Les modèles des experts au service de l'extraction de motifs pertinents *

F. Flouvat[1] J. Sanhes[1] C. Pasquier[1,2] N. Selmaoui-Folcher[1] J-F. Boulicaut[3]

[1] PPME, Université de la Nouvelle Calédonie, BP R4, F-98851 Nouméa, Nouvelle Calédonie
[2] Institut de Biologie Valrose, CNRS UMR7277 - INSERM U1091, F-06108 Nice, France
[3] Université de Lyon, CNRS, INSA-Lyon, LIRIS UMR5205, F-6962, Lyon, France

{frederic.flouvat,jeremy.sanhes,frederic.flouvat,claude.pasquier,nazha.selmaoui}@univ-nc.nc
jean-francois.boulicaut@insa-lyon.fr

Résumé

Pour assister la découverte de connaissances à partir de données, de nombreuses techniques de calcul de motifs ont été proposées. L'un des verrous à leurs dissémination est que nombre des motifs extraits apparaissent triviaux et/ou inintéressants au regard de la connaissance du domaine et des experts. La fouille de données sous contraintes ne permet qu'une prise en compte limitée de ces connaissances et de l'intérêt dit subjectif. Pourtant, il existe souvent des modèles mathématiques qui capturent une partie importante de la connaissance experte. Nous proposons ici d'exploiter de tels modèles pour dériver des contraintes utilisables au cours des processus de fouille et ainsi améliorer la pertinence des motifs calculés tout en gagnant en performances. L'approche est générique mais nous l'étudions empiriquement dans le cas de modèles de l'aléa érosion pour améliorer la pertinence de motifs ensemblistes dans des données réelles.

Mots Clef

Fouille de données, découverte de motifs, connaissance du domaine, modèles experts, contraintes.

Abstract

To support knowledge discovery from data, many pattern mining techniques have been proposed. One of the bottlenecks for their dissemination is the number of computed patterns that appear to be either trivial or uninteresting with respect to available knowledge. Constraint-based data mining supports a limited use of domain knowledge and thus lack to assess pattern relevancy and their subjective interestingness. It turns out that we often have at hand mathematical models that capture part of the domain and expert knowledge. We propose here to exploit such models to derive constraints that can be use during the data mining phase to improve both pattern relevancy and computational efficiency. Even though the approach is generic, it is illustrated on set pattern discovery from real data for studying the erosion risk.

Keywords

Data mining, pattern discovery, domain knowledge, models, constraints.

1 Introduction

Les experts de domaines scientifiques variés (e.g., des géologues, des physiciens ou des épidémiologistes) expriment souvent une partie de leurs connaissances sur certains phénomènes au moyen de modèles mathématiques. Par exemple, les experts en érosion des sols ont développé des modèles permettant d'estimer le risque d'érosion en fonction d'un ensemble de paramètres environnementaux [13, 4]. De même les épidémiologistes estiment le nombre de personnes potentiellement infectées par une maladie [6]. Ces modèles experts présentent l'avantage de synthétiser une partie de la connaissance du domaine dans un contexte donné. Cependant, ce sont fondamentalement des simplifications du réel et, du fait des progrès dans les capteurs et la collecte de données scientifiques, nous disposons souvent des valeurs de nombreux paramètres qui ne sont pas pris en compte dans les modèles disponibles, sans que l'on sache d'ailleurs s'ils devraient l'être ! Dans ces contextes d'explosion des masses de données disponibles, on peut vouloir utiliser les méthodes de fouille de données sous contraintes pour découvrir des motifs qui permettront ensuite d'assister la découverte de connaissances sur les phénomènes étudiés. On utilise différents domaines de motifs comme, e.g., les ensembles ("itemsets") et règles d'association dans des données booléennes, les motifs séquentiels ou les règles d'épisodes dans des (collections de) séquences, ou encore des sous-graphes dans des (collections de) graphes. La fouille sous contraintes a été étudiée pour à la fois permettre l'expression de l'intérêt objectif des motifs (en imposant, e.g., une fréquence minimale [1]) ou l'intérêt subjectif, qu'il s'agisse de celui de motifs (e.g., [14, 16]) ou de collections de motifs (e.g., [17]). Ainsi, la notion subjective d'étonnement peut se spécifier en carac-

*Ce travail a été financé par le contrat ANR-2010-COSI-012-01 FOSTER.

térisant tout de ce qui est attendu ou connu [15, 9]. De fait, la fouille sous contraintes permet non seulement de mieux gérer la pertinence des motifs calculés mais encore, le plus souvent, d'exploiter les propriétés des contraintes (e.g., des propriétés de monotonicité) pour réaliser des extractions complètes et efficaces.

Travailler à l'intégration de la connaissance des experts dans les processus de fouille n'est pas nouveau [2, 3, 7, 5]. Cette connaissance est souvent exprimée sous la forme de règles/contraintes expertes (e.g., de la forme *si ... alors ...*) définies manuellement. Ce type d'explicitation est difficile à obtenir et ne représente que très partiellement la connaissance du domaine : en pratique, elle reste limitée à quelques règles basiques. Très tôt, des algorithmes de fouille ont exploité des taxinomies ou hiérarchies explicitées sur, e.g., les attributs, pour fournir des motifs plus pertinents. Notons également les possibilités offertes par des modèles graphiques comme les réseaux bayésiens : des experts peuvent expliciter certaines dépendances connues entre les variables de sorte que l'on puisse les exploiter comme source de connaissance du domaine [8].

