ECAI 2014
T. Schaub et al. (Eds.)

doi:10.3233/978-1-61499-419-0-327

Improving pattern discovery relevancy by deriving constraints from expert models [1]

Frédéric Flouvat [2] and **Jérémy Sanhes**[2] and **Claude Pasquier**[2, 3] and **Nazha Selmaoui-Folcher**[2] and **Jean-François Boulicaut** [4]

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. Integration of domain knowledge in constraint-based data mining is limited. Relevant patterns still miss because methods partly fail in assessing their subjective interestingness. However, in practice, we often have in the literature mathematical models defined by experts based on their domain knowledge. We propose here to exploit such models to derive constraints that can be used during the data mining phase to improve both pattern relevancy and computational efficiency. Even though the approach is generic, it is illustrated on pattern set discovery from real data for studying soil erosion.

1 Introduction

Experts of a broad range of scientific fields (e.g., geologists, physicists or epidemiologists) often express their knowledge in the form of models. For example, soil erosion experts developed mathematical models (functions of several variables) to assess an erosion risk according to a set of environmental parameters (e.g., vegetation, geology, rainfalls, slope) [16, 19, 4]. Similarly, epidemiologists developed models to estimate the number of people infected by Dengue fever, using the number of inhabitants, life cycle of mosquitoes, and seasons [5, 8, 10]. Such models capture an important expert knowledge in a given context. However, the variables used in such typical models often represent a small part of the variables for which values are nowadays easily collected (thanks to teledetection, remote sensing, etc). Using simultaneously both expert models and available data appears as a timely challenge.

When considering knowledge discovery from data (KDD) based on pattern discovery, the popular framework of constraint-based mining is often used. Various types of patterns have been studied, such as itemsets, sequential patterns, episodes or sub-graphs. In such contexts, constraints can be used to specify a priori the objective and the subjective interestingness of patterns [18]. Objective measures are generally based on frequency and/or statistical properties of patterns (e.g., giving rise to the popular minimal frequency constraint [1]), whereas subjective interestingness has to be specified thanks to expert's goals or needs (see, e.g., [20, 23, 24]). Specifying subjective interestingness by means of constraints concerns the declarative definition of needed properties like, for instance, unexpectedness w.r.t. domain knowledge or available models [21, 13]. Let us recall that constraint-based mining enables to improve the relevancy of computed patterns, but also to use theoretical properties of constraints (e.g., monotonicity property) to perform complete though computationally efficient extractions (see, e.g., [6]).

Integrating expert knowledge in the KDD process is not new [2, 11, 3, 9]. This knowledge is usually expressed as rules/constraints (e.g., "if... then...") that are manually defined by experts (e.g., "if $trail$ and $rains$ then $erosion$"). Such information is hard to obtain and it generally covers a small part of the available domain knowledge: in practice, it is often limited to few basic rules. The use of taxonomies or ontologies improve the definition of useful constraints and the interpretation of extracted patterns [7, 22] even though the bottleneck can then be on their acquisition. We can also use graphical models like bayesian networks. For example, [12] formalizes expert's background knowledge as a bayesian network of causal relations and dependencies between attributes. This network can evolve during the KDD process. It is possible to exploit such a model during a pattern mining phase to extract more interesting patterns. The considered interestingness is defined here as "the divergence between the expected frequency of patterns predicted by the model w.r.t. the observed frequency in the data". Authors in [14] use the Maximum Entropy principle to model statistical informations (mean, variance and histograms) on arbitrary sets of cells as background knowledge. Then, they propose a measure for contrasting the subjective informativeness of patterns.

In many scientific domains, experts have formalized a part of their domain knowledge by means of models (e.g., the so-called physical models). The originality of our contribution is to take advantage of existing models in the literature to build new constraints that can be integrated into a pattern discovery process. Pattern relevancy can be improved, the computation can be more efficient, and experts are not solicited to specify part of the useful domain knowledge for each extraction. We focus on itemset patterns and we consider the many models that can be expressed as a function of several attributes/variables. In Section 2, we introduce several examples of linear, polynomial, and even nonlinear, models that can be used to improve pattern relevancy. We highlight theoretical properties of these model w.r.t. the itemset pattern domain that can be used to safely prune the search space and thus to improve computational efficiency. Our motivating example within the paper concerns expert models to study soil erosion. In this context, our work focuses on patterns likely

[1] This research is partially supported by projet FOSTER ANR-2010-COSI-012-01, and by project AMADOUER (MASTODONS Challenges) funded by CNRS.
[2] PPME, Université de la Nouvelle Calédonie, BP R4, F-98851 Nouméa, Nouvelle-Calédonie, email: FirstName.LastName@univ-nc.nc
[3] Institut de Biologie Valrose, CNRS UMR7277 - INSERM U1091, F-06108 Nice, France
[4] Université de Lyon, CNRS, INSA-Lyon, LIRIS UMR5205, F-69621, Lyon, France, email: jean-francois.boulicaut@insa-lyon.fr

to be related to a strong erosion, while considering the influence of other environmental parameters that are not considered by the models (e.g., parameters related to human activities). It turns out that the constraints based on such models enable to assess or enrich expert's knowledge. It can also highlight contradictory relations.

