Fouille de données orientée motifs, méthodes et usages.

François RIOULT
GREYC - Équipe Données-Documents-Langues
CNRS UMR 6072
Université de Caen Basse-Normandie
France

Résumé

La fouille de données orientée motifs est une discipline récente à l'intersection des domaines des bases de données, de l'intelligence artificielle et de la statistique. Les techniques mises au point permettent l'extraction d'information dans de très volumineuses bases de données, sous la forme de motifs fréquents et de règles d'association. Ces connaissances sont exploitées à des fins de classification supervisée, non supervisée ou de caractérisation de classe.

1 Introduction

L'Extraction de Connaissances dans les Bases de Données (E.C.B.D.) est une discipline récente, à l'intersection des domaines des bases de données, de l'intelligence artificielle, de la statistique, des interfaces homme/machine et de la visualisation. À partir de données collectées par des experts, il s'agit de proposer des connaissances nouvelles qui enrichissent les interprétations du champ d'application, tout en fournissant des méthodes automatiques qui exploitent cette information.

L'ECBD est classiquement décrite comme un processus interactif de préparation des données (sélection de descripteurs, constitution d'une table, discrétisation), d'extraction de connaissances à l'aide d'algorithmes de calcul, de visualisation et d'interprétation des résultats, lors d'interactions avec l'expert (voir figure 1). Les méthodes d'exploration proposent des solutions aux problèmes de recherche d'associations, de classification supervisée et non supervisée.

Plus précisément, la *fouille de données* (data-mining en anglais) concerne l'étape algorithmiquement difficile de ce processus, qui produit des motifs potentiellement intéressants à partir des données booléennes.

Fréquemment exprimée sous forme de règles, la connaissance extraite requiert la mise au point d'algorithmes efficaces pour prendre en compte les difficultés algorithmiques ou liées aux caractéristiques du problème. Les bases de données utilisées comprennent couramment la description de millions d'objets par des milliers d'attributs et l'espace de recherche est de taille exponentielle en nombre d'attributs. Plusieurs

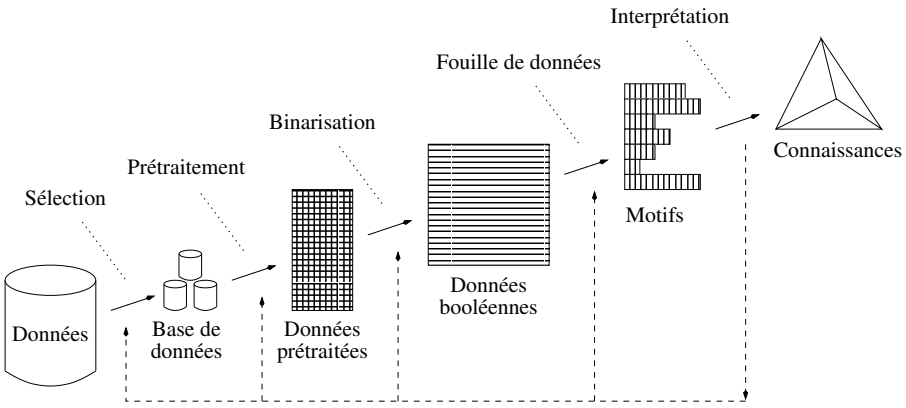

FIG. 1 – Processus d'extraction de connaissances.

problèmes NP-difficiles (pour lesquels on ne dispose pas d'algorithme en temps polynomial) se cachent en particulier derrière la recherche des motifs fréquents (ensembles d'attributs communs à plusieurs objets), étape préalable à la construction de règles associant des motifs.

2 Définitions

Les bases de données considérées ici sont de simples tables contenant l'information, éventuellement construites par jointures à partir de plusieurs relations. L'exemple du tableau 1 répertorie les valeurs de trois *attributs* multi-valués X_1, X_2 et X_3 pour 8 *objets* d'étude, appelés également n-uplets. Dans cet exemple, les deux premiers attributs X_1 et X_2 sont de type symbolique ou qualitatif car leur domaine de définition est discret. *A contrario*, le dernier attribut X_3 est numérique ou quantitatif.

	attributs		
objets	X_1	X_2	X_3
o_1	$+$	$\rightarrow$	0.2
o_2	$-$	$\rightarrow$	0
o_3	$+$	$\rightarrow$	$0,1$
o_4	$+$	$\leftarrow$	$0,4$
o_5	$-$	$\rightarrow$	$0,6$
o_6	$-$	$\rightarrow$	$0,5$
o_7	$+$	$\leftarrow$	1
o_8	$-$	$\leftarrow$	$0,8$

TAB. 1 – Exemple d'une base de données au format attribut/valeur.