En constatant que dans de nombreux contextes de science des données, les experts ont souvent capitalisé une partie de leur connaissance dans des modèles mathématiques, l'originalité de notre approche consiste à exploiter de tels modèles pour dériver de nouvelles contraintes qui vont pouvoir être intégrées aux calculs des motifs sous contraintes et ainsi améliorer la pertinence des extractions. Plus précisément, nous ciblons un domaine de motif de type "itemsets" et nous nous intéressons aux modèles mathématiques qui prennent la forme de fonctions à plusieurs variables. Nous présentons des exemples de modèles linéaires, polynomiaux, mais aussi des modèles non linéaires qui peuvent être utilisés pour améliorer la pertinence des motifs calculés. Nous mettons en évidence certaines propriétés des modèles et du domaine de motif des itemsets qui vont permettre des élagages sûrs dans l'espace de recherche des motifs et donc améliorer l'efficacité des calculs.

Dans cet article, nous prendrons pour exemple les modèles développés par les experts en érosion des sols. Dans ce contexte, notre approche se focalise sur l'extraction de motifs susceptibles d'être liés à une érosion forte tout en permettant d'étudier l'influence d'autres paramètres environnementaux (traduisant, e.g., l'impact anthropique). Les contraintes exprimées à partir de ces modèles permettront donc de compléter la connaissance des experts ou, au contraire, de mettre en avant des relations ou corrélations contradictoires.

La section 2 présente le cas d'étude de l'érosion. La section 3 introduit notre cadre théorique et la section 4 présente notre contribution sur la définition de contraintes à partir de modèles experts. La section 5 décrit des résultats expérimentaux sur des jeux de données réelles. La section 6 conclut et donne quelques perspectives.

2 Le cas d'étude de l'érosion des sols

L'érosion des sols a un très fort impact sur l'Homme dans de nombreuses régions du monde. Les géologues et les géographes ont développé des modèles mathématiques visant à estimer le risque d'érosion d'un sol. Deux grandes classes de modèles peuvent être distinguées : les modèles empiriques et les modèles physiques.
Les modèles empiriques sont construits à partir de connaissance experte et d'expérimentations. Le modèle USLE [1] [18] et le modèle proposé dans [4] en sont des exemples typiques (le premier est un modèle polynomial et le second est linéaire). Les modèles physiques sont des modèles quantitatifs fondés sur des propriétés physiques et calibrés à partir des données expérimentales. Par exemple, les modèles WEPP (*Water Erosion Prediction Project*) [11] et RMMF (*Revised Morgan-Morgan Finney*) [13] sont basés sur plusieurs modèles physiques qui sont non linéaires et non polynomiaux. Le modèle RMMF divise le processus d'érosion en deux phases : détachement par gouttes de pluie (cf. exemple en Figure 1) et détachement par ruissellement. Chaque phase est liée à un sous-modèle physique. Les résultats des deux sous-modèles sont ensuite additionnés pour estimer la perte en sol annuelle.

Paramètres	**Domaine de valeurs**
Indice de détachement du sol (en g/J) x_K	défini en fonction du type de sol
Précipitation annuelle (en mm) x_R	[0 , 12 000]
Proportion de pluie interceptée par la végétation x_A	[0 , 1]
Pourcentage de couverture de la canopée x_{CC}	[0 ,1]
Intensité de la pluie (en mm/h) x_I	{10, 25, 30} en fonction du climat de la zone d'étude
Hauteur de la végétation (en m) x_{PH}	[0 , 130]

$F(x_K, x_R, x_A, x_{CC}, x_I, x_{PH}) = x_k \times [x_R \times x_A \times (1 - x_{CC}) \times (11.9 + 8.7\, log\, x_I) + (15.8 + {x_{PH}}^{0.5}) - 5.87] \times 10^{-3}$

FIGURE 1 – Modélisation du détachement par gouttes de pluie dans RMMF

3 Définitions préliminaires

3.1 Les modèles des experts

Soit $D_f = \{x_1, x_2, ..., x_n\}$ l'ensemble des **variables** du modèle expert. On note $dom(x_j)$, **domaine de valeur** de x_j, ensemble des valeurs possibles pour la variable x_j. Ces variables peuvent être numériques ou catégorielles. Soit $dom_{num}(x_j)$ le **domaine de valeur "numérisé"** de x_j. Si x_j est numérique, $dom_{num}(x_j) = dom(x_j)$. Si x_j est nominale, $dom_{num}(x_j) = \{num_{x_j}(v), v \in dom(x_j)\}$, où $num_{x_j} : dom(x_j) \rightarrow \mathbb{R}$ une fonction définie par l'expert codant les variables nominales de x_j en nombres. Par exemple, l'occupation du sol dans le modèle RMMF (variable x_K) est codé par des entiers représentant le type d'occupation du sol (e.g., "sol nu").
Un **modèle mathématique expert** est une fonction $f : dom_{num}(x_1) \times dom_{num}(x_2) \times ... \times dom_{num}(x_n) \rightarrow \mathbb{R}$ qui

1. Universal Soil Loss Equation

représente la connaissance d'un ou plusieurs experts pour un phénomène.