Section 2 presents the soil erosion case study. Section 3 introduces our theoretical framework. Section 4 presents our contribution to push constraints based on expert models into a pattern mining phase. Section 5 shows experimental results on real data. Finally, section 6 concludes and refers to a few perspectives.

2 A case study on soil erosion

Soil erosion has a deep impact on mankind all over the world. It is worsened by anthropic activities (e.g., deforestation, industries, agriculture), affects environment and economy. Faced with this problem, the scientific community (mainly geologists and geographers) has developed mathematical models to estimate soil erosion risk. Two model classes can be distinguished : empirical and physical models.

Empirical models are constructed empirically based on expert knowledge and experiments. The USLE (*Universal Soil Loss Equation*) model [25] and the model proposed in [4] are typical examples (the first one is a polynomial model and the second one is linear). Physical models are quantitative models based on physical properties that are calibrated from experimental observations. For example, WEPP (*Water Erosion Prediction Project*, [16]) and RMMF (*Revised Morgan-Morgan Finney*, [19]) are both based on several physical models (nonlinear and non-polynomial). RMMF divides the erosion process in two steps: raindrop detachment (see figure 1) and runoff detachment. Each step is based on a physical sub-model. Their respective results are summed to finally assess the annual soil loss.

Parameters	domain of values
Soil detachment index (in g/J) x_K	depends on soil type
Annual rainfall (in mm) x_R	[0 , 12 000]
Proportion of rain stopped by vegetation x_A	[0 , 1]
Canopy cover percentage x_{CC}	[0 ,1]
Rainfall intensity (in mm/h) x_I	{10, 25, 30} depending on the climate of the studied area
Vegetation height (in 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. Raindrop detachment model in RMMF

3 Preliminary definitions

3.1 Expert models

Let $D_{model} = \{x_1, x_2, ..., x_n\}$ be the set of **variables/attributes** of the expert model. We denote by $dom(x_j)$, **domain** of x_j, the set of possible values of attribute x_j. A **mathematical model** is a function $f : dom(x_1) \times dom(x_2) \times ... \times dom(x_n) \rightarrow \mathbb{R}$, $x = (x_1, x_2, \ldots, x_n) \mapsto f(x)$. It represents the knowledge of one or several experts about a phenomenon, and it can be found in the literature of the domain.

These models are numerical functions. Thus, when experts want to integrate categorical/nominal data in such numerical models (e.g. soil type), they have to transform these values to numbers. This mapping is done by experts based on their domain knowledge. It is given with the expert model. For example, the RMMF model [19] takes in input the soil type which is a nominal value (e.g. "sand", "loam"). In their paper, the experts associate each soil type to a numerical value (called soil detachment index) based on past experiments (e.g. "sand" is 1.2 and "loam" is 0.8). In the same way, the model proposed in [4] integrates land cover data (e.g. "dense forest", "sugar cane farming"). The numerical value associated to each land cover is given in the contribution. It is proportional to the impact of the land cover on soil erosion (e.g. "dense forest" is 1 and "sugar cane farming" is 4).

In the rest of the paper, we will use the following example :
$f : [1, 16] \times [0, 4\pi] \times [1, 100] \subset \mathbb{R}^3 \rightarrow [0, 5] \subset \mathbb{R}$
$f(x_1, x_2, x_3) = \sqrt{x_1} - \cos(x_2)/2 \times \log(x_3)$

3.2 Itemset mining

We link now the itemset definition of Agrawal and Srikant [1] with the previous models.

Let $D_{DB} = \{d_1, d_2, ..., d_{n'}\}$ be the **dimensions/attributes** of the database DB (e.g., geology, rainfall or vegetation). D_{DB} must cover at least one attribute of the expert model, i.e., $D_{DB} \cap D_{model} \neq \emptyset$. The domain of $d_{j'}$ in DB, written $dom(d_{j'})$, is a set of categorical values. Domains of numerical attributes are discretized (i.e., decomposed into disjoint intervals). For example, the domain of the "annual rainfall" attribute x_R in Figure 1 is $dom(x_R) = [0, 12000]$. In the database, it can be discretized as follows $\{$"$x_R \in [0, 2000]$", "$x_R \in [2001, 3200]$", "$x_R \in [3201, 12000]$"$\}$. In the remainder of this paper, we consider that each original value is transformed into a categorical value in which both the attribute and the value are represented (e.g., value "ultramafic soil" of attribute x_K will be noted "$x_K = ultramafic\ soil$"). Formally, values of $dom(d_{j'})$ can be seen as pairs (attribute, value). To facilitate the interpretation of itemsets in examples, we only integrate the name of attributes in categorical values if they are related to the expert model.

Let $\mathcal{I} = \bigcup_{d_j \in D_{DB}} dom(d_j)$ be the set of all values in the database DB. A value $i \in \mathcal{I}$ is called an **item**. The pattern language is $\mathcal{L} = \{X \in 2^{\mathcal{I}} \mid \nexists i, i' \in X \ s.t. \ i, i' \in dom(d_j), d_j \in D_{DB}\}$. In other words, a pattern is a combination of categorical values from different attributes. A set of items $X = \{i_1, i_2, ..., i_k\} \in \mathcal{L}$, with $k \leq n'$ (the number of attributes), is an **itemset**. The **attributes of an itemset** X **over** $\mathbf{D_{model}}$, written $\mathbf{Atts(X, D_{model})}$, is the set of attributes of X's items that belong to D_{model}. For example, $X = \{$"$x_K = ultramafic\ soil$", "$x_R = [2001, 3200]$", "$mine$", "$trail$"$\}$ is an itemset. Its attributes w.r.t. attributes of the previous $RMMF$ model are $Atts(X, RMMF) = \{x_K, x_R\}$.