Cet article se concentre sur l'extraction de *motifs* ensemblistes, où un motif est un

ensemble d'attributs booléens. Cela nécessite de discrétiser les attributs numériques, afin de disposer de données booléennes. Il sort du cadre de cet exposé de discuter précisément des méthodes de discrétisation qui permettent d'obtenir de tels contextes booléens à partir d'attributs multi-valués ou continus. Disons simplement que cette étape de prétraitement des données est difficile dans le cas d'attributs numériques : il faut regrouper ensemble des valeurs différentes qui expriment la même information, ou définir des intervalles. Les connaissances des experts se révèlent indispensables pour effectuer les bons choix lors de cette opération délicate. Nous supposons donc obtenir (tableau 2) une matrice booléenne qui indique pour chaque objet les attributs qu'il contient. Ce format est usuellement qualifié de *transactionnel*. Dans ces contextes booléens, un attribut est souvent appelé *item* et un objet *transaction*.

	attributs						
objets	a_1	a_2	a_3	a_4	a_5	a_6	a_7
o_1	×		×		×		
o_2		×	×		×		
o_3	×		×		×		
o_4	×			×		×	
o_5		×	×			×	
o_6		×	×			×	
o_7	×			×			×
o_8		×		×			×

TAB. 2 – Exemple d'une base de données au format transactionnel.

Une base de données booléenne r est notée sous la forme d'un contexte formel $(\mathcal{A}, \mathcal{O}, R)$ où $\mathcal{A} = \{a_1 \dots a_m\}$ est l'ensemble des attributs, $\mathcal{O} = \{o_1 \dots o_n\}$ celui des objets et R une relation binaire entre $\mathcal{A}$ et $\mathcal{O}$.

3 Exemples de données

Les méthodes de fouille de données trouvent leur origine dans l'analyse des achats effectués par les consommateurs des grands magasins. Dans ce cadre, les bases de données étudiées concernent des objets-consommateurs décrits par des attributs relatifs aux produits achetés. La particularité de ces données est leur grande dimension : on dispose couramment de renseignements sur des millions de consommateurs à propos de milliers de références. En revanche, ces données sont peu denses, car les consommateurs effectuent leur choix dans une faible portion de l'ensemble des articles disponibles.

Les données médicales fournissent de nombreuses données à étudier. Les objets sont dans ce cas les patients d'une étude thérapeutique et les attributs sont leurs caractéristiques : âge, sexe, poids, taille, résultats d'examens, analyses biologiques, etc. Les dimensions de ces données sont nettement plus faibles que pour les transactions commerciales : quelques milliers d'objets, une centaine d'attributs. Cependant, ces données sont très denses et posent de nombreux problèmes algorithmiques.

Plus récemment, l'analyse du génome fournit de grandes quantités de données relatives à l'expression des gènes dans différentes situations biologiques. Ces données sont obtenues à l'aide de puces à ADN, où un grand nombre de réactions entre des gènes et un échantillon cellulaire sont analysées par des procédés optiques. Les données correspondantes concernent des objets qui sont les situations biologiques analysées, caractérisés par des attributs qui définissent l'expression des gènes dans cette situation. Du fait du coût financier important de ces techniques, il y a un nombre faible d'objets d'étude (une centaine) mais un grand nombre (plusieurs milliers) d'attributs représentant les gènes. Ce format inhabituel de données, plus larges que hautes, pose de grandes difficultés algorithmiques notamment résolues par des opérations à base de transposition de matrice.

Enfin, l'émergence des techniques de fouille de données permet de toujours plus élargir le champ d'application de ces méthodes. Par exemple, l'analyse de procédés industriels, le traitement du signal numérique bénéficient de plus en plus de l'apport de ces techniques novatrices.

