DISDAMIN: Algorithmes de Data Mining Distribués

Valerie FIOLET[1,2] - Bernard TOURSEL [1]

[1] Equipe PALOMA - LIFL - USTL - LILLE (FRANCE)

[2] Service Informatique - UMH - MONS (BELGIUM)

- Clermont Ferrand - Janvier 2003

Plan

- **Data Mining**
- **Le projet DISDAMIN**
- **Schéma général pour la recherche de règles d'association**
- **Clustering "incrémental"**
- **Expérimentations**
- **Conclusions et Perspectives**

Data Mining

- **Découverte** de **Connaissances** utiles et “inédites” dans une grande quantité de données.

  - Règles d ’association, Classification (supervisée ou non - clustering), Segmentation, ...

- **Grande quantité de données**

  - grande quantité de travail

  - recours au parallèlisme

    algorithmes hautes performances (parallèles et distribués)

- **Données complexes**

  - travail de structuration des données complexes avant traitement

  - traitement = travail supplémentaire

    => parallélisme

Le projet DISDAMIN:
Distributed Data Mining

- Problématique initiale:
  - base de données médicales (180 attributs)
  - règles d'association.
- Proposer des algorithmes distribués pour le Data Mining.
- Nécessité de prendre en compte:
  - les spécifités du Data Mining
  - les spécificités de traitements parallèles et distribués
    - * paraléllisme de type SMP/CC_NUMA

      (mémoire commune et réseau interne rapide, machines **très coûteuses**)

    * actuellement: - réseau de stations (clusters/grappes de PCs)

      - à plus grande échelle =>**Grid Computing**

      (pas de mémoire partagée, réseau lent)
- But: - traiter de grandes quantités de données pour les problèmes de Data Mining
  - exploiter les ressources disponibles (grille, réseaux de stations)

Règles d'association
Problématique

- But: - trouver des relations de type **A et B => C**
  - fournir une mesure de leur pertinence

- Principe

  Recherche d 'ensemble d 'attributs fréquents
  Nombre de sous-ensembles d'attributs: 2^n

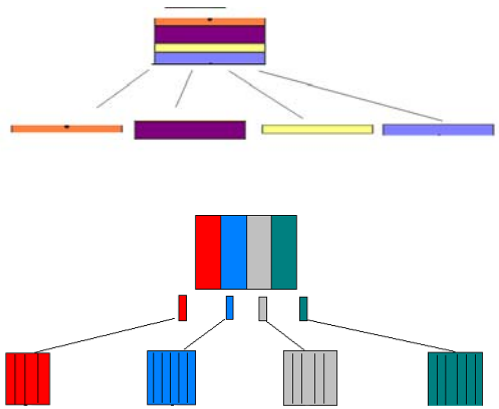
- Algorithme A Priori (Agrawal et al)

Tout sur-ensemble d'un sous-ensemble infréquent est infréquent.

=> **limitation de l'espace de recherche.**

- Algorithmes parallèles existants
  - duplication des sous-ensembles (CD, Parallel Partition, ...)
  - partitionnement des sous-ensembles (DD, IDD, DMA..)
  - => adaptés à des machines paralléles coûteuses