The problem of constraint-based pattern mining is to find the set of patterns satisfying a selection predicate q in the data (where q can be a conjunction of primitive constraints). This set, sometimes called the theory of DB with respect to $\mathcal{L}$ and q, is denoted by $\mathcal{T}h(\mathcal{L}, DB, q)$ [17]. Formally, $\mathcal{T}h(\mathcal{L}, DB, q) = \{X \in \mathcal{L} \mid q(X, DB)$ is true$\}$. In our work, the predicate q is a conjunction of constraints based on objective and subjective measures. The first constraint used in this paper is the minimal frequency constraint, i.e., a pattern is selected iff it occurs in the database more than a user-defined threshold (often denoted by $minsup$). The second constraint is a threshold constraint w.r.t. an expert model, denoted by $q_{f\geq}$, i.e., a pattern is selected iff its value w.r.t. an expert model f is greater than a user-defined threshold (denoted by $minf$).

4 From patterns to models

A simple approach to take advantage of domain knowledge is to derive from it primitive constraints, and to use them during pattern mining. We propose to use constraints that are derived from expert models instead of ad-hoc constraints manually defined by experts. These constraints represent much more than basic "if ... then ..." rules since a single expert model can denote a huge number of such rules.

Various types of constraints could be derived according to data, to expert models, but also to the studied problem. In this paper, we focus on a constraint relatively that is to the minimal frequency constraint even if its theoretical properties are quite different.

We can define a constraint that filters patterns whose value by f is greater or equal than a given threshold. In other words, this constraint keeps patterns for which values predicted by the expert model are greater or equal than a given threshold. Depending on what f expresses, this constraint will have different meanings. For example, if f estimates soil loss (in $kg.m^{-2}$ by year) such as in RMMF model, this constraint will filter patterns corresponding to a potential soil loss greater than a given quantity. In the absence of "ground truth" (i.e., data on real soil losses), this constraint will enable to highlight if such losses are likely to be frequent in the studied area and in which situations (thanks to the values of the other environmental parameters described in the pattern). Moreover, it would enable to show (if the pattern is frequent) with which other factors, not covered by the model, these soil losses are frequently related in the data. In the presence of "ground truth", this constraint will enable to compare predictions of the expert model with the "truth" of collected data. Patterns in accordance with the model are interesting because they are validated twice: once by ground truth and once by domain knowledge (i.e., represented by the expert model). Moreover, additional items of the extracted pattern can complement explanations of the expert model. Patterns in contradiction with the prediction of the expert model are also interesting, because they enable to identify correlations not considered by the expert model used (which can be used to improve the model).

Let $X \in \mathcal{L}$ be an itemset, f be a model already developed by experts, and $minf \in \mathbb{R}$ a user-defined threshold. Our threshold constraint derived from expert model f is defined as:

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

4.1 Value of an itemset X by an expert model f

The previous constraint implies that we can calculate the value predicted by an expert model for a given itemset, i.e., $f(X)$. Let us consider the model $f(x_1, x_2, x_3) = \sqrt{x_1} - \cos(x_2)/2 \times \log(x_3)$ introduced in Section 3.1. If pattern X is {"$x_1 \in [3, 5]$", "$x_2 = 3$", "$x_3 = A$"}, what is the value of $f(X)$? In other words, what is the prediction of the expert model f for $x_1 \in [3, 5]$, $x_2 = 3$, and $x_3 = A$ values ?

For an itemset such as {"x_1=1", "x_2 = 3", "x_3=A", "mine"}, we only have to calculate $f(1, 3, 10)$, if we suppose that "$x_3 = A$" is associated to 10 by experts. This case is simple because all the attributes of the model appear in the pattern. Moreover, these attributes are associated to a single value in each items (and not an interval of values). Note that we do not need to consider the $mine$ item in f calculus, since this information is not considered by the model. Nevertheless, this item is interesting because it gives an additional information w.r.t. the knowledge captured by the model. More formally, if $Atts(X, D_{model}) = D_{model}$ and $\forall i \in X$, item i represents a single value, then $f($ X=$\{i_1, i_2, ..., i_n, ..., i_k\}) = f(i_1, i_2, ..., i_n)$. However, we have to deal with two problems in a more general case.

