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

Abdourahamane Baldé*, Yves Lechevallier*,
Brigitte Trousse**, Marie-Aude Aufaure***

* INRIA Rocquencourt (Projet AxIS)
Domaine de Voluceau Rocquencourt, B.P. 105, F-78153 Le Chesnay Cedex, France
{abdourahamane.balde, yves.lechevallier}@inria.fr
** INRIA Sophia Antipolis (Projet AxIS)
Route des Lucioles, B.P. 93, F-06902 Sophia Antipolis Cedex, France
brigitte.trousse@inria.fr
***Supélec - Plateau du Moulon - Service Informatique
F-91192 Gif-sur-Yvette Cedex
marie-aude.aufaure@supelec.fr

Résumé. Les résultats des méthodes de fouille de données sont difficilement interprétables par un utilisateur n'ayant pas l'expertise requise. Dans ce papier nous proposons un outil permettant aux utilisateurs d'interpréter les résultats issus des méthodes de classification non supervisée. Cet outil est basé sur des métadonnées utilisées pour formaliser le processus d'interprétation automatique. Ces métadonnées vont servir à l'utilisateur pour comprendre dans quelles circonstances les données originales ont été collectées et de quelle manière elles ont été agrégées puis classifiées. L'intérêt de ce travail porte sur la souplesse qu'auront les utilisateurs à pouvoir interpréter facilement les classes obtenues. Nous développons notre approche basée sur l'utilisation des métadonnées. Nous traduirons notre méthodologie par un exemple concret.

1 Introduction

La fouille de données définie comme étant l'extraction à partir de données brutes de connaissances potentiellement exploitables, n'en demeure pas moins un processus complexe dès lors qu'il s'agit d'interpréter les résultats fournis. Les techniques de fouille de données représentent une étape fondamentale du processus d'Extraction de Connaissances dans les Bases de Données connu sous le nom ECD ou KDD (Knowledge Discovery in Databases) (Han 2001).

Dans ce papier nous nous intéressons à l'une de ces techniques : la classification non supervisée. Celle-ci est définie comme un ensemble de processus aptes à être exécutés sur ordinateur pour constituer des hiérarchies de classes ou de simples partitions établies à partir de tableaux de données (Jambu 1978). Les règles d'interprétation des structures classificatoires obtenues (hiérarchies, partitions, etc.) à l'issue de ces classifications n'ont pas la simplicité des méthodes descriptives uni-dimensionnelles.

Notre objectif, dans ce travail, est de proposer une aide aux utilisateurs afin d'interpréter les résultats des méthodes de classification. En effet, les modules de classification proposent des techniques de visualisation des résultats très intéressantes et conviviales (Song 1998), (Sprenger et al. 2000), (Wills 1998) mais la plupart ont fait l'impasse sur la structuration des résultats.

En partant de ce constat, nous proposons une nouvelle approche qui consiste à utiliser les métadonnées comme moyen de représentation des connaissances capitalisées au cours du processus de classification.

Les métadonnées sont souvent définies comme étant des données sur les données (Grossmann 1999), (Kent et al. 1997). Elles sont aussi définies comme un ensemble d'informations pertinentes pour la collecte, le traitement, la diffusion, l'accès, la compréhension et l'utilisation des données (Zeila, 2004). En ce sens, elles peuvent aider à comprendre dans quelles circonstances les données originales ont été collectées et de quelle manière elles ont été agrégées puis classifiées.

Dans ce travail, nous proposons une architecture basée sur notre modèle de métadonnées (Baldé et Aufaure, 2005) pour réaliser le processus de production de métadonnées et d'aide à l'interprétation de classes. Nous présentons notre outil de traitement de métadonnées, basé sur le langage Xquery. Pour valider notre travail, nous montrons les résultats de notre expérimentation réalisée sur les données issues des fichiers logs de l'INRIA[1].

2 Notre approche

L'aide à l'interprétation consiste en toute technique ou calcul qui permet d'éprouver le bien fondé des classes obtenues en rendant raison de la formation de celles-ci (Jambu 1978). Dans la section suivante nous allons présenter notre architecture basée sur le modèle de métadonnées.

