Utilisation de métadonnées pour l'aide à l'interprétation de classes

Abdourahamane Baldé
Université Paris Dauphine – INRIA Rocquencourt
Directeur de thèse : E. Diday
Encadreurs : Y. Lechevallier, B.Trousse
Participation : M-A. Aufaure

Plan

- Introduction
- Problématique
- Objectifs
- Etat de l'art
- Critiques de l'existant
- Notre approche
  - Modèle de métadonnées
  - Architecture
  - Cas d'utilisation
- Conclusions et perspectives

Introduction (1/2)

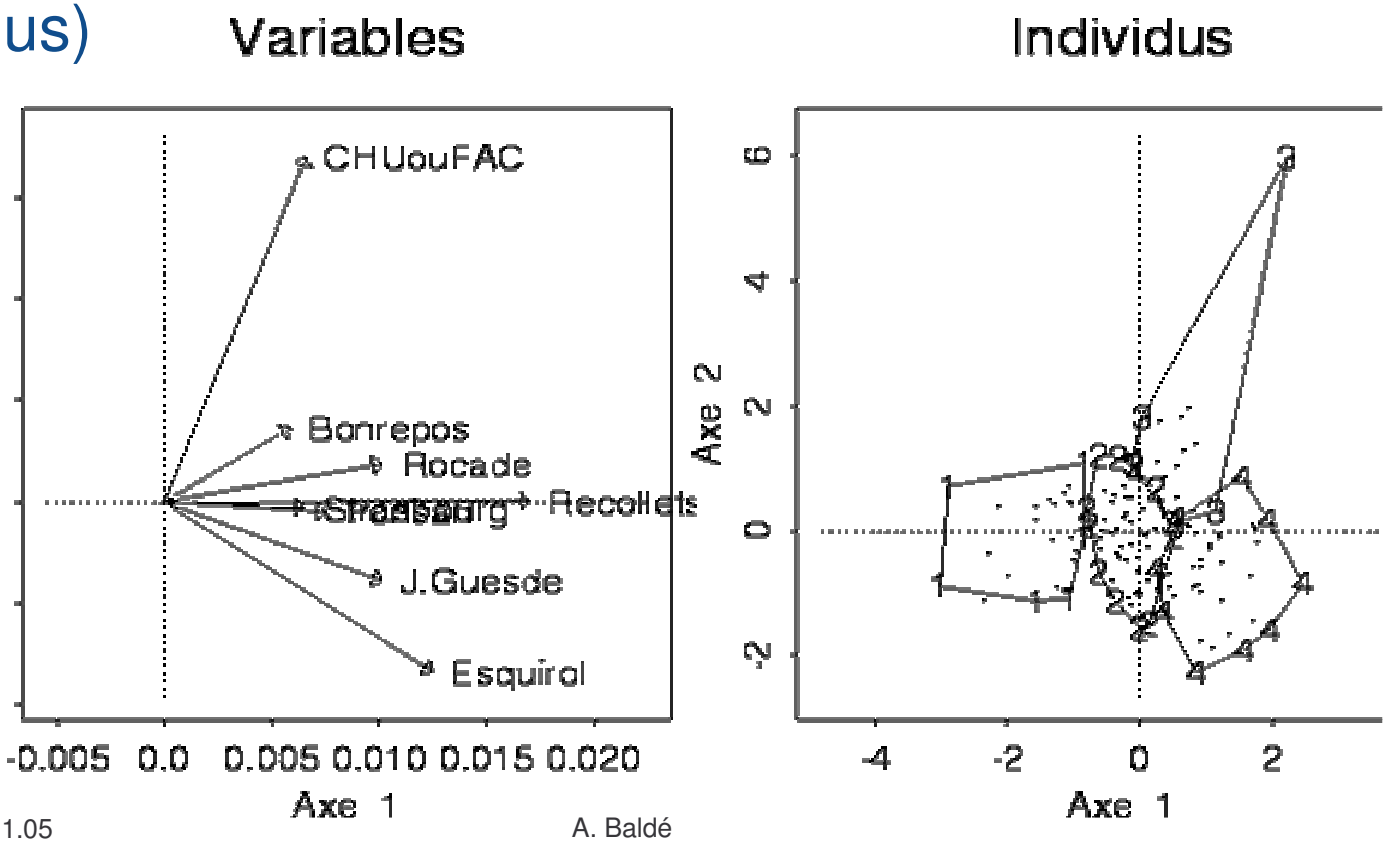
- ***Métadonnées* :** Ensemble d'informations pertinentes sur la *collecte, le traitement, la diffusion, l'accès, la compréhension et l'utilisation* des données ***[K.Zeila, 2004]***
- Elles doivent répondre aux questions:
  - *Qui ? : qui est créateur ou responsable des données,…*
  - *Quoi ? : quelles sont les données traitées ou collectées, …*
  - *Comment ? : la manière dont les données ont été traitées ou collectées ou classifiées,*
  - *Pourquoi ? : la raison pour laquelle telle classe est plus intéressante que telle autre, ….*
  - *Quand ? : la date à laquelle les données ont été collectées ou traitées, …*
  - *Où ? : le lieu de collecte des données, …*