First, some attributes of the model may not be available in the data ($D_{model} \not\subseteq D_{DB}$). In the same way, some attributes of the model may not be expressed in the pattern. Let us consider the itemset X'= {"x_1=1", "x_3=A", "mine"}. It does not have all the attributes of the model (x_2 is not expressed). We can find upper and lower bounds for $f(X')$ by considering the values of x_2 for which f is maximal/minimal. In our example, if $x_2 = \pi$ or 3π, then $f(1, x_2, 10) = 1.5$. This value of f is the greatest possible value given x_1 and x_3 values. On the other hand, if x_2 = 0, 2π or 4π, then $f(1, x_2, 10) = 0.5$. This value of f is the smallest possible value given x_1 and x_3 values. We can easily deduce that $0.5 \leq f(X') \leq 1.5$ even if x_2 is not represented in X'. More formally, let $X = \{i_1, i_2, ..., i_n, ..., i_k\}$ be an itemset, $f(x_1, ..., x_j, ...x_n)$ be an expert model. For $\forall x_j \in D_{model}, x_j \notin Atts(X, D_{model})$:

$$\min_{\forall i_j \in dom(x_j)} f(i_1, ..., i_j, ...i_n) \leq f(X)$$

$$f(X) \leq \max_{\forall i_j \in dom(x_j)} f(i_1, ..., i_j, ...i_n)$$

Next, domains of values of the model and those of itemsets may be different. The mathematical model is based on numerical values, whereas itemsets are based on categorical values (using a discretization method if necessary). We often have itemsets representing a mix of intervals, numerical values and categorical values. For example, let us consider the itemset $X'' = \{$"$x_1 = 4$", "$x_2 \in [0, 2\pi[$", "$x_3 = A$"$\}$. It associates an interval of values to attribute x_2. This item "$x_2 \in [0, 2\pi[$" comes from a data preprocessing step in which the domain of x_2 has been discretized in several (disjoint) intervals. Such as previously, it is possible to find $f(X'')$ by studying upper and lower bounds w.r.t. $x_2 \in [0, 2\pi[$. If we study the cosine function on $[0, 2\pi[$, then we know that $f(X'')$ is maximal when $x_2 = \pi$ (in this case, $f(4, \pi, 10) = 2.5$), and minimal when $x_2 = 0$ (in this case, $f(4, 0, 10) = 1.5$). Thus, we can deduce that $1.5 \leq f(X'') \leq 2.5$. The previous formula can be generalized to any item $i_j \in X$ that represents an interval $[inf_j, sup_j]$ of an attribute x_j of model f :

$$\min_{\forall i_j \in [inf_j, sup_j]} f(i_1, ..., i_j, ...i_n) \leq f(X)$$

$$f(X) \leq \max_{\forall i_j \in [inf_j, sup_j]} f(i_1, ..., i_j, ...i_n)$$

As a consequence, the value of an itemset X by an expert model f can be an interval of values. The definition of our model-based threshold constraint has to be extended in the following way:

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

It is now important to study the theoretical properties of this constraint to improve computational efficiency. It is one of the necessary condition to design complete and correct patten mining algorithms. For example, the frequency is a monotonic decreasing function. Thus, if an itemset is not frequent, then all its supersets are also not frequent. This "anti-monotonic" property has been extensively used in frequent pattern mining algorithms to provide scalability. We now discuss properties of models that can also be used to prune the search space well.

4.2 Theoretical properties of models w.r.t. itemsets

4.2.1 Properties between an itemset and its supersets

Let $X, Y \in \mathcal{L}$ be two itemsets such that $X \subset Y$. If X and Y have the same attributes of the model f, then they have the same items

for these attributes, and thus $f(Y) = f(X)$. In other words, Y only differs from X because it has additional items not considered by the model, which does not impact $f(Y)$ calculus. In this case, if $f(X) < minf$, then $f(Y) < minf$. For example, itemset $X'' = \{$"$x_1 = 4$", "$x_2 \in [0, 2\pi[$", "$x_3 = A$"$\}$ has the same value by f as $Y_1'' = \{$"$x_1 = 4$", "$x_2 \in [0, 2\pi[$", "$x_3 = A$", "$mine$"$\}$ and $Y_2'' = \{$"$x_1 = 4$", "$x_2 \in [0, 2\pi[$", "$x_3 = A$", "$mine$", "$trail$"$\}$. Indeed, $f(4, 0, 10) \leq f(X'') \leq f(4, \pi, 10)$ such as $f(Y_1'')$ and $f(Y_2'')$, since attributes x_1, x_2 and x_3 of f have the same values. As a consequence, if $f(X'') < minf$, then $f(Y_1'')$ and $f(Y_2'')$ are also lower than $minf$.

Property 1. *Given $X \in \mathcal{L}$. If $q_{f\geq}(X)$ is false then $\forall Y \in \mathcal{L}$ s.t. $X \subset Y$ and $Atts(X, D_{model}) = Atts(Y, D_{model})$, $q_{f\geq}(Y)$ is false.*

It is more complex when f attributes are not expressed in X but are expressed in Y (the only other possibility if we consider $X \subset Y$ hypothesis). For example, let us consider itemset $X = \{$"$x_2 \in [0, 2\pi[$", "$x_3 = A$"$\}$ and one of its supersets $Y_1 = \{$"$x_1 = 16$", "$x_2 \in [0, 2\pi[$", "$x_3 = A$"$\}$. Attribute x_1 is expressed in Y_1 but not in X. We know that $0.5 = f(1, 0, 10) \leq f(X) \leq f(16, \pi, 10) = 4.5$ since $dom(x_1) = [1, 16]$. If the minimum threshold for f is 2, we may have $f(X) < minf$, although $f(Y_1) = [3.5, 4.5] > minf$. On the other hand, if $minf = 5$, then we are sure that all the supersets of X satisfy $f(X) < 5$. For any $x_1 \in [1, 16]$, minimum and maximum values of f are 0.5 and 4.5. Thus, all supersets of X have a value by f between 0.5 and 4.5. If the upper bound of $f(X)$ is lower than the threshold, then all the supersets of X are also in such a case. This property is known as "bounds consistency" of a constraint in the constraints community.

