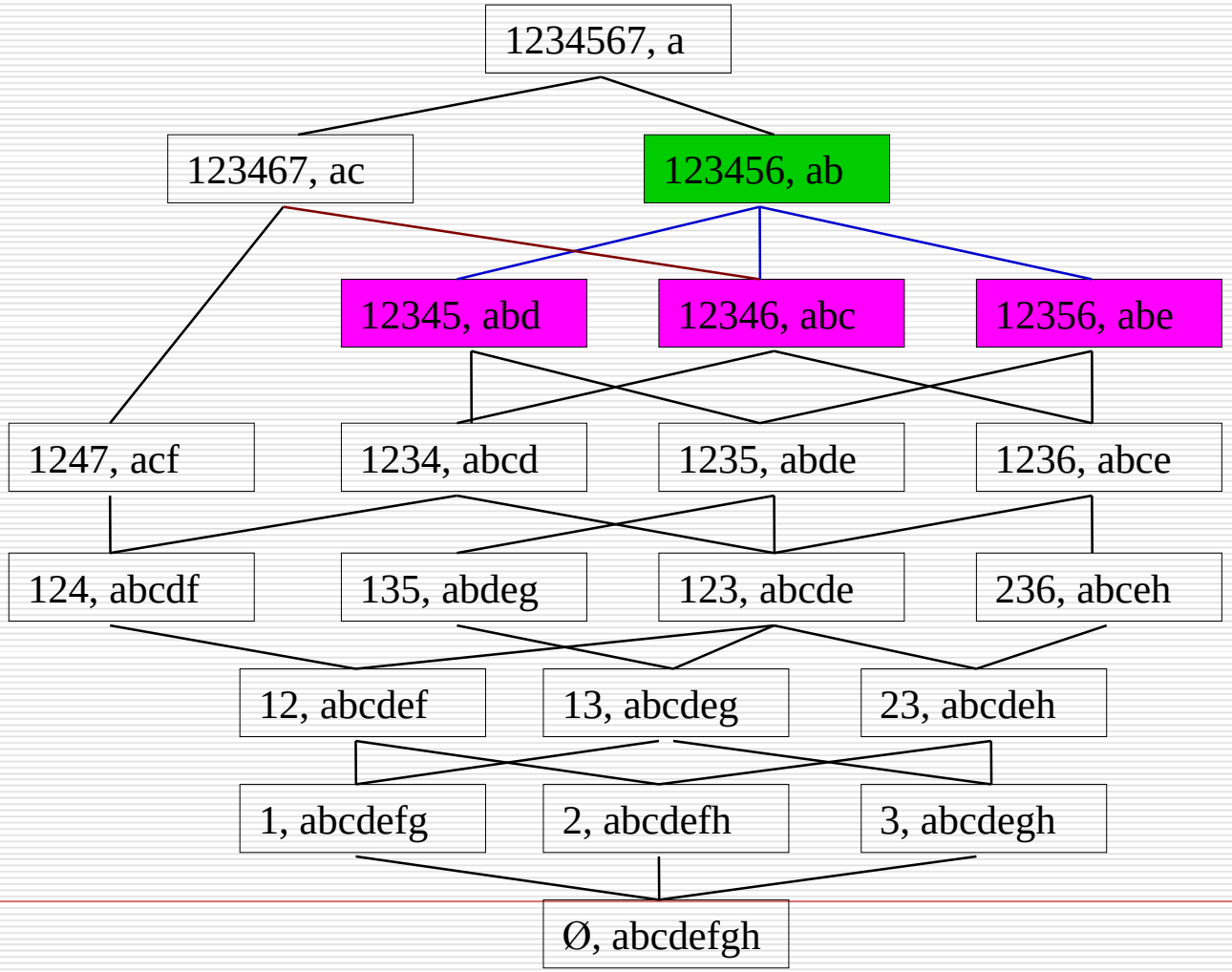
Notions de base:

- **Contexte (O, A, I),**
- **Correspondance de Galois (f, g),**
- **Concept (O_i, A_i), à L**
- **Relation d'ordre (≤),**
- **Treillis de concepts (L, ≤)**

O1 = {6, 7} ⇒ f(O1) = {a, c}

	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		

Treillis de Galois : rappel



Treillis de Galois : Rappel

- $h = g \cdot f$ et $h' = f \cdot g$, sont:
  - isotones : $O_1 \subseteq O_2 \Rightarrow h(O_1) \subseteq h(O_2)$
  - extensives $O_1 \subseteq h(O_1)$
  - idempotentes $h(O_1) = h \cdot h(O_1)$
- h (resp h′) fermeture dans $\mathcal{P}(O)$ (resp $\mathcal{P}(A)$)

- Unicité du treillis pour un contexte
- Treillis = autre image du contexte

- Espace de recherche exhaustif et concis

- Concept ≈ Extension + Intension

Treillis de Galois et FD

- 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)

Treillis de Galois et FD

- 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 Treillis de Galois 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
  - (Australie), February/March ?, 2004.