Pour illustrer les concepts utilisés, nous considérons le modèle "jouet" suivant :

$f : [1, 16] \times [0, 4\pi] \times [1, 100] \subset \mathbb{R}^3 \rightarrow [0, 5] \subset \mathbb{R}$

$f(x_1, x_2, x_3) = \sqrt{x_1} - cos(x_2)/2 \times log(x_3)$

3.2 L'extraction d'itemsets sous contraintes

Dans cette partie, nous allons faire le lien entre la définition d'itemsets ([1]) et les modèles définis précédemment.

Soit $D_{\mathcal{D}} = \{d_1, d_2, ..., d_{n'}\}$ les **dimensions d'analyse** présentes dans la base de données $\mathcal{D}$ (e.g., "type de sol" ou "végétation"). $D_{\mathcal{D}}$ doit recouvrir une partie des variables du modèle expert, soit $D_{\mathcal{D}} \cap D_f \neq \emptyset$. Le domaine des valeurs de $d_{j'}$ dans $\mathcal{D}$, est noté $dom(d_{j'})$. Les algorithmes d'extraction de motifs ne traitant que des attributs catégoriels, ces domaines de valeur sont généralement discrétisés. Nous notons $dom_{categ}(d_{j'})$ le **domaine de valeur discrétisé** de la dimension $d_{j'}$. Si le domaine de $d_{j'}$ est nominal ou numérique discret, $dom_{categ}(d_{j'}) = dom(d_{j'})$. Par exemple, le domaine de la dimension "précipitations" x_R en Figure 1 est $dom(x_R) = [0, 12000]$ et son domaine discrétisé est $dom_{categ}(x_R) = \{$"$x_R \in [0, 2000]$", "$x_R \in [2001, 3200]$", "$x_R \in [3201, 12000]$"$\}$. Dans la suite, afin de faciliter la compréhension des itemsets pris en exemple, nous intégrons systématiquement la dimension/variable dans les items issues de variables du modèle (e.g., "$x_K = sol\ nu$" à la place de "sol nu").

Une **relation d'ordre total**, notée $\prec_{x_j}$, est définie pour chaque dimension $x_j \in D_{\mathcal{D}} \cap D_f$. Si $i, i' \in dom_{categ}(x_j)$ sont des valeurs nominales, $i \prec_{x_j} i'$ ssi $num_{x_j}(i) < num_{x_j}(i')$. Si $i, i' \in dom_{categ}(x_j)$ sont des valeurs numériques discrètes, $i \prec_{x_j} i'$ ssi $i < i'$. Enfin, si $i, i' \in dom_{categ}(x_j)$ représentent des intervalles, i.e., $i =$ "$x_j \in [inf_i, sup_i]$" et $i' =$ "$x_j \in [inf_{i'}, sup_{i'}]$", $i \prec_{x_j} i'$ ssi $sup_i < inf_{i'}$. Par exemple, "$x_R \in [0, 2000]$" $\prec_{x_R}$ "$x_R \in [2001, 3200]$" car $2000 < 2001$.

Soit $\mathcal{I} = \bigcup_{d_j \in D_{\mathcal{D}}} dom_{categ}(d_j)$ l'ensemble des valeurs de la base de données $\mathcal{D}$. Une valeur $i \in \mathcal{I}$ est appelée un **item**. Le langage des motifs est $\mathcal{L} = 2^{|\mathcal{I}|} \backslash \{X \in 2^{|\mathcal{I}|} \mid \exists i, i' \in X, i \neq i',\ i, i' \in dom_{categ}(d_j), d_j \in D_{\mathcal{D}}\}$. Un ensemble d'items $X = \{i_1, i_2, ..., i_k\} \in \mathcal{L}$, $k \leq n'$, est appelé un **itemset**. La **couverture d'un itemset X par rapport à** D_f, notée $cover(X, D_f)$, est l'ensemble des variables/dimensions des items de X appartenant à D_f. La couverture de X par rapport à D_f est totale ssi $cover(X, D_f) = D_f$. Par exemple, $Z = \{$"$x_K = sol\ nu$", "$x_R = [2001, 3200]$", "$mine$", "$piste$"$\}$ est un itemset. Sa couverture par rapport aux variables du modèle REP précédent est $cover(Z, RMMF) = \{x_K, x_R\}$.

Il est possible d'étendre la relation d'ordre $\prec_{x_j}$ à deux itemsets $X, Y \in \mathcal{L}$. On obtient ainsi une relation d'ordre partiel entre deux itemsets par rapport à une variable/dimension x_j. L'itemset X est inférieur à l'itemset Y sur la variable $x_j \in D_{\mathcal{D}} \cap D_f$ (noté $X \prec_{x_j} Y$), ssi $i \prec_{x_j} i'$ avec $i \in X$, $i' \in Y$, $i, i' \in dom_{categ}(x_j)$. Ainsi, si $X = \{$"$x_K = sol\ nu$", "$x_R = [0, 2000]$", "$mine$"$\}$ et $Y = \{$"$x_R = [2001, 3200]$", "$mine$", "$piste$"$\}$, on a $X \prec_{x_R} Y$.