Introduction (2/2)

- ***Deux types*** de métadonnées :
  - fournies par les utilisateurs
  - liées aux résultats et aux données
- ***Classification:*** Découper une population d'objets en plusieurs classes, en tenant compte des variables qui les caractérisent et de la mesure de ressemblance choisie.

Exemple d'une analyse factorielle(1/2)

- Analyser les trajets de bus (selon leur rapidité) (Splus)



Exemple d'une analyse factorielle(2/2)

- Individus à gauche ont des faibles valeurs de variables : *trajets rapides* contrairement aux individus de la classe 4
  - **Besoin d'un expert**
  - **Difficile de connaître le contexte (heure de pointe)**
  - **Pas facile à comparer avec d'autres données d'autres villes portant sur le même sujet**

    Mettre à la disposition de l'utilisateur des informations pertinentes : variables discriminantes, classes les plus homogènes, etc…

Problématique

- **Constat :** chaque outil de classification propose ses propres critères d'interprétation (V_test):
  - Impossible de modifier les règles d'interprétation
  - Besoin d'un expert de l'outil pour l'interprétation
- **Extraire** des informations pertinentes (métadonnées) au cours du processus de classification
- **Utiliser ces métadonnées** pour aider les utilisateurs à interpréter les résultats obtenus
  - Proposer un outil pour faciliter l'interprétation des classes

Objectifs

- Métadonnées exploitables algorithmiquement
  - Créer un modèle structuré servant de cadre pour faciliter l'interprétation de classes
  - Utiliser peu de typage pour notre structure
- Aide à l'interprétation des classes(définir des scénarios) en tenant compte des métadonnées fournies:
  - Par l'utilisateur
    - Les unités de mesure, nombre de classes, etc.
  - Par les données et les résultats
    - paramètres de la méthode, l'usage des variables, etc.
- Création d'une ontologie de la classification facilitant l'interprétation de résultats : **Aspect sémantique**