Property 2. *Let be $X \in \mathcal{L}$ such that $inf_X \leq f(X) \leq sup_X$. If $q_{f\geq}(X)$ is false and $sup_X < minf$, then $\forall Y \in \mathcal{L}$ s.t. $X \subset Y$, $q_{f\geq}(Y)$ is false.*

4.2.2 Property of itemsets sharing the same attributes

The previous properties show links between an itemset and its supersets. These links enable to bound the values of f for the supersets of a given itemset. Analyzing function f enables to highlight other properties between itemsets w.r.t. the model. However, this can be complex due to the nature of studied functions (possibly nonlinear functions of several attributes). It is difficult to study globally the monotonicity of a function over several attributes. Our solution consists in analyzing the curve of the function w.r.t. one attribute at the same time (the others being considered as constants). This solution is equivalent to studying the partial derivative of f on each attribute. Given one attribute, the main objective is to identify the intervals in which the function is monotonic. Then, for each interval, it is possible to derive properties that enable to prune the search space.

Let us consider itemsets X={"x_1=4", "$x_2 \in [\pi/2, \pi[$", "x_3=A"} and Y={ "x_1=4", "$\boldsymbol{x_2 \in [0, \pi/2[}$", "$x_3$ =A"}. Note that $Atts(X, D_{model}) = Atts(Y, D_{model})$. The analysis of function f w.r.t. x_2 shows that it is strictly increasing on $[0, \pi]$ (i.e., $\frac{\partial f}{\partial x_2} > 0$ on $[0, \pi]$). Since X is greater than Y w.r.t. x_2 (i.e., $Y \prec_{x_2} X$), we have $f(Y) <f(X)$. Indeed, $f(X) = [2, 2.5[$ and $f(Y) = [1.5, 2[$. As a consequence, if $f(X) < minf$, then $f(Y) < minf$ (even if $X \not\subset Y$).

In the same way, let us consider $Y'' = \{$"$\boldsymbol{x_1 = 1}$", "$\boldsymbol{x_2 \in [0, \pi/2[}$", "$x_3 = A$"$\}$. We know that $f(Y'') < f(X)$ because $\frac{\partial f}{\partial x_1} > 0$ on $dom(x_1)$ and $Y'' \prec_{x_1} X$.

More formally, a **total order relation**, denoted by $\prec_{x_j}$, can be defined for each attribute $x_j \in D_{DB} \cap D_{model}$. If $i, i' \in dom(x_j)$ represent intervals, i.e., $i =$ "$x_j \in [a, b]$" and $i' =$ "$x_j \in [c, d]$", then $i \prec_{x_j} i'$ iff $b < c$. For example, "$x_R \in [0, 2000]$" $\prec_{x_R}$ "$x_R \in [2001, 3200]$" because $2000 < 2001$. If $i, i' \in dom(x_j)$ are nominal values associated to numerical values num_i and $num_{i'}$ by experts, then $i \prec_{x_j} i'$ iff $num_i < num_{i'}$. For example, "$x_K = volcanic\ soil$" $\prec_{x_j}$ "$x_K = ultramafic\ soil$" because experts associate "volcanic soil" to 8 and "ultramafic soil" to 10. It is possible to extend the order $\prec_{x_j}$ to itemsets $X, Y \in \mathcal{L}$. Thus, we have a partial order relation between itemsets w.r.t an attribute x_j. The itemset X is less than Y according to x_j (written $X \prec_{x_j} Y$) iff $i \prec_{x_j} i'$ with $i \in X$, $i' \in Y$, and $i, i' \in dom(x_j)$. Thus, if $X = \{$"$x_K = ultramafic\ soil$", "$x_R = [0, 2000]$", "$mine$"$\}$ and $Y = \{$"$x_R = [2001, 3200]$", "$mine$", "$trail$"$\}$, we have $X \prec_{x_R} Y$.

Based on these definitions, the previous property can be formalized as follows :

Property 3. *Let $X, Y \in \mathcal{L}$ be two itemsets such that $Atts(X, D_{model}) = Atts(Y, D_{model})$. Let us denote by $X.x_j$ the value of attribute x_j in X. Given that for all attributes $x_j \in D_{model}$ of X and Y, we have $\frac{\partial f}{\partial x_j} > 0$ over [a,b] $\wedge$ $X.x_j, Y.x_j \in [a, b]$ $\wedge$ $Y \prec_{x_j} X$ OR $\frac{\partial f}{\partial x_j} < 0$ over [a,b] $\wedge$ $X.x_j, Y.x_j \in [a, b]$ $\wedge$ $X \prec_{x_j} Y$. If $q_{f\geq}(X)$ is false, then $q_{f\geq}(Y)$ is false.*

Note that the impact of this property depends on the discretization. We cannot deduce such a thing with itemsets $X = \{$"$x_1 = 4$", "$\boldsymbol{x_2 \in [\pi, 2\pi[}$", "$x_3 = A$"$\}$ and $Y = \{$"$x_1 = 4$", "$\boldsymbol{x_2 \in [0, \pi[}$", "$x_3 = A$"$\}$, because function f is increasing in x_2 on $[0, \pi]$ and decreasing on $[\pi, 2\pi]$. Let x_j be an attribute such that its domain is discretized in intervals in which f is monotonic. The larger the number of items in each interval is, the more effective the previous property is.