Le problème de fouille que nous traitons consiste à calculer l'ensemble des motifs qui satisfont un prédicat de sélection q dans les données. Cet ensemble, parfois appelé la théorie de $\mathcal{D}$ par rapport à $\mathcal{L}$ et q, est noté $\mathcal{T}h(\mathcal{L}, \mathcal{D}, q)$ [12], avec $\mathcal{T}h(\mathcal{L}, \mathcal{D}, q) = \{X \in \mathcal{L} \mid q(X, \mathcal{D})$ est vrai$\}$. Le prédicat de sélection est, en général, une conjonction de contraintes primitives parmi lesquelles nous aurons souvent une contrainte de fréquence minimale (i.e., sélection des itemsets dont la fréquence d'apparition est supérieure à un seuil fixé par l'utilisateur). Dans les sections suivantes, nous allons introduire une nouvelle contrainte basée sur un modèle des experts, et montrer comment elle peut être intégrée dans ce cadre général.

4 Des modèles aux contraintes

Une approche simple pour exploiter la représentation de la connaissance du domaine est d'en dériver des contraintes primitives à prendre en compte lors des extractions. Nous proposons ici de définir des contraintes dérivées des conditions des modèles à appliquer aux motifs plutôt que des contraintes définies "à la main". Ces contraintes représentent bien plus que de simples règles "si ... alors ..." car un seul modèle expert peut condenser l'information sur un nombre considérable de telles règles. Rappelons que les contraintes dérivables des modèles dont nous parlons dans cette section seront utilisées conjointement avec celles qui auront été spécifiées par les analystes lorsqu'ils définissent l'intérêt a priori des motifs.

Différents types de contraintes vont pouvoir être dérivés en fonction des données, des modèles utilisés, mais aussi du problème étudié. Dans le cadre de cet article, nous nous focalisons plus particulièrement sur une contrainte qui apparaît similaire à une contrainte de fréquence minimale même si ses propriétés sont très différentes.

Tout comme la contrainte de fréquence minimale, nous pouvons définir une contrainte filtrant les motifs X tel que $f(X)$ (i.e., la prévision du modèle f pour l'itemset X) est supérieure ou égale à un seuil. En fonction de ce que modélise f, cette contrainte aura différentes sémantiques.

Par exemple, si f estime la perte en sol (en kg/m^2 par an) comme c'est le cas dans le modèle RMMF, cette contrainte permettra de ne conserver que les motifs susceptibles de représenter une perte en sol (et donc une érosion) supérieure à une certaine quantité. Dans cette application, on peut distinguer deux cas en fonction de la disponibilité de données "terrain" sur le phénomène modélisé (la perte en sol).

En l'absence de "vérité terrain" (i.e., de mesures sur la perte en sol), cette contrainte permettra de mettre en avant si de telles pertes risquent d'être fréquentes dans la zone d'étude et dans quelles situations (grâce aux valeurs des autres paramètres environnementaux décrits par les motif extraits). Elle peut également permettre d'identifier (si les motifs sont fréquents) à quels autres facteurs, non couverts

par le modèle, ces pertes en sols sont fréquemment liées dans les données.
En présence de "vérité terrain", cette contrainte permettra de comparer la prévision du modèle des experts à la réalité des données collectées. Les motifs confirmant le modèle sont intéressants car ils sont doublement validés par la vérité terrain et la connaissance du domaine (i.e., le modèle expert). De plus, les items additionnels du motif peuvent compléter les explications du modèle. Les motifs contredisant le modèle des experts sont tout aussi intéressants car ils permettent de mettre en avant certaines spécificités non prises en compte dans les modèles experts utilisés, déterminant donc des possibilités de révision.
Soit $X \in \mathcal{L}$ un itemset, f un modèle déjà développé par des experts et un seuil dans $\mathbb{R}$, la forme générale de la contrainte que nous voulons étudier sera définie comme :