Règles d'association
Proposition Distribuée

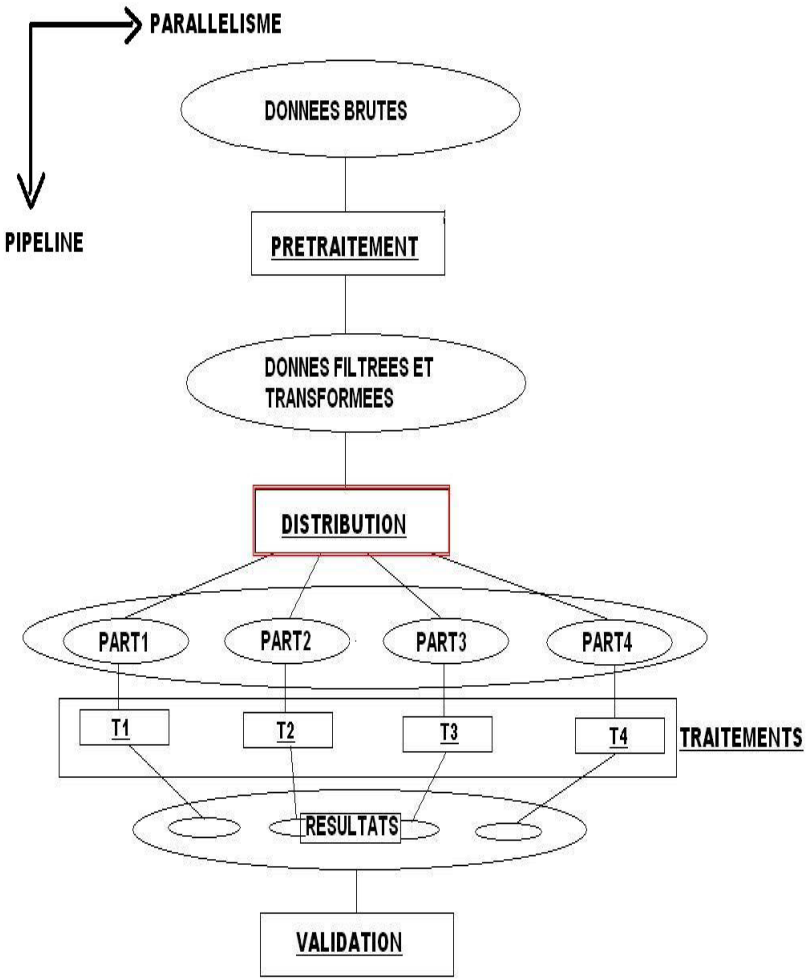


- Possibilités de distribution des données:
  - Horizontale (par enregistrements - base distribuée)
  - Verticale (par attribut - multibase)

- Critère de distribution retenue
  => Distribution horizontale
  identifier des groupes d 'enregistrements à traiter ensemble

- Distribution « intelligente »
  similarité des enregistrements
  - les plus similaires ensemble
  - les moins similaires séparés

Règles d'association
Schéma Global d'Exécution



Plusieurs étapes

- Pré-traitement des données

  filtre, nettoyage, formatage pour le problème.

- Distribution des données

  pb: « intelligemment »
  recherche de profils d 'enregistrements
  (critères de clustering)

- Traitements distribués

  algorithme Apriori
  séquentiel sur chaque fragment

- Validation des résultats

  sur l 'ensemble des données

- Chaque étape peut être distribuée

Règles d'association
Difficultés et Avantages d'une Vision Distribuée

- Méthode basée sur des critères globaux (support, fréquence...)
  - vue locale (partielle) des données sur les sites
    nécessiter de trouver de nouvelles méthodes (heuristiques)
  - limiter les communications, synchronisations
- Exploiter au maximum la puissance de calculs
  - traitement de fragments de données le plus indépendamment possible sur chaque site
  - utiliser le parallèlisme à d 'autres niveaux du schéma
  - approche pipeline entre les étapes
- Sécurité des données
  - format interne pour les communications admises
- Répartition des étapes de pré-traitement et de distribution

  DISTRIBUTION=> **Clustering Incrémental**

Clustering
Problématique

- But
  - identifier des groupes dans les données (des classes)
- Critères
  - dans une classe (un cluster), les données sont les plus similaires possible
  - entre deux classes, les données sont les moins similaires possible
- Principe
  - se baser sur des notions de distance
- Algorithmes
  - Clustering Agglomératif (méthode exacte - problème de complexité)
  - méthode des k-moyennes (complexité acceptable - heuristique)
- Efficacité
  - appel au parallélisme

Clustering “Incrémental”
Proposition Distribuée

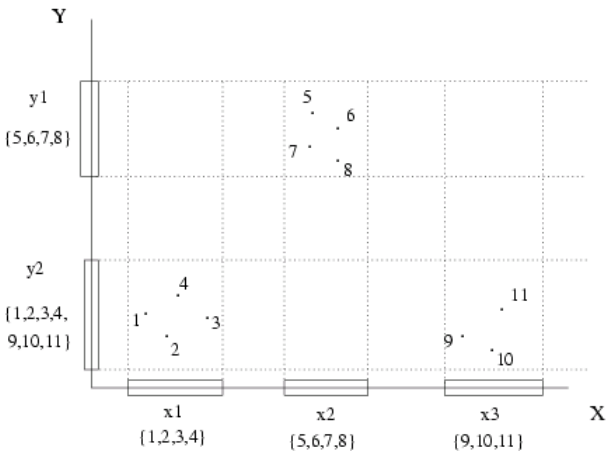
- Origine
  - nécessité d ’un clustering des enregistrements
    (dans le schéma général de recherche de règles)
  - utilisation du parallèlisme
  - résultats de clustering unidimensionnels disponibles