4.3 Pushing expert models into pattern mining

The proposed constraint is relatively simple to integrate in pattern mining algorithms, since it has similar properties to the ones classically used to extract itemsets (e.g., the minimal frequency constraint). Only few modifications have to be done since checking the constraint does not access to the database or other resources. Only the generation of candidate patterns is impacted. The interest of using this threshold constraint based on expert models during extraction (and not in post-processing step) is to quickly prune uninteresting patterns, and this improves performance and scalability.

As an example, we have integrated in our experiments such constraint in the algorithm Close-By-One [15]. Initially developed for formal concept analysis, this algorithm is used in our context to extract closed frequent itemsets. The principle of this algorithm is to perform a depth-first search in the lattice to compute closed patterns. At each step, the algorithm extends currently generated patterns by adding one item and then processes its closure. A canonicity test is also done to avoid redundancy. This algorithm assumes that there is a linear order on the sets of attributes and items. In our context, attributes of the model are enumerated in lexicographic order first, followed by the other attributes of the database (also in in lexicographic order). For each attribute, items are ordered by their value.

Algorithms 1 and 2 describe this approach. Parameter (X, T) of Algorithm 2 represents the current closed pattern X (to extend) and the set of transactions T in which it appears. Parameter A is the itemset used to generate the current pattern X. Parameter i is the last item

added in pattern X, and parameter B is the set of attributes that can be used for extension. Line 1 is the canonicity test used to avoid generating the same pattern twice. Line 5 saves the current pattern in the solutions (with its frequency $|T|$). Lines 8 and 10 enumerate each item that can be used to extend X. Line 9 processes the possible attributes for next extensions. Lines 11-14 calculate the closure of the extensions of X, their transactions and run the next iterations. The notation "$cl(\ )$" in Lines 12-13 represents the closure operator. The closure of a set of items is the set of transactions in which the set of items appears (Line 12). The closure of a set of transactions is the set of items that are common to all input transactions (Line 13).

The only difference between this algorithm and the original one in [15] relies in Lines 2-4, and 7. Line 2 checks the frequency constraint. Line 3 represents *Property 2* of our model-based constraint, i.e., if the upper bound of $f(X)$ (or its value if it is not associated to an interval) is lower than $minf$, then all its supersets can be pruned. Lines 4 and 7 represent *Property 1* of our model-based constraint, i.e., if the extension of X is false w.r.t the model threshold constraint, then all supersets sharing the same attributes w.r.t. D_{model} can be pruned.

Algorithm 1: *CBOwithModelConstrainst(DB, minsup, f)*

Output: The set of closed frequent itemsets $Closed$ whose values by the model f are greater than $minf$ (with their frequency)

1 $Closed \leftarrow \emptyset$
2 **foreach** $d_k \in D_{DB}$ **do**
3 $B \leftarrow \{d_l \in D_{DB} \mid d_k < d_l\}$
4 **foreach** $i \in dom(d_k)$ **do**
5 $Process(\ (cl(cl(\{i\})), cl(\{i\}))\ , \{i\}, i, B, Closed\ )$
6 **return** $Closed$

Algorithm 2: *Process((X,T), A, i, B, Closed)*

1 **if** $\{h \mid h \in X \backslash A \ and \ h \prec i\} = \emptyset$ **then**
2 **if** $|T| \geq minsup$ **then**
3 **if** $sup_X(\ f(X)\ ) < minf$ **then** break
4 **if** $f(X) \geq minf$ **then**
5 $Closed \leftarrow Closed \cup \{(X, |T|)\}$
6 $B_{tmp} \leftarrow B$
7 **else** $B_{tmp} \leftarrow B \backslash \{d \in D_{DB} \mid d \notin D_{model}\}$
8 **foreach** $d_k \in B_{tmp}$ **do**
9 $B \leftarrow B \backslash \{d_l \in D_{DB} \mid d_l \leq d_k\}$
10 **foreach** $j \in \{h \mid h \in dom(d_k) \ and \ i \prec h\}$ **do**
11 $Z \leftarrow X \cup \{j\}$
12 $U \leftarrow T \cap cl(\{j\})$
13 $Y \leftarrow cl(U)$
14 $Process(\ (Y, U), Z, j, B, Closed\ )$

Our approach is totally generic. Most pattern mining algorithms (s.t. Apriori, FP-growth, Eclat) could have been used instead of Close-By-One. However, depending on the algorithm strategy, exploiting some of these properties to prune the search space may not be easy. For example it is difficult to take advantage of property 3 in Close-By-One (due to its candidate generation strategy based on closure), while it is easier for algorithms s.t. Apriori, FP-growth, or Eclat (since itemsets are extended by one item at a time).