$$q_{f\geq}(X) \equiv f(X) \geq seuil$$

4.1 Valeur d'un itemset X par un modèle f

La contrainte précédente nécessite de pouvoir calculer la valeur du motif par le modèle, i.e., $f(X)$. Par exemple, en considérant $f(x_1, x_2, x_3) = \sqrt{x_1} - cos(x_2)/2 \times log(x_3)$ donné en section 3.1, si le motif X est {"$x_1 \in [3, 5]$", "$x_2 = 3$", "$x_3 = A$"}, quelle est la valeur de $f(X)$? Autrement dit, quelle est la prévision du modèle f pour les valeurs $x_1 \in [3, 5]$, $x_2 = 3$, et $x_3 = A$?
Pour un itemset comme {"x_1=1", "x_2 = 3", "x_3=A", "mine"}, il suffit de calculer $f(num_{x_1}(\text{“}x_1 = 1\text{”}), num_{x_2}(\text{“}x_2 = 3\text{”}), num_{x_3}(\text{“}x_3 = A\text{”})) = f(1, 3, 10)$, en supposant que $num_{x_3}(\text{“}x_3 = A\text{”}) = 10$. Ce cas est simple car toutes les variables du modèle sont exprimées dans le motif. De plus, elles sont associées à une unique valeur dans chacun des items. A noter que l'item $mine$ n'a pas besoin d'être considéré pour le calcul de f car il n'est pas pris en compte par le modèle. Cet item est néanmoins intéressant car il apporte une information supplémentaire par rapport à la connaissance du modèle. Plus formellement, si $cover(X, D_f) = D_f$ et $\forall i \in X$, l'item i représente une valeur unique, alors $f($ X={$i_1, i_2, ..., i_n, ..., i_k$}) = $f(num_{x_1}(i_1), num_{x_2}(i_2), ..., num_{x_n}(i_n))$. Néanmoins, dans un cas plus général, cette opération soulève deux problèmes.
Premièrement, certaines variables du modèle peuvent ne pas être présentes dans les données ($D_f \not\subseteq D_{\mathcal{D}}$). De même, d'autres peuvent être absentes du motif car il n'implique qu'un sous-ensemble des valeurs apparaissant dans les données. Considérons l'itemset X'= {"x_1=1", "x_3=A", "mine"}. Il ne couvre pas toutes les variables du modèle (x_2 n'est pas exprimée). Il est tout de même possible de borner $f(X')$ en considérant les valeurs de x_2 pour lesquelles f est maximale/minimale. Dans notre exemple, si $x_2 = \pi$ *ou* 3π, alors f(1,x_2,10)=1.5. Cette valeur de f est la plus grande valeur possible étant donné les valeurs de x_2 et x_3. A l'opposé, si $x_2 = 0$, 2π ou 4π, alors f(1,x_2,10)=0.5. Cette valeur de f est la plus petite valeur possible étant donné les valeurs de x_2 et x_3. On en déduit que $0.5 \leq f(X') \leq 1.5$, même si x_2 n'est pas dans X'. Plus formellement, soit $X = \{i_1, i_2, ..., i_k\}$ un itemset, un modèle $f(x_1, ..., x_j, ...x_n)$, et $i'_l = num_{x_l}(i_l), \forall l \in [1, n], i_l \in X$. Pour $\forall x_j \in D_f, x_j \notin cover(X, D_f)$, on a

$$\min_{\forall i'_j \in dom(x_j)} (f(i'_1, ..., i'_j, ...i'_n)) \leq f(X)$$

$$f(X) \leq \max_{\forall i'_j \in dom(x_j)} (f(i'_1, ..., i'_j, ...i'_n))$$

Deuxièmement, les domaines de valeur du modèle et ceux des motifs sont différents. En effet, le modèle mathématique s'appuie sur des valeurs numériques alors que les motifs sont construits à partir de valeurs catégorielles (discrétisées sinon). Il est courant d'avoir des motifs représentant des mélanges d'intervalles, de valeurs numériques et de valeurs nominales. Considérons par exemple l'itemset $X'' = \{\text{“}x_1 = 4\text{”}, \text{“}x_2 \in [0, 2\pi[\text{”}, \text{“}x_3 = A\text{”}\}$. Il associe un intervalle de valeurs à la variable x_2. Cet item “$x_2 \in [0, 2\pi[$” est issu d'un prétraitement des données ayant discrétisé le domaine de valeur de x_2 en différents intervalles. Pour calculer $f(X'')$, il est possible d'appliquer une approche similaire à celle utilisée précédemment en bornant $f(X'')$ par rapport à $x_2 \in [0, 2\pi[$. En étudiant la fonction cosinus sur $[0, 2\pi[$, on sait que $f(X'')$ est maximale lorsque $x_2 = \pi$ (dans ce cas, $f(4, \pi, 10) = 2.5$) et minimale lorsque $x_2 = 0$ (dans ce cas, $f(4, 0, 10) = 1.5$). Nous pouvons donc en déduire que $1.5 \leq f(X'') \leq 2.5$. Cette formule peut donc être généralisée à tout item $i_j \in X$ représentant un intervalle $[inf_j, sup_j]$ d'une variable x_j du modèle f.

$$\min_{\forall i'_j \in [inf_j, sup_j]} (f(i'_1, ..., i'_j, ...i'_n)) \leq f(X)$$

$$f(X) \leq \max_{\forall i'_j \in [inf_j, sup_j]} (f(i'_1, ..., i'_j, ...i'_n))$$

On constate donc que la valeur d'un itemset X par f peut être un intervalle de valeurs. La définition de la contrainte de seuil sur les prévisions du modèle doit donc être étendue de la manière suivante :

Soit $f(X) = [inf_X, sup_X]$,
$q_{f\geq}(X) \equiv f(X) \geq seuil \equiv inf_X \geq seuil$

Il est important d'étudier certaines propriétés des prédicats et contraintes afin de les exploiter pour optimiser les calculs de motifs. C'est l'une des conditions nécessaires pour la définition d'algorithmes de fouille de données sous contraintes corrects et complets. Par exemple, la mesure de fréquence a la propriété d'être décroissante, i.e. si un itemset est non fréquent alors tous ses sur-ensembles sont aussi non fréquents. Cette propriété de fréquence minimale, dite d'"anti-monotonie" a été largement exploitée par de nombreux algorithmes d'extraction de motifs fréquents. Nous présentons maintenant des propriétés des modèles pouvant être utilisées dans les algorithmes d'extraction pour également contribuer à l'élagage dans les espaces de recherche.