TG et Règles d'association

- Démarche Génération RA:
  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 le 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]

TG et Règles d'association

- 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 ("frequent closed itemsets") - **TG**
  - Contraintes sur les règles d'associations
  - Méta-règles pour guider la génération de règles d'association

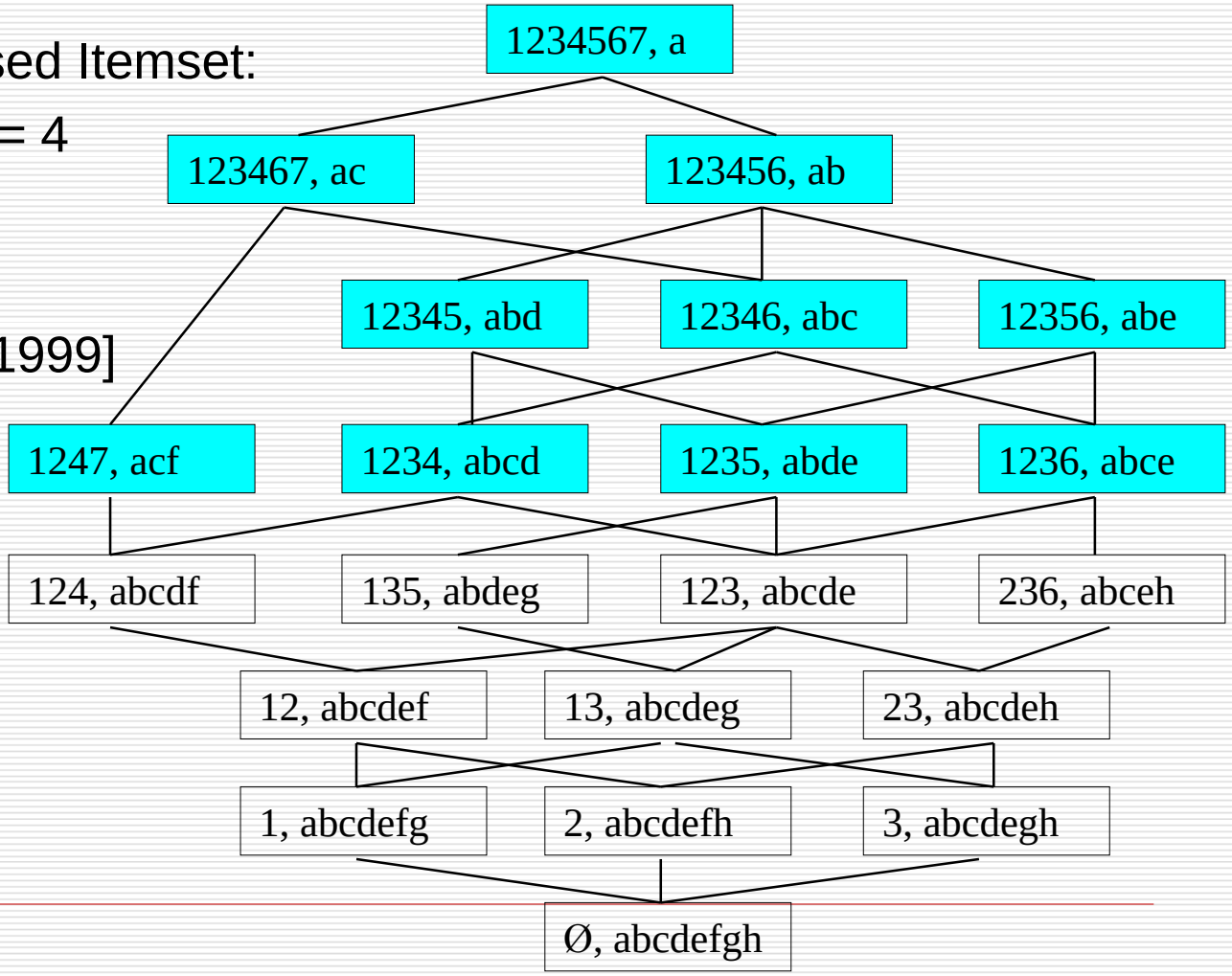
TG et Règles d'association

- Frequent Closed Itemset:
- Seuil support = 4
- **|L| = 10**

- [Lakhal et al, 1999]