Etat de l'art

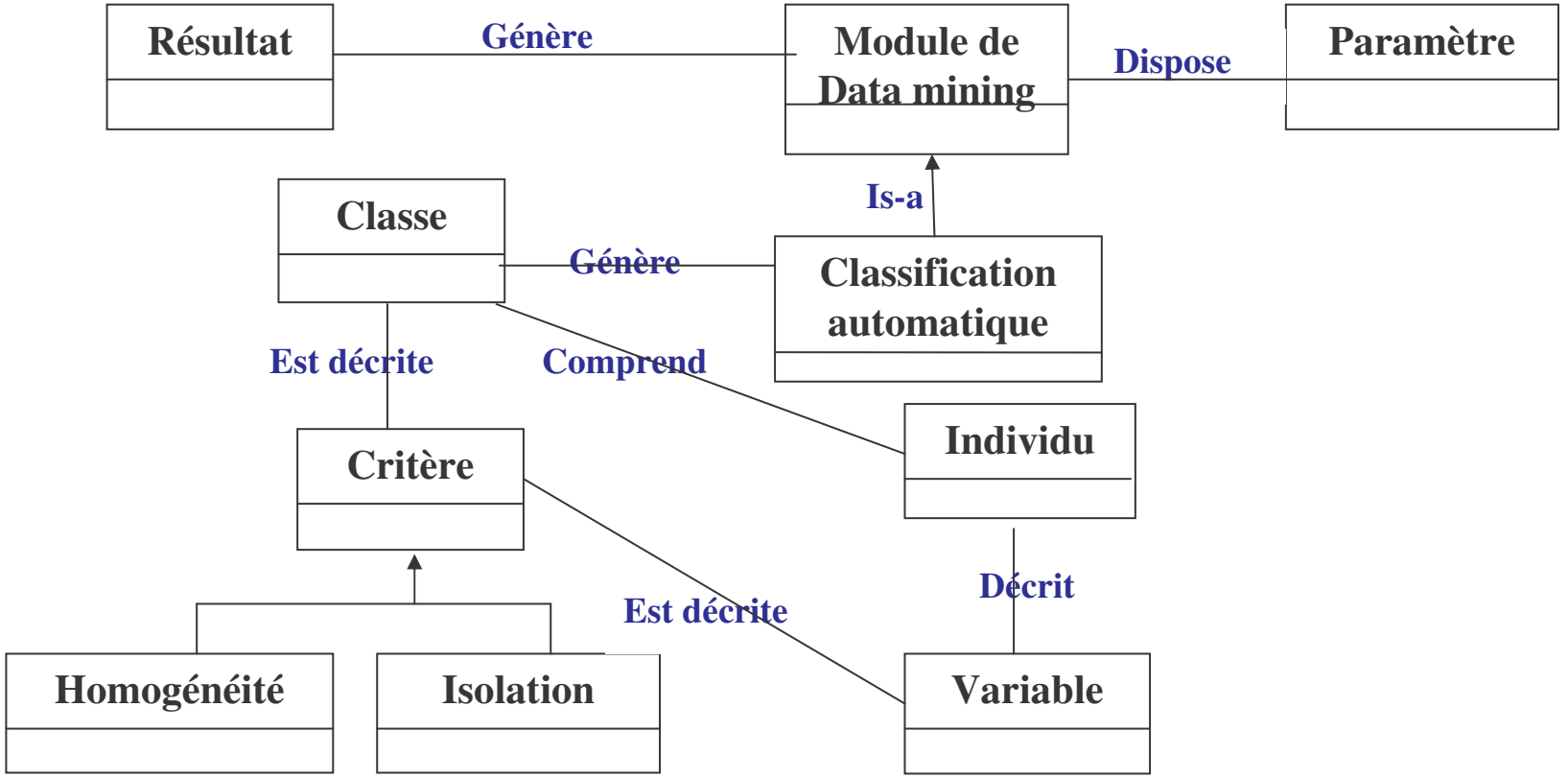
- Métadonnées en base de données (CWM, …)et en statistique (SDMX, SDMS, …)
- Métadonnées en web sémantique et en documentation électronique (DC, DCE, RDF,…)
  Peu d'automatisation (formulaires à remplir) ou extraction à partir de pages web
- Métadonnées en fouille de données (PMML, …)

Tous ces modèles sont généralement orientés sur l'aide à la recherche d'information

Critiques de l'existant

- PMML : décrit uniquement les méthodes de fouille de données, les entrées, le prétraitement sans décrire les résultats des méthodes de classification
- Weka : décrit par un vecteur d'indicateurs globaux les expérimentations des méthodes de classification mais ne stocke pas les informations locales sur les variables ou les classes.

Modèle de métadonnées (1/2)