4.2 Liens entre un itemset et ses sur-ensembles

Soit deux motifs $X, Y \in \mathcal{L}$ tel que $X \subset Y$. Si X et Y couvrent les mêmes variables de f, alors ils auront les mêmes items pour ces variables, et par conséquent $f(Y) = f(X)$. En fait, Y ne se différencie de X que par des items non pris en compte dans le modèle, et n'influencent pas le calcul de $f(Y)$. Dans ce cas, si $f(X) < seuil$, alors $f(Y) < seuil$. Par exemple, l'itemset $X'' = \{$"$x_1 = 4$", "$x_2 \in [0, 2\pi[$", "$x_3 = A$"$\}$ a la même valeur par f que $Y_1'' = \{$"$x_1 = 4$", "$x_2 \in [0, 2\pi[$", "$x_3 = A$", "$mine$"$\}$ et $Y_2'' = \{$"$x_1 = 4$", "$x_2 \in [0, 2\pi[$", "$x_3 = A$", "$mine$", "$piste$"$\}$. En effet, $f(4, 0, 10) \leq f(X'') \leq f(4, \pi, 10)$, tout comme $f(Y_1'')$ et $f(Y_2'')$, car les variables x_1, x_2 et x_3 de f ont les mêmes valeurs. Par conséquent, si $f(X'') < seuil$, alors $f(Y_1'')$ et $f(Y_2'')$ aussi.

Propriété 4.1. *Soit $X \in \mathcal{L}$. Si $q_{f\geq}(X)$ est faux alors $\forall Y \in \mathcal{L}$, $X \subset Y$, $cover(X, D_F) = cover(Y, D_F)$, $q_{f\geq}(Y)$ est faux.*

Le cas est plus complexe lorsque des variables de f ne sont pas exprimées dans X mais le sont dans Y (seule autre possibilité si l'on conserve l'hypothèse $X \subset Y$). Prenons l'exemple du motif $X = \{$"$x_2 \in [0, 2\pi[$", "$x_3 = A$"$\}$ et du sur-ensemble $Y_1 = \{$"$x_1 = 16$", "$x_2 \in [0, 2\pi[$", "$x_3 = A$"$\}$. La variable x_1 est exprimée dans Y_1 mais pas dans X. Or $0.5 = f(1, 0, 10) \leq f(X) \leq f(16, \pi, 10) = 4.5$, car $dom_{num}(x_1) = [1, 16]$. Si le seuil minimum pour f est 2, $f(X) < seuil$, pourtant $f(Y_1) = [3.5, 4.5] > seuil$. Par contre, si le seuil minimum est 5, on est sûr que tous les sur-ensembles de X vérifient $f(X) < 5$. En effet, tous ces sur-ensembles ont leur valeur comprise entre 0.5 et 4.5 (car ce sont les valeurs maximales et minimales de f pour tout $x_1 \in [1, 16]$). Or, étant donné que la borne supérieure de $f(X)$ est inférieure au seuil, il en sera de même pour tous les sur-ensembles de X.

Propriété 4.2. *Soit $X \in \mathcal{L}$ et $inf_X \leq f(X) \leq sup_X$. Si $q_{f\geq}(X)$ est faux et $sup_X < seuil$ alors $\forall Y \in \mathcal{L}$, $X \subset Y$, $q_{f\geq}(Y)$ est faux.*

4.3 Itemsets partageant la même couverture

Les propriétés précédentes mettent en avant des liens entre un itemset et ses sur-ensembles. Ces liens permettent de borner la valeur par f des sur-ensembles d'un itemset. L'étude détaillée de f permet d'exposer d'autres propriétés entre les motifs. Toutefois, elle peut être complexe de part la nature des fonctions considérées (des fonctions à plusieurs variables non nécessairement linéaires). Il est difficile d'étudier globalement la monotonie d'une fonction à plusieurs variables. Notre solution consiste à analyser la fonction par rapport à une variable à la fois (les autres étant considérées comme des constantes). Cela revient à étudier ses dérivées partielles pour identifier les intervalles dans lesquels la fonction est monotone par rapport à chaque variable. Dans chacun de ces intervalles (pour chacune des variables), il est possible de dériver des propriétés permettant d'élaguer l'espace de recherche.

Prenons par exemple les itemsets X={"x_1=4", "$x_2 \in [\pi/2, \pi[$", "x_3=A"} et Y={ "x_1=4", "$\boldsymbol{x_2 \in [0, \pi/2[}$", "$x_3$ =A"}. On remarque que $cover(X, D_f) = cover(Y, D_f)$. Etant donné que f est strictement croissante par rapport à x_2 sur $[0, \pi]$ (i.e., $\frac{\partial f}{\partial x_2} > 0$ sur $[0, \pi]$) et que X est plus grand que Y par rapport à x_2 (i.e., $Y \prec_{x_2} X$), alors $f(Y)$ <$f(X)$. En effet, f(X) = [2, 2.5[et f(Y) = [1.5, 2[. Par conséquent, si f(X) < seuil, alors $f(Y)$ aussi (même si $X \not\subset Y$).