2.1 Architecture

L'architecture que nous présentons exploite les métadonnées produites au cours du processus de classification automatique (figure 1).

Cette architecture est constituée d'un ensemble de couches définies ci-après :

- un modèle de métadonnées;
- un gestionnaire de métadonnées : qui va servir de tampon entre notre modèle de métadonnées et les manipulations qui y seront effectuées;
- une couche gérant des requêtes d'utilisateurs en interrogeant le gestionnaire de métadonnées. C'est cette dernière couche qui traite les requêtes des utilisateurs exprimées en Xquery. Cette couche traite les requêtes exprimées en Xquery (Chamberlain 2004).

Pour implémenter ces requêtes nous avons utilisé le processeur Saxon[2]. Contrairement à d'autres processeurs comme Berkeley DB XML[3], Saxon est un ensemble d'outils dédiés aux traitements des documents XML et est performant en terme de rapidité et conforme aux spécifications du W3C.

[1] Institut National de Recherche en Informatique et Automatique
[2] http://www.saxonica.com
[3] http://www.sleepycat.com/products/xml.shtml

Notre approche présente l'avantage qu'il ne soit plus nécessaire de procéder à des modifications de certains critères (relatifs à l'homogénéité par exemple) et de relancer le module de classification pour observer le gain ou la perte d'homogénéité. Le but est, à partir des résultats des requêtes, de pouvoir comparer plusieurs scénarios d'interprétation.

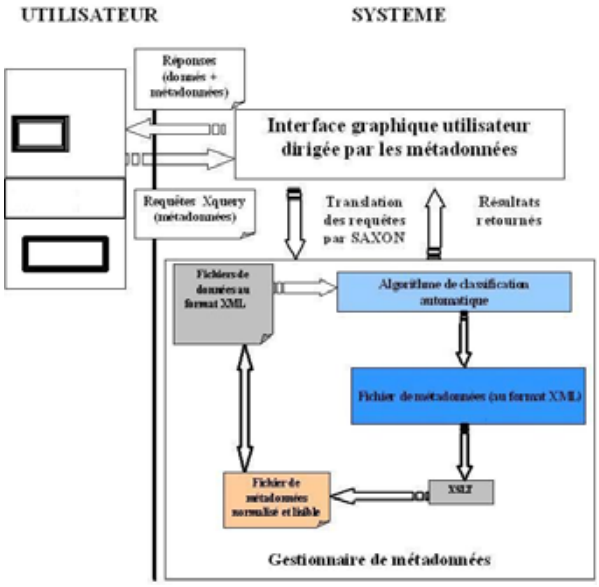

FIG. 1 – *Architecture de production et de traitement des métadonnées*

Notre objectif étant d'aider les utilisateurs dans l'interprétation de leurs résultats, nous avons mis à leur disposition un certain nombre de scénarios d'interprétation (exprimés en Xquery) définis par les experts du domaine. Ceux-ci ont permis d'interpréter des résultats de modules de classification tels que Sclust[4] (Chavent et al. 2003). Cependant l'utilisateur a la possibilité de modifier les critères utilisés pour réaliser son propre scénario. A travers cet outil nous lui donnons la possibilité d'interpréter les résultats dont il dispose et surtout de manipuler automatiquement ceux-ci.

2.2 Processus d'extraction

Les éléments de métadonnées extraits sont de deux types : les métadonnées correspondant aux informations fournies par l'utilisateur et celles qui sont associées aux données et aux résultats de la classification.
Voici des exemples de métadonnées fournies par l'utilisateur : le nombre de classes, les informations sur l'auteur, la description de la méthode de classification, la distance utilisée, l'unité de mesure utilisée pour certaines valeurs de données, etc…

[4] Sclust est un module du logiciel SODAS développé dans le cadre du projet Européen ASSO: http://www.ceremade.dauphine.fr/touati/sodas-pagegarde.htm