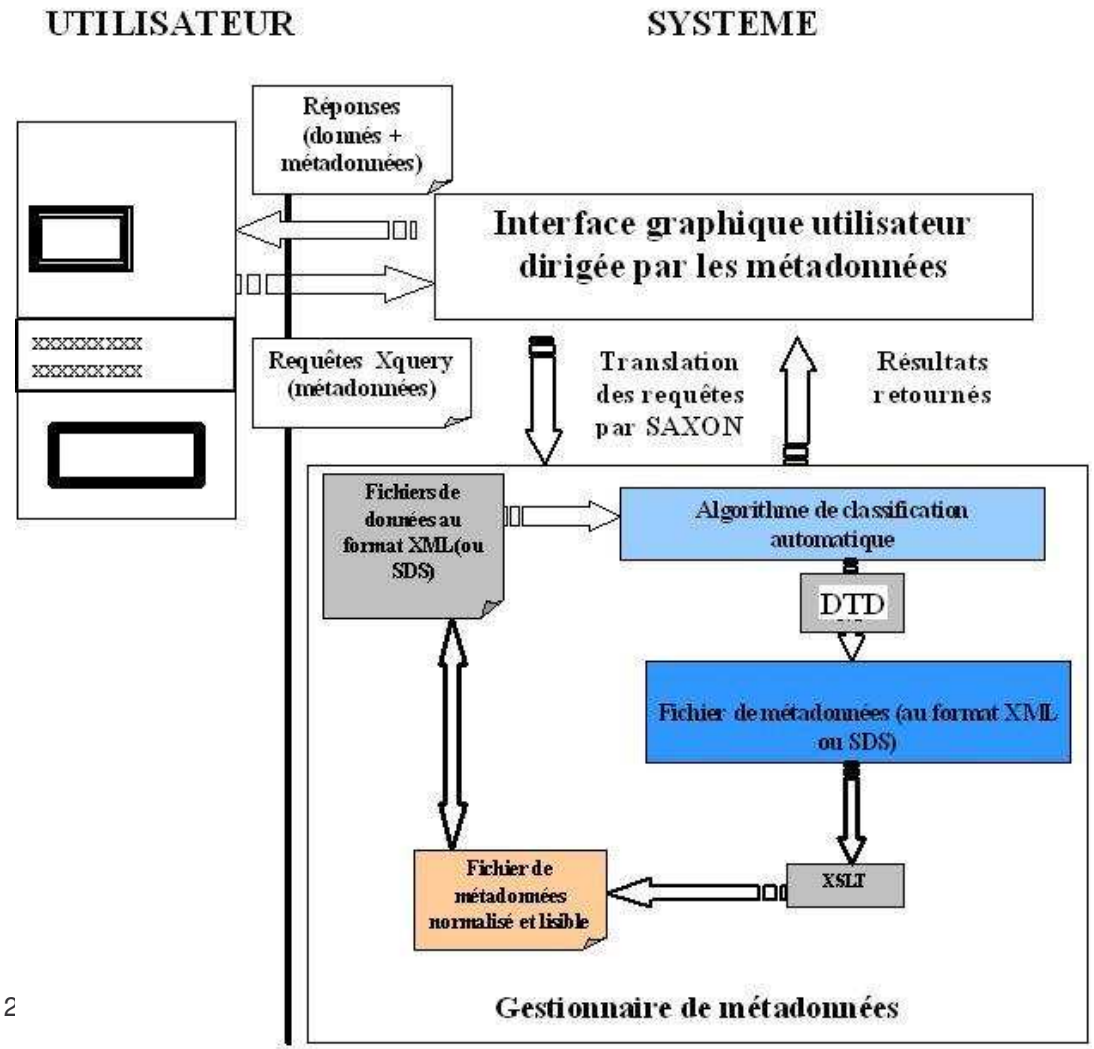


Modèle de métadonnées (2/2)

- Pour **réaliser ce modèle** j'ai opté pour l'utilisation du langage **XML** qui offre certains avantages

Notre modèle est composé de métadonnées fournissant des informations sur les **données** et sur les **classes**

Architecture (1/2)



Architecture (2/2)

Notre modèle permet

- La recherche des données (individus d'une classe)
- L'analyse des données (unités de mesure, …)
- L'interprétation des données et des classes :
  - En donnant la possibilité aux utilisateurs de poser des requêtes XQuery ***(aspect manuel)***
  - En se basant sur l'ontologie du domaine pour faire une interprétation automatique devant être validée par un expert ***(aspect semi-automatique) : non réalisé pour le moment***

Critères d'interprétation

- Soit ***T*** l'inertie totale, ***W*** l'inertie intra-classe et ***B*** l'inertie inter-classe.

T=B+W

- ***Les critères de qualité***
  - **Variable : $Q_J(P)$** mesure le pouvoir discriminant de la variable ***j*** de la partition ***P***.