Algorithmes:

- Titanic
- Close
- Closet
- Charm
- …



TG et Règles d'association

- Fermés fréquents
  - CLOSET
    - J. Pei, J. Han et R. Mao, ACM DMKD'00
  - Incremental Mining,
    - P. Valtchev, R. Missaoui et al., JETAI'02

TG et Classification supervisée

- “Classification” en anglais
- Définition : Classification supervisée
  - Processus à deux phases:
    1. Apprentissage : construire un modèle (ou classifieur) qui décrit un ensemble prédéterminé de classes de données
    2. Classement : utiliser le classifieur pour affecter une classe à un nouvel objet

TG et Classification supervisée

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

  Données :

  - f : fonction caractéristique de l'ensemble d'apprentissage; **inconnue**
  - O : ensemble d'apprentissage de taille fini, $n \in N$, 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), $m \in N$
  - $x_i = (x_{i1}, \ldots, x_{im})$, tel que x_{ij} = valeur de x_i pour l'attribut j.

  But :

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

TG et Classification supervisée

- **Problème de classement** :

  Données :

  - f^ : classifieur; **modèle appris**
  - x_k : exemple

  But :

  - Déterminer $y^_k = f^(x_k)$, classe d'un nouvel exemple x_k

Question :

Comment apprécier la différence entre f et f^ ?

- Réponse: calcul du taux de précision ou du taux d'erreur

TG et Classification supervisée

- Critères de comparaison de classifieurs :
  1. Taux de précision : capacité à prédire correctement
  2. Temps de calcul : temps nécessaire pour apprendre et tester f^
  3. Robustesse : précision en présence de bruit
  4. **Volume de données** : efficacité en présence de données de grande taille
  5. 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

TG et Classification supervisée

- Taux de précision du classifieur :
  - Pourcentage des exemples de l'ensemble test qui sont correctement classés par le modèle
  - Taux d'erreur = 1 – 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"

TG et Classification supervisée

- 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]

TG et Classification : 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
  - NBTree (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

TG et Classification : 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 Ci étant donné un exemple o. $P(Ci|o) = P(Ci) \times P(o|Ci) / P(o)$
  - Prédiction de la classe ayant la plus grande probabilité
- Classifieur k-PPV (k-Plus proches voisins)
  - Recherche des k plus proches voisins étant donné une similarité
  - Prédiction de la classe majoritaire parmi les k-voisins

TG et Classification : CLNN & CLNB

- Classifieur Contextuel Composé
  - Règle contextuelle r : H **à** CLS
    - H est un concept formel (Oi, Ai)
    - CLS classifieur de base induit sur l'extension de H
  - Si $\wedge$ intension(H) alors CLS
  - Plusieurs règles contextuelles $r_1 : H_1$ **à** 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

TG et Classification : CLNN & CLNB

- Contraintes pour réduire la recherche
  - Support
    - $||Ext(H)|| >= \alpha \times ||O||$
    - Si $H_2 \leq H_1$ alors $||Ext(H)|| \geq \sigma / (1 - acc(H_1$ **à** $CLS_1))$
  - Précision
    - Si $H_2 \leq H_1$ alors $acc(r_2) > acc(r_1) + \delta * \log(||ext(H_1)||/||ext(H_2)||)$
  - Rejet
    - Si $int(H_2) \subset int(H_1)$ et $||ext(H_1)|| > \gamma * ||ext(H_2)||$ alors supprimer r_1

- Valeurs par défaut
  - $\alpha = 0.05$ $\sigma = 3$ $\delta = 0$ $\gamma = 0.9$

TG et Classification : CLNN & CLNB

- Stratégie de vote pour classement
  - Marquer tous les classifieurs contextuels activés par o
  - Etant donné 2 règles contextuelles activées r1 et r2,
    - désactiver r1, si $int(H_2) \subset int(H_1)$
    - 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

TG et Classification : 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

TG et Classification : 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)

TG et autres FD

- Dépendances fonctionnelles et approximatives
  - Lopes et al., JETAI'02

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

- Pistes à explorer
  - 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

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

Miscellaneous

- **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)**

Atelier TG-IA, 4 Juillet 2003, Laval

- Invités:
  - Gerd Stumme : Ontology Engineering with FCA
  - Vincent Duquenne: …
- Exposés:
  - 7 papiers longs, 2 courts et 1 résumé
  - Quelques titres:
    - GALICIA : plateforme ouverte pour l'AFC
    - A fast scalable algorithm to build closed itemsets for large data -- ScalingNextClosure
    - Treillis de Galois Alpha et regroupement conceptuel
    - Algorithmique combinatoire dans les bases de données massives
    - …