Les métadonnées liées aux données et aux résultats sont par exemple : les paramètres de la méthode de classification, le nombre d'individus dans une classe, la source des données originales, la description des variables, l'usage des variables (active, prédictive,…), les valeurs des critères d'hétérogénéité et/ou d'isolation pour chaque classe, la contribution de chaque variable dans la construction de chaque classe, etc…
Ces métadonnées sont extraites au cours de l'exécution de l'algorithme de classification. En sortie de l'algorithme, nous obtenons le fichier de métadonnées. Ce fichier va nous servir de base à la réalisation des requêtes d'interprétation. L'utilisateur pourra effectuer les calculs qu'il souhaite afin de mieux affiner son interprétation. Cette flexibilité permettra d'interpréter au mieux les résultats suivant le domaine visé par la classification. Par exemple un utilisateur privilégiera le critère f-mesure ou le critère d'inertie intraclasse ou encore la contribution marginale des variables, etc…

Pour rendre ces fichiers de métadonnées plus lisibles et mieux compréhensibles par les utilisateurs nous avons utilisé le processeur XSLT (Kay 2001). Pour le traitement des requêtes posées par les utilisateurs, nous avons utilisé le processeur Saxon.

La description des critères utilisés dans ce travail est dans (Jambu 1978). Il faut savoir que la relation fondamentale dans une interprétation, basée sur l'inertie, est : $T = B + W$ où T est l'inertie totale indépendante de la partition, B l'inertie de nuage des centres de gravité munis de poids (l'inertie interclasse) et W l'inertie d'une classe k par rapport à son propre centre de gravité (l'inertie intra classe).

3 Expérimentation

Pour cette étude de cas, le module de classification utilisé est Sclust. Ce module est utilisé pour partitionner un ensemble d'individus décrits par des données symboliques (Bock et Diday 2000) en un nombre k de classes homogènes. Pour interpréter ces classes obtenues, l'utilisateur dispose uniquement d'un fichier listing inexploitable algorithmiquement et mal structuré.

Les données qui sont traitées proviennent des navigations recensées sur les deux serveurs de l'INRIA (siège et Sophia) pour la période du 1er au 15 janvier 2003. Ces navigations ont été prétraitées suivant un certain nombre de critères (El Golli et al.2005) avant de procéder à leur classification. Les informations sur l'outil de prétraitement sont décrites dans (Tanassa et Trousse 2004) ou sur le site web www-sop.inria.fr/axis/axislogminer. Le but de la classification est de voir si les projets de recherche qui ont des activités scientifiques ou tout au moins des centres d'intérêt communs se retrouvent en analysant uniquement le parcours des internautes sur ces deux sites. Le tableau de données est composé de 100 groupes de navigations décrits par 127 variables.

Supposons que l'utilisateur soit intéressé par la contribution de chacune des variables dans la formation des classes, alors en utilisant notre outil il obtiendra l'ensemble des variables ayant une contribution supérieure au seuil qu'il a fixé. Par exemple, l'utilisateur cherche des variables ayant une contribution supérieure à 1.5 fois la moyenne. Il obtient :

```
<Resultat>
  <name>www-sop/robotvis</name>
  <name>www/actualites</name>
  <name>www/index.fr.html</name>
  <name>www/inria</name>
  <name>www/multimedia</name>
```

```
<name>www/personnel</name>
<name>www/travailler</name>
<name>www/</name>
</Resultat>
```

L'utilisateur peut faire évoluer ce seuil, par exemple il peut vouloir chercher les variables qui ont un taux de contribution supérieur à 1.6 fois la moyenne. Alors nous aurons les variables:

```
<Resultat>
<name>www/travailler</name>
<name>www/actualites</name>
<name>www/multimedia</name>
</Resultat>
```

4 Conclusion et travaux futurs

Dans ce papier nous avons présenté une architecture exploitant les métadonnées en vue d'une aide à l'interprétation des résultats de classification automatique. Nous avons présenté l'utilisation de notre modèle dans le cas d'une méthode de classification non supervisée (Sclust) et montré que celui-ci peut s'adapter à d'autres modules de classification. De plus, nous envisageons d'inclure le traitement des fichiers PMML[5] par notre outil. En effet, PMML étant devenu un standard de représentation très populaire, utilisé dans l'API JDM[6], il nous semble utile de l'intégrer dans notre processus de traitement.