4 Méthodes d'exploration des données

Les méthodes d'exploration des données fournissent à l'expert des solutions pour l'aide à la décision. On distingue généralement deux types de données :

supervisées : chaque objet étudié est étiqueté par une valeur de *classe*. Par exemple, s'il s'agit de données médicales concernant des patients, la classe définit le degré d'atteinte de la maladie. Pour des produits de fabrication industrielle, la classe est déterminée par la qualité de fabrication ;

non supervisées : aucune classe n'est attribuée *a priori*.

Suivant le type des données, les décisions concernent :

la classification supervisée : elle consiste à proposer une valeur de classe pour un objet dont la classe est inconnue. Un médecin peut ainsi adapter le traitement d'un patient en fonction de ses attributs ;

la classification non supervisée (ou *clustering*) : cette méthode permet de constituer des groupes homogènes d'objets, pour par exemple grouper des patients qui ont le même comportement ;

le calcul d'associations : elles formulent les corrélations présentes dans les données et sont utilisées à des fins de classification ou de caractérisation de classe.

4.1 Classification supervisée

Avant l'apparition récente des techniques de fouille de données à base de motifs, les méthodes reconnues concernent l'utilisation de réseaux de neurones, de réseaux bayésiens et les arbres de décision.

4.1.1 Réseaux de neurones

Les réseaux neurones (ou réseaux connexionnistes) utilisent l'analogie avec l'architecture physiologique du cerveau humain : les neurones sont des entités élémentaires qui reçoivent des signaux en entrée et transmettent à d'autres neurones des signaux de sortie qui résultent d'une combinaison des signaux d'entrée. Les premiers neurones d'entrée sont reliés aux valeurs des attributs d'un objet. Par exemple pour la reconnaissance d'images, ce sont les pixels allumés ou éteints. Les neurones de sortie indiquent la valeur finale de la décision, c'est-à-dire la classe de l'objet. Des neurones intermédiaires sont organisés en couches et l'ensemble constitue un réseau.

Pendant sa phase d'apprentissage, des objets sont présentés au réseau et lorsque la réponse diffère de la classe supervisée, un algorithme de rétro-propagation modifie les comportements des neurones intermédiaires. Techniquement, un tel réseau calcule des équations d'hyper-plan séparateurs des classes, selon un algorithme de descente de gradient.

Les réseaux de neurones ont connu un rapide succès, particulièrement pour le traitement des images. Cependant, il est très difficile d'expliquer comment la décision est rendue par ces réseaux, du fait de la grande complexité de leur architecture.

4.1.2 Réseaux bayésiens

Fondés sur la notion de probabilité conditionnelle, les réseaux bayésiens permettent de calculer la distribution jointe sur un ensemble de données à l'aide de procédés stochastiques. Leur architecture est obtenue par apprentissage à partir des données, mais cette étape reste la partie difficile de la mise en œuvre de cette technique. En revanche, les décisions rendues par ces réseaux sont plus compréhensibles que celles fournies par les neurones.

4.1.3 Arbres de décision

Un arbre de décision permet de représenter les objets étudiés sous une forme arborescente, selon une hiérarchie des attributs déterminée par un calcul d'entropie. Ces méthodes sont populaires pour la présentation synthétique des données qu'elles fournissent, ainsi que pour la clarté des explications concernant la décision rendue.

4.2 Classification non-supervisée

Les algorithmes classiques de classification non-supervisée sont les méthodes à base de k-moyenne ou de nuées dynamiques et permettent de segmenter l'ensemble des objets en un nombre défini de classes homogènes. À partir de centroïdes choisis aléatoirement, les classes sont déterminées par fusion ascendante, de manière à minimiser la distance inter-classes et à maximiser la distance extra-classes. De façon duale, une classification peut être réalisée selon une technique hierarchique descendante : l'ensemble des objets est découpé en classes de plus en plus fines par l'utilisation d'une distance.

La mise au point d'une distance adéquate reste un problème épineux pour ces méthodes. Elles souffrent également d'un manque de lisibilité des décisions rendues.