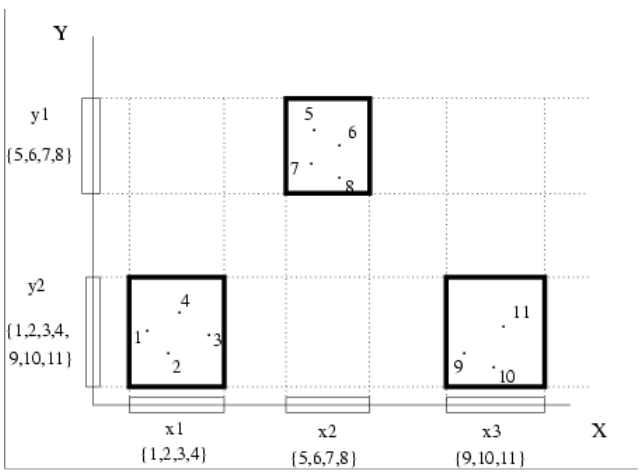
- Lien entre résultats unidimensionnels et multidimensionnels
  - clusterings sur partitions verticales de la base
  - lien avec un clustering sur l ’ensemble de la base

- Plusieurs étapes
  - clusterings unidimensionnels indépendants
    clusterings locaux en parallèle (sur des partitions verticales)
  - regroupement (par croisement) = agglomération
  - (limitation du nombre de clusters)

Clustering “Incrémental”

Principe de fonctionnement





- Croisement des résultats unidimensionnels
  => Clusters candidats
- Elagage des groupes vides
  => clusters multidimensionnels
- Eventuellement:
  Clustering des clusters mutidimensionnels
  => rapprochement
  => limitation du nombre de données

Expérimentations préalables
Recherche de règles d'association

- Données réelles (médicales)
  limitation à 30 attributs
- Premier stade expérimental par clustering centralisé
  (des distances à un modèle)
- Taux de conservation des itemsets fréquents: 95%
- Mise en œuvre des différents principes distribués:
  - pipeline pour recouvrement des communications
  - asynchronisme…
- Vérification de la viabilité d 'une démarche par morceaux
- Paramètres à ajuster mais résultats encourageants.

Expérimentations
Clustering “Incrémental”

- Données synthétiques
- Comparaison des différents algorithmes de clustering
  - clustering agglomératif
  - méthode des k-moyennes
- Pour les différentes phases
  - clustering unidimensionnel
  - clustering de regroupement (limitation)
- Qualité
  - Groupes obtenus
  - Temps - Complexité
  - Potentiel de distribution

Expérimentations
Clustering “Incrémental”

Bilan

Croisement	Kmoyennes		Agglomératif	
Unidimensionnel	Avantages	Inconvénients	Avantages	Inconvénients
Kmoyennes	Temps unidimensionnel			Temps limitation
	Temps limitation	Moins de groupes	-	
	Qualité résultats			Qualité résultats
Agglomératif		Temps unidimensionnel		Temps unidimensionnel
	Temps limitation		-	Temps limitation
		Qualité résultats		Qualité résultats

Comparaison des influences des algorithmes de clustering utilisés

Version	Temps séquentiel	Qualité des groupes	Taux de paraléllisation
Multidimensionnel Agglomératif	- - - -	++	0
Multidimensionnel k-moyennes	+++	++	0
Incrémental (unidimensionnel)	++	++	+++
Incrémental (macroitératif)	+	++	+++

Bilan des avantages des méthodes de clustering incrémental retenues

Conclusions

- Résultats encourageants pour la recherche de règles d ’association
  - même avec une distribution non optimale
  - (critères utilisés, parallèlisations)

- Clustering incrémental = heuristique acceptable de clustering
  - inspiré du schéma général
  - acceptable en dehors de celui-ci

- Clustering de distribution
  - pour le problème posé, il n ’est pas nécessaire d ’avoir un clustering optimale
  - qualité suffisante dans le schéma général de recherche de règles
  - distribuable

- Structuration des données en parallèle
  = phase préalable

Perspectives

- Intégration du clustering «incrémental»
  - comme phase de distribution dans le schéma général (en cours)

- Optimiser l ’étape de distribution

  grâce aux possibilités du traitement distribué
  - pipeline / recouvrement des communications
  - asynchronisme
  - …