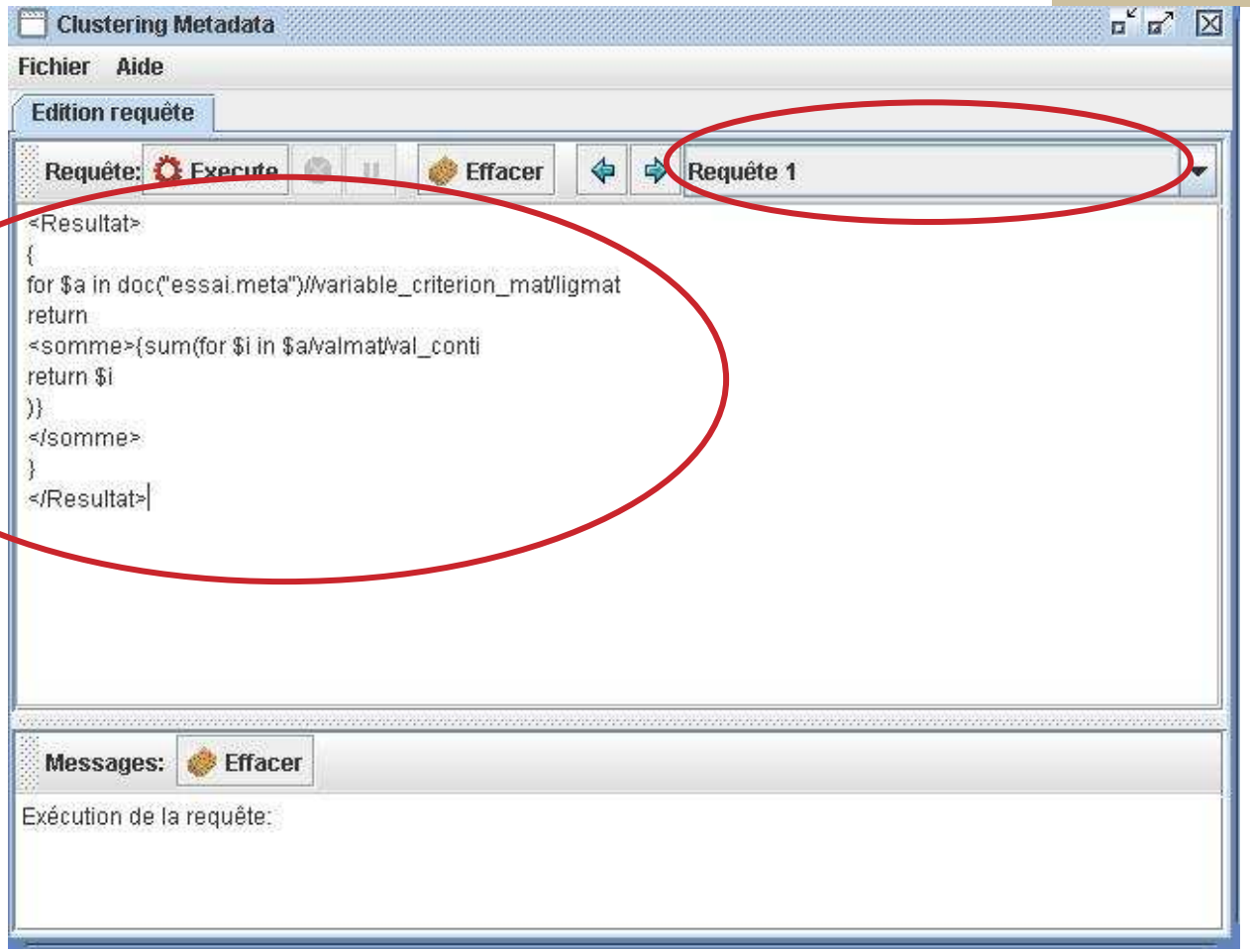
$$Q_j(P) = \frac{T^j - W^j}{T^j} = \frac{B^j}{T^j}$$

  - **Partition : Q(P)** moyenne pondérée des valeurs de **$Q_J(P)$**

Mesure l'importance de la variable dans la partition (à comparer avec **$Q_J(P)$**)

$$Q(P) = \frac{B}{T} = 1 - \frac{W}{T}$$

Outil d'interprétation



Plan de génération des métadonnées

- Fichier de base (.meta) généré au moment de la classification ou à partir de fichiers résultats structurés
- Fichier intermédiaire (.data) généré au moment du calcul des critères locaux
- Fichier final (.final) contenant le résultat de l'interprétation

Etude de cas : Sclust

- Exemple : métadonnées générées au moment de la classification des données issues des fichiers logs: fichiers .meta, .data et .final

Conclusions & perspectives

- Nous avons présenté :
  - La définition d'un modèle de métadonnées
  - L'extraction des métadonnées selon un format cible
  - La définition d'une structure de métadonnées pour la classification
  - La définition de scénarios à travers des requêtes XQuery
- Nous envisageons :
  - utiliser les métadonnées pour la détermination du bon nombre de classes
  - de définir un "langage" de description des métadonnées
  - de définir une ontologie du domaine de la classification
  - d'étendre Weka pour aider à interpréter les résultats