5 Fouille de données

5.1 Motivations

La fouille de données est motivée par les arguments suivants :
- les données traitées concernent à la fois des attributs qualitatifs et quantitatifs, ce qui justifie les étapes de discrétisations pour obtenir des contextes booléens ;
- les données sont volumineuses : des objets par millions, des attributs par milliers. Ces caractéristiques posent de nombreux problèmes algorithmiques ;
- la fouille de données poursuit un but d'exhaustivité des connaissances découvertes. À la différence des techniques statistiques, ce ne sont pas seulement les tendances globales des données qui sont recherchées mais également des propriétés locales qui concernent un petit nombre d'objets ;
- dans l'optique des méthodes d'exploration qui permettent d'aider l'expert dans sa prise de décision, il est souhaitable que l'aide fournie soit clairement justifiée, expliquée et compréhensible.

5.2 Définitions

Notre travail porte sur l'extraction de motifs d'attributs dans les bases de données, qui sont des sous ensembles de $\mathcal{A}$, mais nous parlons également de motifs d'objets (sous ensemble de $\mathcal{O}$) dans ce document. Lorsqu'aucune précision n'est indiquée, un *motif* est un motif d'attributs. Pour alléger les écritures, ces motifs seront notés sous forme de chaîne plutôt que sous forme d'ensembles (*i.e.* a_1a_2 au lieu de $\{a_1, a_2\}$) et le signe *union* $\cup$ sera omis dans les expression (XY remplace $X \cup Y$). Dans un contexte formel, nous considérons indifféremment un objet o comme un élément de $\mathcal{O}$ ou comme un motif d'attributs : $o = \{a \in \mathcal{A} \mid R(a, o) = present\}$.

Les notions de support et de fréquence d'un motif sont essentielles pour caractériser la représentativité des motifs découverts dans les données : un objet $o \in \mathcal{O}$ supporte le motif d'attributs X (ou X est présent dans o) si $X \subseteq o$ (ou $\forall a \in X, R(a, o) = present$). Le support $supp(X)$ d'un motif X d'attributs est l'ensemble des objets qui le supportent. Sa fréquence $\mathcal{F}(X)$ est le cardinal du support.

La notion de motif *fréquent* est centrale pour la fouille de données et désigne les motifs dont la fréquence dépasse un seuil γ fixé par l'utilisateur. Sur notre exemple, les objets o_4 et o_7 contiennent le motif a_1a_4, son support est o_4o_7. Sa fréquence vaut 2, et si le seuil de fréquence est fixé à 1 ou 2, ce motif est fréquent.

Pour exprimer les corrélations entre les attributs, les *règles d'association* basées sur $Z = XY$ sont des expressions de la forme $X \rightarrow Y$. X est la prémisse, Y est la conclusion. La fréquence de la règle est celle de XY. La confiance, notée $conf(X \rightarrow Y)$, est la proportion d'objets contenant X qui contiennent aussi Y : $conf(X \rightarrow Y) = \mathcal{F}(X \cup Y)/\mathcal{F}(X)$. Sur notre exemple, la règle $a_1a_3 \rightarrow a_5$ est exacte (les fréquences de a_1a_3 et de $a_1a_3a_5$ sont égales et valent 2). La règle $a_2a_3 \rightarrow a_6$ a une confiance de $2/3$ ($\mathcal{F}(a_2a_3) = 3$ et $\mathcal{F}(a_2a_3a_6) = 2$).

Même si dans la pratique les possesseurs de bases sont plus intéressés par les règles présentes dans les données que par les motifs fréquents, celles-ci sont dépendantes de l'obtention de ces motifs. Le calcul de ces règles n'est pas difficile lorsque l'on connaît

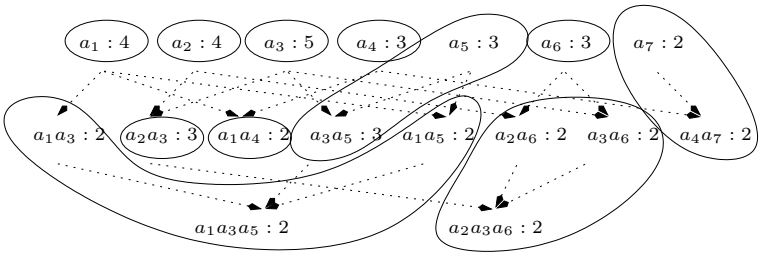

FIG. 2 – Calcul des motifs fréquents sur l'exemple de la table 2.

les motifs fréquents. L'extraction des motifs fréquents est clairement identifiée dans la communauté fouille de données comme l'étape algorithmiquement difficile préalable à la formation de règles.