5 Experimentations

In our experiment, we have considered a conjunction of two constraints: the minimal frequency constraint and our constraint based on the expert model of Atherton [4]. This constraint has been integrated in the algorithm Close-By-One [15].

Our dataset is based on a satellite image of more than 8 millions of pixels (a 500 Mb SPOT image). The satellite image has been transformed into a transactional database in which each transaction represents informations of one pixel. The attributes of the database correspond to the radiometric properties of pixels (red, green, blue, Brightness Soil Index and Normalized Difference Vegetation Index) that have been discretized. We also add other attributes such as trails, slope, soil type and soil occupation. These last attributes are the ones used by the Atherton model. At the end, we have a dataset with 74 different items and 8 millions of transactions (each transaction being composed of 7 items). Experiments have been done on a PC with 8 Gb of RAM and a 3.20 GHz processor.

In these experiments, we study execution time and number of solutions extracted for different frequency thresholds (x axis), and for different model thresholds ($minf$). Altogether, 16 minimum frequency thresholds have been tested w.r.t. 5 different model thresholds. The experiments with "no model constraint" represent the performance of the state-of-the-art Close-By-One algorithm with the frequency constraint only. Note that 1.5 is the lowest value for the model and 17 is the highest. Thus, $minf = 3$ represents a low threshold. It means that we prune patterns that are not related at all to soil erosion. On the other hand, $minf = 15$ is a very high threshold. It means that, in these experiments, we prune all patterns that are not associated to a strong soil erosion risk by the model.

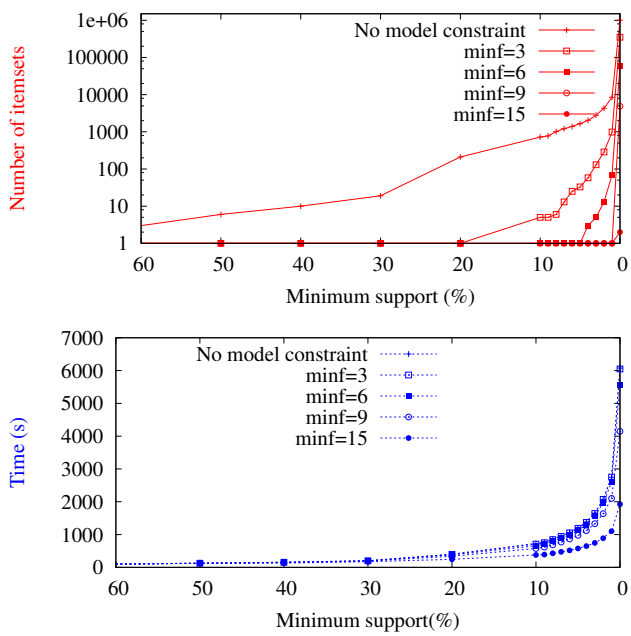

Figure 2. Number of extracted patterns and execution time

The plot at the top of Figure 2 presents the number of extracted patterns for these experiments. Without our model-based constraints, the number of solutions can exceeds 1000 itemsets, for a frequency threshold of 10%. With our model-based constraint, no more than 10 patterns are extracted with the lowest model threshold ($minf = 3$).

These results show that, thanks to expert models, we can easily prune lot of patterns not related at all to the studied phenomenon. Thus, experts have a limited number of patterns related to their problematic to analyze. Note that below 10%, the number of patterns quickly explodes in all cases (even if an important difference remains).
The plot at the bottom presents execution time for the previous experiments. As shown by this figure, for lower frequency thresholds, execution time can exceed 6000 seconds without our model-based constraint (in this case, only the frequency constraint is used). If the model-based constraint is used, then execution time never exceeds 2000 seconds. As expected, the model constraint reduces the number of solutions which accelerates pattern extraction. It also shows that the cost of processing the model does not exceed its benefits from a performance point of view.

Thanks to this constraint, a frequent itemset related to a strong soil erosion risk ($minf = 15$) has been extracted. This itemset is {"Geology=Serpentinites", "LandCover=ultramafic soil on volcano-sedimentary substrat", "Slope=[61,100]", "Red=(14.2,28.4]", "Green=[0.0,36.1]", "NDVI=(-0.071,0.115]", "Blue=[0.0,24.5]"}. This pattern shows that less than 1% of the studied area is associated to a strong soil erosion risk. These high risk areas are characterized by serpentinite soils covered by volcano-sedimentary substrat and have an important slope. Radiometric attributes (not considered by the expert model) confirm this information. Their value shows that we are in presence of low green and NDVI indices, typical of a sparse vegetation. Radiometric attributes have an other interest. These values can be used on other satellite images to identify high risk areas, even if we do not have the geology and the land cover (i.e., input data of the model) on these images. Another example of pattern is {"Geology=Thick laterites on peridotites", "LandCover=Ligno-herbaceous scrub", "Slope=[3.6;30] "}. This pattern is associated by the model to a moderate soil erosion risk ($minf = 6$). Its frequency shows that 4-5$\%$ of the area is characterized by such erosion risk.