De même, si l'on considère le motif $Y'' = \{$"$\boldsymbol{x_1 = 1}$", "$\boldsymbol{x_2 \in [0, \pi/2[}$", "$x_3 = A$"$\}$, on peut dire que $f(Y'') < f(X)$, car $\frac{\partial f}{\partial x_1} > 0$ sur $dom(x_1)$ et $Y'' \prec_{x_1} X$. Cette propriété peut être formalisée ainsi :

Propriété 4.3. *Soit $X, Y \in \mathcal{L}$ deux itemsets tels que $cover(X, D_f) = cover(Y, D_f)$. $\forall x_j \in cover(X, D_f) \cap cover(Y, D_f)$, tel que $((\frac{\partial f}{\partial x_j} > 0 \wedge Y \prec_{x_j} X) \vee (\frac{\partial f}{\partial x_j} < 0 \wedge X \prec_{x_j} Y))$. Si $q_{f\geq}(X)$ est faux, alors $q_{f\geq}(Y)$ est faux.*

Notons que l'impact de cette propriété dépend beaucoup de la discrétisation. En effet, il n'aurait pas été possible de déduire cela si nous avions eu les itemsets $X = \{$"$x_1 = 4$", "$\boldsymbol{x_2 \in [\pi, 2\pi[}$", "$x_3 = A$"$\}$ et $Y = \{$"$x_1 = 4$", "$\boldsymbol{x_2 \in [0, \pi[}$", "$x_3 = A$"$\}$, car f est croissante par rapport à x_2 sur $[0, \pi]$ et décroissante sur $[\pi, 2\pi]$. Soit une variable x_j dont le domaine est découpé en intervalles dans lesquels f est monotone. Plus le nombre d'items appartenant à chacun de ces intervalles est important, plus cette propriété sera efficace.

4.4 Itemsets sous-contraintes du modèle

Il est relativement simple d'intégrer dans les algorithmes d'extraction de motifs des contraintes exhibant des propriétés similaires à celles traditionnellement utilisées pour extraire des itemsets (e.g, la contrainte de fréquence). Les modifications engendrées sont minimes car le test n'implique pas d'accès aux données ou à toute autre ressource. Elles impactent simplement la génération des motifs candidats. L'intérêt d'appliquer cette contrainte de seuil du modèle pendant l'extraction permet de n'examiner que certains motifs et par conséquent d'améliorer la performance de l'algorithme et la pertinence des solutions.

5 Expérimentations

Dans nos expérimentations, nous avons choisi d'intégrer les contraintes dans l'algorithme Close-By-One [10]. Nous avons utilisé un jeu de données réelles extraites d'une image satellitaire de plus de 8 millions de pixels (une image SPOT de 500 Mo) représentant le Sud de la Nouvelle-Calédonie. L'image a été transformée en base transactionnelle représentant les informations sur les pixels. Les attributs sont les propriétés radiométriques des pixels (rouge, verte, bleue et NDVI) discrétisées et les variables nécessaires au modèle Atherton [4] que l'on a utilisé pour ces

expérimentations (pente, nature et occupation du sol). Au final, nous obtenons une base de données transactionnelles composée de 8 millions de transactions, où chaque transaction est constituée de 7 items. Ce jeu de données contient au total 74 items différents (regroupés selon les 7 attributs/variables cités précédemment). Nous avons exécuté l'algorithme sur une machine de 8Go de RAM et un processeur cadencé à 3.20GHz.

La Figure 2 donne les performances en temps ainsi que le nombre de solutions trouvées pour différents seuils de fréquence en abscisses, et pour différentes contraintes de modèle. Nous avons répété l'expérience sur une partie de l'image originale (coin Nord Ouest), ne faisant qu'un million de pixels. Conformément à nos attentes, l'utilisation des contraintes du modèle réduit grandement (par plusieurs ordres de grandeur) le nombre de solutions, et accélère ainsi la fouille. Par exemple, sans notre contrainte basée sur le modèle Atherton, le nombre de solutions peut excéder 1000 itemsets pour une fréquence minimale de 10%. Avec la contrainte basée sur ce modèle, le nombre de solutions ne dépasse jamais 10. De la même manière, le temps d'exécution peut atteindre 6000 secondes sans notre contrainte, alors qu'ils ne dépasse pas 2000 secondes avec la contrainte $f \geq 15$ (en conservant la même contrainte de fréquence). Grâce à la réduction du nombre de solutions, nous avons pu plus facilement recueillir le seul itemset présentant une érosion particulièrement forte ($f \geq 15$) : {Geology=Serpentinites, LandCover=Sol nu sur substrat volcano-sédimentaire, Slope=[61,100], Red=(14.2,28.4], Green=[0.0,36.1], NDVI=(-0.071,0.115], Blue=[0.0,24.5]}. Les variables radiométriques (qui ne font pas partie du modèle) précisent que nous sommes en présence de faibles indices de Vert et de NDVI, traduisant une absence de végétation, confirmant à son tour la pertinence du modèle.