5.3 Calcul de motifs fréquents

Extraire les motifs fréquents d'une base de données est un problème de recherche. Il est caractérisé par un espace à parcourir, pour découvrir des éléments satisfaisant la contrainte de fréquence. Cet espace est celui des parties de $\mathcal{A}$, de taille exponentielle en la taille de $\mathcal{A}$. Il est donc illusoire de pouvoir le parcourir exhaustivement et il faudra utiliser une stratégie, dérivée des propriétés de la contrainte de fréquence.

Cette contrainte de fréquence est anti-monotone, c'est-à-dire que si un motif est fréquent, ses sous-ensembles le sont également. Réciproquement, si un motif n'est pas fréquent, tous ses sur-ensembles ne le sont pas non plus. Une stratégie couramment employée consiste donc à parcourir l'espace de recherche par niveaux, en examinant les motifs potentiellement fréquents par ordre croissant de leurs tailles. La figure 2 schématise ce procédé sur les données de la table 2 pour une fréquence minimale de 2 : on commence par calculer la fréquence des singletons, puis celle des paires. Les paires infréquentes sont éliminées et leurs sur-ensembles élagués. Enfin les motifs de longueur trois sont obtenus par fusion des paires fréquentes.

5.4 Génération de règles d'association

À partir d'un motif fréquent $Z = XY$, on construit les règles d'association $X \rightarrow Y$ en respectant l'exigence de confiance minimale. Sur notre exemple, $a_1a_3a_5$ est fréquent et si la confiance minimale est $2/3$, on obtient les règles $a_1a_3 \rightarrow a_5$ (confiance 1), $a_1a_5 \rightarrow a_3$ (1), $a_3a_5 \rightarrow a_1$ (2/3) et $a_5 \rightarrow a_1a_3$ (2/3). Les règles $a_1 \rightarrow a_3a_5$ (2/4) et $a_3 \rightarrow a_1a_5$ (2/5) ne sont pas construites car leur confiance est trop faible.

Cette technique produit de nombreuses règles redondantes. Par exemple, la règle $a_5 \rightarrow a_3$ est exacte (confiance 1), donc pour tout attribut a_i, $a_ia_5 \rightarrow a_3$ est aussi exacte, mais redondante. On cherchera donc à optimiser la production de règles en ne conservant que celles dont les prémisses sont minimales, et les conclusions maximales. Ces règles particulières sont qualifiées d'*informatives*. Pour cela, remarquons que nous avons représenté sur la figure 2 des classes d'équivalence de support, c'est-à-dire des

classes qui regroupent les motifs présents dans les mêmes objets. Ces classes possèdent des éléments minimaux définissant les prémisses des règles informatives, et un élément maximum qui est la conclusion.

6 Applications

Les applications de la fouille de données concernent la classification supervisée, non supervisée, et la mise en évidence d'association à des fins de caractérisation de classe.

La classification supervisée est rendue possible à l'aide de règles d'association qui concluent sur une valeur de classe. Grâce à des méthodes de vote, la classe d'un objet nouveau peut être déterminée par les règles dont la prémisse est contenue par l'objet.

La classification non supervisée peut être réalisée grâce aux motifs maximaux des classes d'équivalence, appelés motifs *fermés*. Ces motifs sont des rectangles maximaux dans la matrice des données et définissent des *concepts*, associant un motif fermé d'attributs et un motif fermé d'objets. Les concepts définissent donc des groupes homogènes d'objets qui partagent les mêmes attributs. La classification non supervisée finale est effectuée grâce à des mesures d'intérêt de ces rectangles, par exemple fondées sur des notions de divergence d'aires relatives (cf. figure 3).