- Adapter les paramètres du schéma général

- Générer les règles d ’association à partir des itemsets fréquents
  - de manière distribuée également

- Phase de validation des résultats
  - traitements totalement indépendants à re-vérifier entièrement
  - traitements collaboratifs (échange d ’informations)

  ↗ qualité des résultats

Projets

- Traitements d ’images satellites:

  données spatiales, temporelles, spectrales…

- Reconnaissances de phénomènes flous et dynamiques

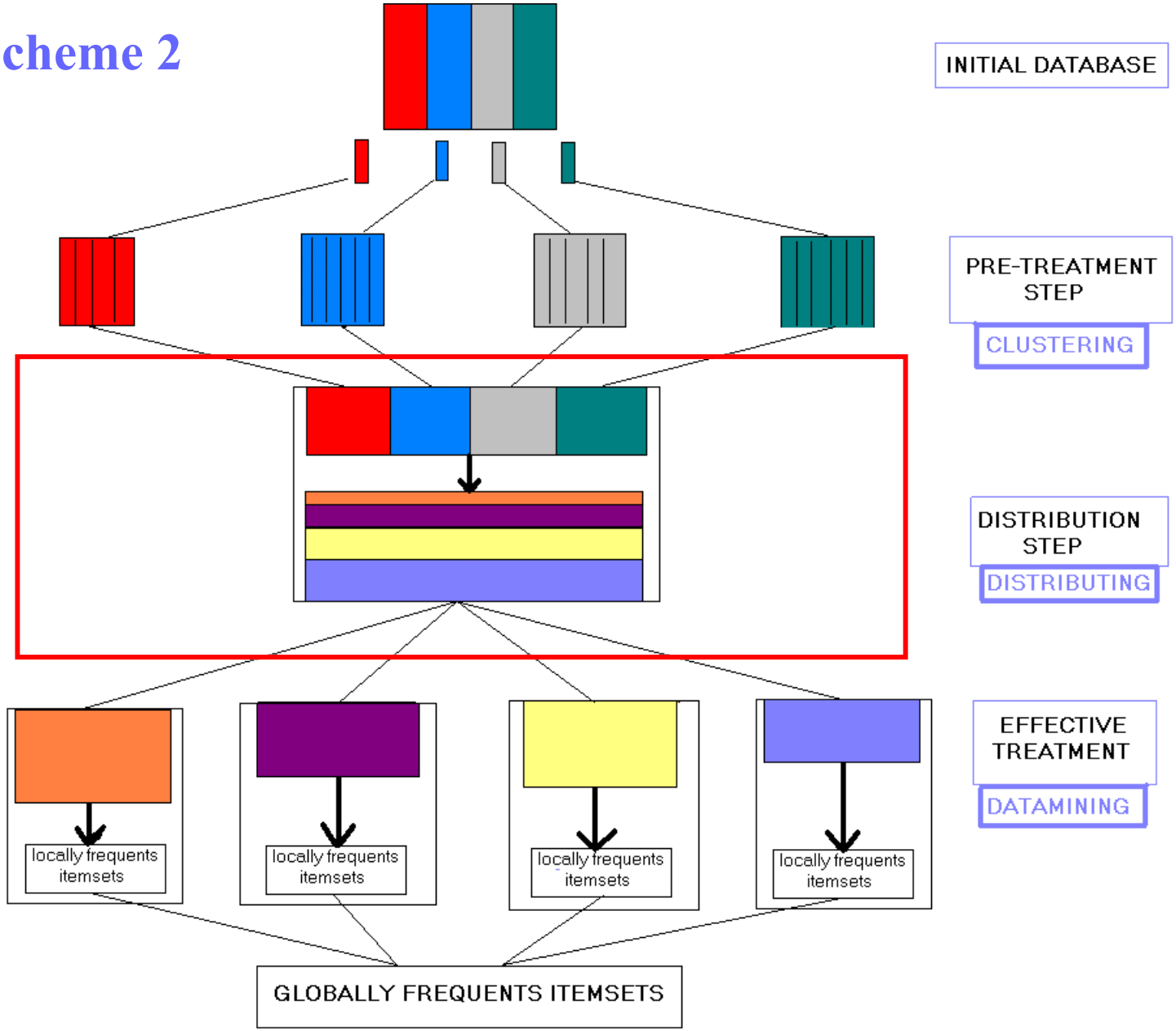
- Associer des compétences:

  - en imagerie
  - en data mining
  - en traitements distribués hautes performances