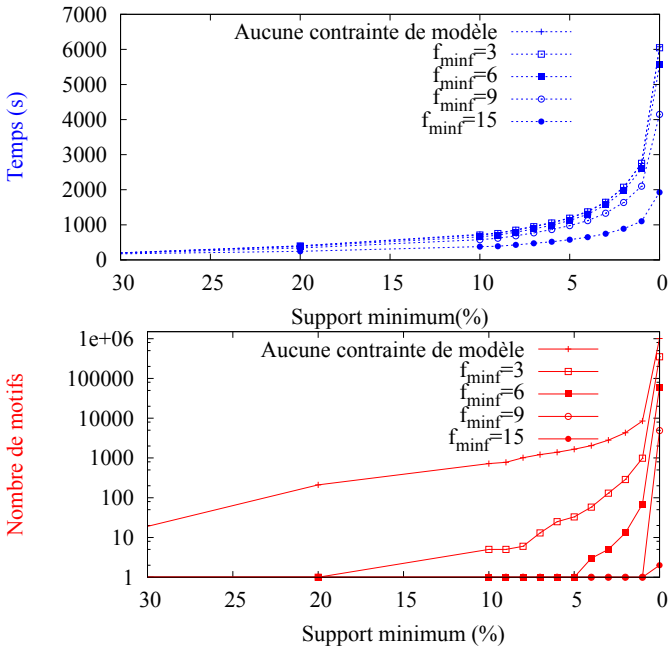

FIGURE 2 – Temps d'exécution et nombre d'itemsets

6 Conclusion et perspectives

Cet article a mis en avant l'intérêt d'exploiter les modèles des experts dans le processus d'extraction des connaissances. En effet, ils représentent un condensé de la connaissance du domaine et sont donc bien plus riches que de simples règles *"si ... alors ..."*. Dans notre étude, nous avons utilisé ces modèles comme contraintes lors de l'extraction de motifs. Ces contraintes permettent de mieux cibler l'analyse, tout en améliorant les performances grâce à des propriétés des modèles. Nous obtenons ainsi des motifs plus pertinents, venant compléter ou contredire les connaissances des phénomènes étudiés.

Une perspective naturelle est de dériver des contraintes combinant plusieurs modèles, chacun étant pondéré par l'expert en fonction du contexte d'application. Une autre perspective serait de comparer plus globalement la connaissance induite par un ou plusieurs modèles avec la connaissance extraite par fouille de données. Pour finir, il serait aussi intéressant de combiner, faire une synthèse, de la connaissance de plusieurs modèles. Les modèles des experts deviendraient ainsi les données sur lesquelles la fouille de données serait appliquée. Ce type d'approche pourrait, par exemple, permettre d'extraire des corrélations fréquemment exprimées par les modèles d'un domaine donné.

Références

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487–499, 1994.

[2] S. S. Anand, D. A. Bell, and J. G. Hughes. The role of domain knowledge in data mining. In *CIKM*, pages 37–43. ACM Press, 1995.

[3] C. Antunes. Mining patterns in the presence of domain knowledge. In *ICEIS (2)*, pages 188–193, 2009.

[4] J. Atherton, D. Olson, L. Farley, and I. Qauqau. Fiji Watersheds at Risk : Watershed Assessment for Healthy Reefs and Fisheries. Technical report, Wildlife Conservation Society - South Pacific, Suva, Fiji, 2005.

[5] L. Cao. Domain-driven data mining : Challenges and prospects. *IEEE Transactions on Knowledge and Data Engineering*, 22(6) :755–769, 2010.

[6] L. de Castro Medeiros, C. Castilho, C. Braga, W. de Souza, L. Regis, and A. Monteiro. Modeling the dynamic transmission of dengue fever : investigating disease persistence. *PLoS neglected tropical diseases*, 5(1), Jan. 2011.

[7] P. Domingos. Toward knowledge-rich data mining. *Data Mining and Knowledge Discovery*, 15(1) :21–28, Apr. 2007.

[8] S. Jaroszewicz, T. Scheffer, and D. A. Simovici. Scalable pattern mining with bayesian networks as background knowledge. *Data Mining and Knowledge Discovery*, 18(1) :56–100, 2009.

[9] S. Jaroszewicz and D. A. Simovici. Interestingness of frequent itemsets using bayesian networks as background knowledge. In *SIGKDD*, pages 178–186, 2004.

[10] S. O. Kuznetsov and S. Obiedkov. Comparing performance of algorithms for generating concept lattices. *Journal of Experimental and Theoretical Artificial Intelligence*, 14 :189–216, 2002.

[11] L. Lane and M. Nearing. *Water Erosion Prediction Project : Hillslope Profile Model Documentation*. USDA.ARS.NSERL Report. US Department of Agriculture Science and Education Administration, Washington, USA, 1989.

[12] H. Mannila and H. Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3) :241–258, 1997.

[13] R. Morgan. A simple approach to soil loss prediction : a revised Morgan-Morgan-Finney model. *Catena*, 44(4) :305–322, July 2001.

[14] R. T. Ng, L. V. S. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning optimizations of constrained associations rules. *ACM SIGMOD Record*, 27(2) :13–24, June 1998.

[15] B. Padmanabhan and A. Tuzhilin. A belief-driven method for discovering unexpected patterns. In *KDD*, pages 94–100, 1998.

[16] J. Pei, J. Han, and L. V. S. Lakshmanan. Mining frequent item sets with convertible constraints. In *ICDE*, pages 433–442, 2001.

[17] L. D. Raedt and A. Zimmermann. Constraint-based pattern set mining. In *SDM*, pages 237–248, 2007.

[18] W. H. Wischmeier and D. D. Smith. *Predicting Rainfall Erosion Losses : A Guide to Conservation Planning*, volume 537 of *Agricultural Handbook*. US Department of Agriculture Science and Education Administration, Washington, USA, 1978.