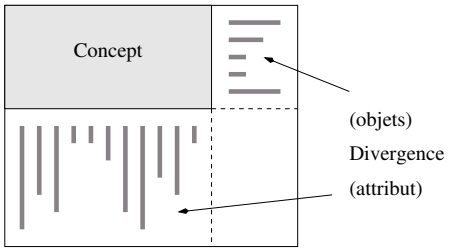

FIG. 3 – Mesures d'intérêt pour les concepts.

Les concepts sont également particulièrement utiles pour traiter les bases de données comportant un grand nombre d'attributs devant le nombre d'objets. En effet, ces rectangles maximaux sont invariants par transposition (cf. figure 4). Il est donc possible de réaliser les extractions de motifs dans la transposée de la base de données qui possède des dimensions moins rédhibitoires. Cette technique est applicable dans le cas des données génomiques.

Enfin, dans un but de caractérisation de classes, les motifs sont caractérisés par un taux de croissance qui représente le rapport de leur fréquence dans une classe et dans le reste de la base. Ces motifs *émergents* permettent de définir une signature de l'appartenance à une classe (voir figure 5).

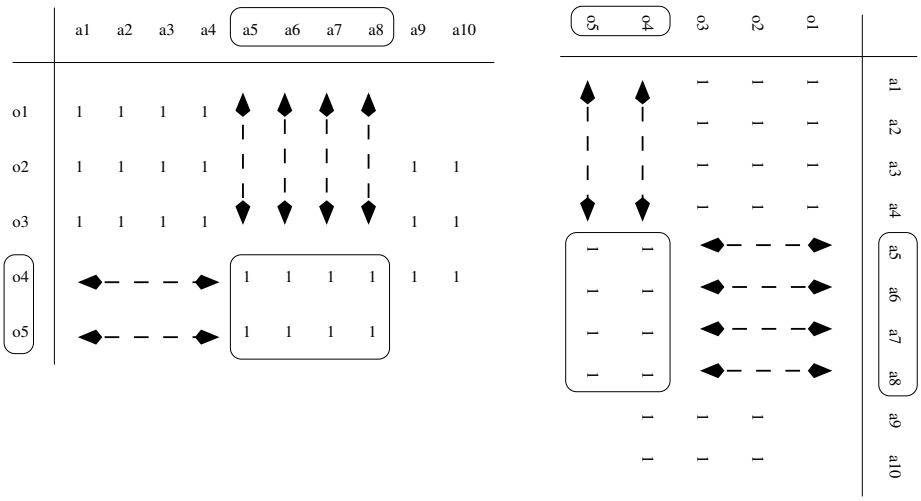

FIG. 4 – Transposition de base de données.

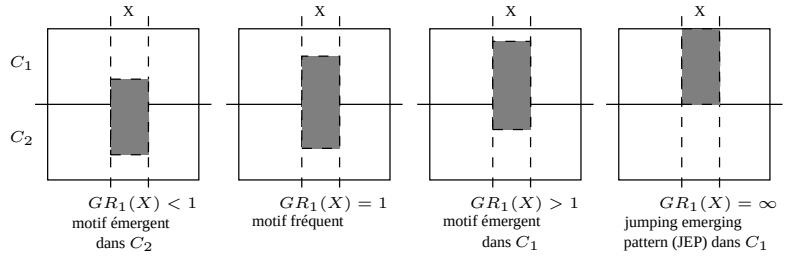

FIG. 5 – Motifs émergents pour la caractérisation de classes.

7 Conclusion - perspectives

La fouille de données orientée motifs est désormais bien maîtrisée pour les contextes de données classiques : transactions commerciales, données médicales et biologiques, processus industriels, etc. Les usages des motifs et règles découverts sont nombreux et autorisent la mise au point de méthodes de classification supervisée, non supervisée, ou de recherche d'associations à des fins de caractérisation de classe.

La communauté scientifique se tourne désormais vers des problèmes nouveaux. Il s'agit essentiellement du traitement de l'information temporelle et spatiale. La recherche de motifs séquentiels est cruciale pour l'analyse des fluctuation des cours de bourse, des séquences d'ADN, le traitement des connexions à des services internet ou l'analyse des rapports de systèmes d'alarme et de sécurité. Dans ce cadre, l'exploitation de grands flux de données, par exemple en provenance de capteurs sismiques, revêt une importance toujours plus grande.