Pour conclure, nous pouvons dire que les métadonnées peuvent assister les utilisateurs dans la recherche de l'information, dans l'interprétation de leur contenu et elles peuvent aider dans les post-traitements des classes construites.

Nos travaux futurs s'articuleront autour de la création d'une ontologie du domaine de la classification. Pour ce faire nous nous appuierons sur les travaux réalisés au niveau des méthodes de fouille de données par (Cannataro 2003). Cette ontologie permettrait une interprétation automatique des classes et des partitions obtenues par des modules de classification.

Une autre perspective à ce travail serait d'utiliser des techniques de visualisation pour améliorer le processus d'interprétation des résultats. Les métadonnées pourront aussi servir à aider dans la détermination du bon nombre de classes.

Références

Baldé, A. et M-A. Aufaure (2005). How can metadata contribute to add semantic information to clusters? Journal of Symbolic Data Analysis, 3(1), pp 32-44.

Bock, H. et E. Diday (2000). Analysis of Symbolic Data.Exploratory methods for extracting statistical information from complex data, Springer Verlag.

Cannataro, M. et C. Comito (2003). A Data Mining Ontology for Grid Programming, In Workshop on Semantics in Peer-to-Peer and Grid Computing.

[5]Predictive Model Markup Language du Data Mining Group : http://www.dmg.org

[6] Java Data Mining est une API dédiée aux développeurs d'applications orientées en fouille de données. http://jcp.org/aboutJava/communityprocess/mrel/jsr073/

Chamberlin, D. et al. (2004). Xquery from the Experts: A guide to the W3C XML Query Language, Addison-Wesley.

Chavent, M., F.A.T De Carvalho, Y. Lechevallier et R. Verde (2003). Trois nouvelles méthodes de classification automatique des données symboliques de type intervalle. Revue de Statistique Appliqué. Paris (France) : v. LI, n. 4, pp 5-29.

El Golli, A., F. Rossi, B. Conan-Guez, et Y. Lechevallier (2005). Une adaptation des cartes auto-organisatrices pour des données décrites par un tableau de dissimilarités, RSA.

Grossmann, W. (1999). Metadata, In Enciclopedia of statistical, John Wiley and sons, pp. 811-815.

Han, J. et M. Kamber (2001). Data Mining: Concepts and Techniques, San Francisco, California, Editions Morgan Kaufmann, pp. 550.

Jambu, M. (1978). Classification automatique pour l'analyse des données, Dunod, pp. 245-286.

Kay, M. (2001). XSLT Programmer's Reference, Wrox, 2nd edition.

Kent, J.P. et M. Shuerhoff. (1997). Some thoughts about a metadata management system, In Proceedings of the Ninth International Conference on Scientific ans Statistical Databases Systems, pp. 174-185.

Song. M. (1998). BiblioMapper: A Cluster-Based Information Visualization Technique, infovis, p. 130, IEEE Symposium on Information Visualization.

Sprenger, T.C., R. Brunella et M.H. Gross. (2000). H-BLOB: A Hierarchical Visual Clustering Method Using Implicit Surfaces, vis, 11th IEEE Visualization 2000 (VIS'00).

Tanassa, D., et B. Trousse (2004). Advanced Data Preprocessing for Intersites Web Usage mining, IEEE Intelligent Systems, 19(2), pp.59-65.

Wills, G.J. (1998). An Interactive View for Hierarchical Clustering, infovis, p. 26, IEEE Symposium on Information Visualization.

Zeila, K. (2004). Metadata driven integrated statistical data processing and dissemination system. In Proc. International Conference Statistics :investment in the future, Prague.

Summary

A huge volume of data is produced by many applications. Data mining techniques are used to extract knowledge from this mass of information since it is no longer possible to manually examine this data. In the meantime, the interpretation of the results obtained by applying data mining techniques is not easy. In this paper, we focus on unsupervised learning, and we propose a tool in order to help the end-user to interpret the clusters obtained. Our objective is to facilitate the interpretation process and to point out that metadata can play a major role for this purpose. Metadata will help the user to understand how the original data has been collected, aggregated and then classified. One of the characteristics of this work is that users have the possibility of carrying out the calculations that they wish. These calculations were done by using Xquery queries. To validate our work, we present an example.