- Aborder le problème dès le départ avec une vision distribuée

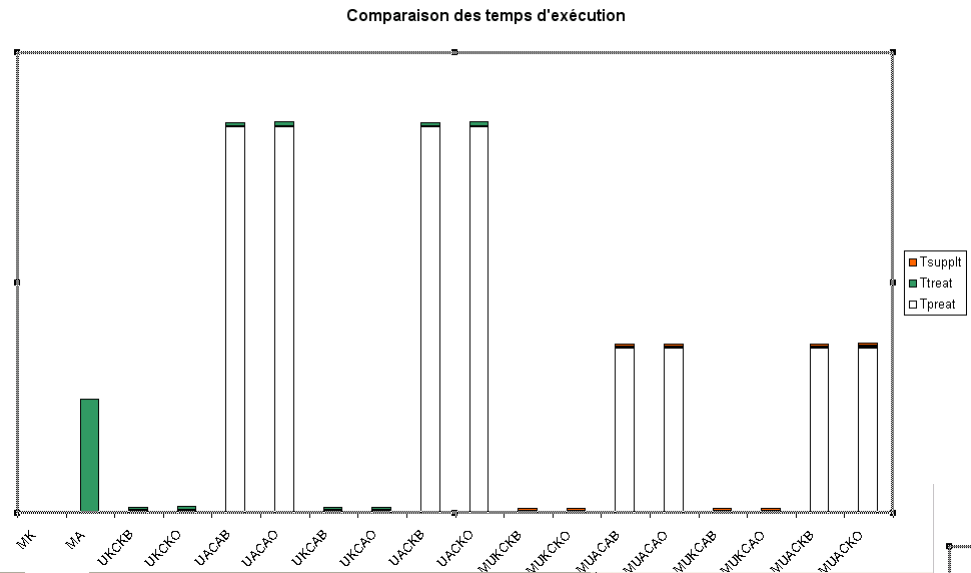
  les meilleurs algorithmes séquentiels donnent rarement de bons algorithmes distribués.

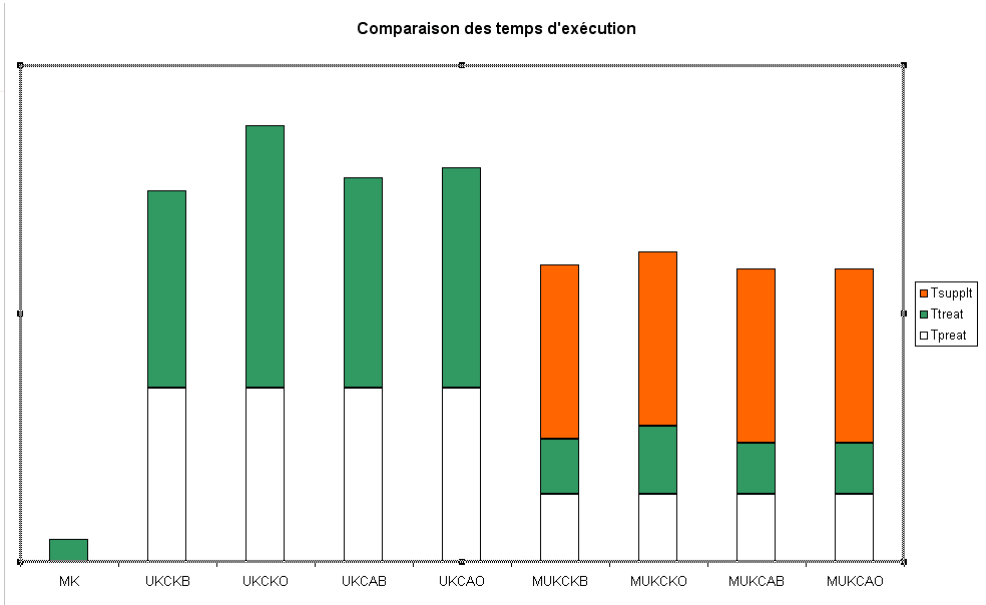
Global Scheme 2



Expérimentations

Clustering “Incrémental”





Clustering “Incrémental”
Vision Distribuée

•Possibilités de distribuer la phase de clustering unidimensionnel
=> résultats indépendants intéressants

•Digressions possibles pour les regroupements
(ordonnés, en structure d ’arbre, selon la disponibilité)

•Macro-Itérations sur les sites distants.