6 Conclusion and perspectives

This paper highlights the interest of using domain models in KDD. These models represent a synthesis of the knowledge of a given field and are much richer than basic *"if ... then ..."* rules. In our work, we pushed these models as constraints during pattern mining. They allow a finer analysis while improving performances thanks to some of their properties. Hence, we obtain more relevant patterns that can extend or contradict previous knowledge of studied phenomena.

The perspectives of this work are numerous. A first perspective is to combine several models, each one being weighted by experts depending on application context. Another perspective would be to compare more globally knowledge of one or several models with knowledge extracted using data mining. Finally, it would be interesting to combine, synthesize, knowledge derived from several models. Thus, expert models would become input data for data mining. This kind of approach would allow to extract correlations that are frequently expressed in models of a given field.

REFERENCES

[1] Rakesh Agrawal and Ramakrishnan Srikant, 'Fast algorithms for mining association rules in large databases', in *VLDB*, pp. 487–499, (1994).

[2] Sarabjot S. Anand, David A. Bell, and John G. Hughes, 'The role of domain knowledge in data mining', in *CIKM*, pp. 37–43. ACM Press, (1995).

[3] Cláudia Antunes, 'Mining patterns in the presence of domain knowledge', in *ICEIS (2)*, pp. 188–193, (2009).

[4] James Atherton, David Olson, Linda Farley, and Ingrid Qauqau, 'Fiji Watersheds at Risk: Watershed Assessment for Healthy Reefs and Fisheries', Technical report, Wildlife Conservation Society - South Pacific, Suva, Fiji, (2005).

[5] N.T.J. Bailey, 'Mathematical Theory of Infectious Diseases', *Mathematics in Medicine Series*, (1987).

[6] Jean-François Boulicaut and Baptiste Jeudy, 'Constraint-based data mining', in *Data Mining and Knowledge Discovery Handbook*, eds., Oded Maimon and Lior Rokach, 339–354, Springer, (2010).

[7] Laurent Brisson, Martine Collard, and Nicolas Pasquier, 'Improving the knowledge discovery process using ontologies', in *1st international workshop on Mining Complex Data in conjunction with ICDM (5th IEEE International Conference on Data Mining) Conference, November 27, Houston, USA*, (2005).

[8] M N Burattini, M Chen, A Chow, F A B Coutinho, K T Goh, L F Lopez, S Ma, and E Massad, 'Modelling the control strategies against dengue in Singapore', *Epidemiology and infection*, **136**(3), 309–19, (March 2008).

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

[10] Líliam de Castro Medeiros, César Castilho, Cynthia Braga, Wayner de Souza, Leda Regis, and Antonio Monteiro, 'Modeling the dynamic transmission of dengue fever: investigating disease persistence.', *PLoS neglected tropical diseases*, **5**(1), (January 2011).

[11] Pedro Domingos, 'Toward knowledge-rich data mining', *Data Mining and Knowledge Discovery*, **15**(1), 21–28, (April 2007).

[12] Szymon Jaroszewicz, Tobias Scheffer, and Dan A. Simovici, 'Scalable pattern mining with bayesian networks as background knowledge', *Data Mining and Knowledge Discovery*, **18**(1), 56–100, (2009).

[13] Szymon Jaroszewicz and Dan A. Simovici, 'Interestingness of frequent itemsets using bayesian networks as background knowledge', in *SIGKDD*, pp. 178–186, (2004).

[14] Kleanthis-Nikolaos Kontonasios, Jilles Vreeken, and Tijl De Bie, 'Maximum entropy models for iteratively identifying subjectively interesting structure in real-valued data', in *ECML/PKDD (2)*, eds., Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Zelezný, volume 8189 of *Lecture Notes in Computer Science*, pp. 256–271. Springer, (2013).

[15] Sergei O. Kuznetsov and Sergei Obiedkov, 'Comparing performance of algorithms for generating concept lattices', *Journal of Experimental and Theoretical Artificial Intelligence*, **14**, 189–216, (2002).

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

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

[18] Ken McGarry, 'A survey of interestingness measures for knowledge discovery', *The Knowledge Engineering Review*, **20**(01), 39, (December 2005).

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

[20] Raymond T. Ng, Laks V. S. Lakshmanan, Jiawei Han, and Alex Pang, 'Exploratory mining and pruning optimizations of constrained associations rules', *ACM SIGMOD Record*, **27**(2), 13–24, (June 1998).

[21] Balaji Padmanabhan and Alexander Tuzhilin, 'A belief-driven method for discovering unexpected patterns', in *KDD*, pp. 94–100, (1998).

[22] Christine Parent, Stefano Spaccapietra, Chiara Renso, Gennady Andrienko, Natalia Andrienko, Vania Bogorny, Maria Luisa Damiani, Aris Gkoulalas-Divanis, Jose Macedo, Nikos Pelekis, Yannis Theodoridis, and Zhixian Yan, 'Semantic trajectories modeling and analysis', *ACM Comput. Surv.*, **45**(4), 42:1–42:32, (August 2013).

[23] Jian Pei, Jiawei Han, and Laks V. S. Lakshmanan, 'Mining frequent item sets with convertible constraints', in *ICDE*, pp. 433–442, (2001).

[24] Luc De Raedt and Albrecht Zimmermann, 'Constraint-based pattern set mining', in *SDM*, pp. 237–248, (2007).

[25] 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.