Soft Threshold Constraints for Pattern Mining

Willy Ugarte, Patrice Boizumault, Samir Loudni, and Bruno Crémilleux

GREYC (CNRS UMR 6072) – University of Caen Basse-Normandie,
Campus II, Côte de Nacre, 14000 Caen, France
firstname.lastname@unicaen.fr

Abstract. Constraint-based pattern discovery is at the core of numerous data mining tasks. Patterns are extracted with respect to a given set of constraints (frequency, closedness, size, etc). In practice, many constraints require threshold values whose choice is often arbitrary. This difficulty is even harder when several thresholds are required and have to be combined. Moreover, patterns barely missing a threshold will not be extracted even if they may be relevant. In this paper, by using Constraint Programming we propose a method to integrate soft threshold constraints into the pattern discovery process. We show the relevance and the efficiency of our approach through a case study in chemoinformatics for discovering toxicophores.

1 Introduction

Extracting knowledge from large amounts of data is at the core of the Knowledge Discovery in Databases. This involves different challenges, such as designing efficient tools to tackle data and the discovery of patterns of a potential user's interest. Many authors [9,10] have promoted the use of constraints to represent background knowledge and to focus on the most promising knowledge by reducing the number of extracted patterns to those of a potential interest given by the final user. The most popular example with local patterns is the minimal frequency constraint based on the frequency measure: it addresses all patterns having a number of occurrences in the database exceeding a given minimal threshold.

In practice, data mining tasks require to deal both with pattern characteristics (e.g., frequency, size, contrast [11]) and background knowledge (e.g., price in the traditional example of supermarket databases, chemical features such as aromaticity in chemoinformatics). Then several measures have to be handled and combined leading to entail choosing several threshold values.

This notion of thresholding has serious drawbacks. Firstly, unless specific domain knowledge is available, the choice is often arbitrary and relevant patterns are missed or lost within a lot of spurious patterns. This drawback is obviously even deeper when several measures have to be combined and thus several thresholds are needed. A second drawback is the stringent aspect of the classical constraint-based mining framework: a pattern satisfies or does not satisfy the set of constraints. But, what about patterns that respect only some thresholds,

J.-G. Ganascia, P. Lenca, and J.-M. Petit (Eds.): DS 2012, LNAI 7569, pp. 313–327, 2012.

especially if only very few constraints are slightly violated? There are very few works such as [3] which propose to introduce a softness criterion into the mining process as we will see in Section 6. This thresholding issue is also present in pattern set mining [15] where the goal is to mine for a set of patterns with constraints combing several local patterns. An example is the top-k pattern approaches [7,16]: by associating each pattern with a rank score, these approaches return an ordered list of the k patterns with the highest score to the user [16]. However, the performance of top-k approaches are sensitive to both the threshold value k and the thresholds of the aggregated measures in the score function. This paper deals with these issues.

The key contribution of this paper is that we show how constraint relaxation, developed for Constraint Programming, can be applied to propose a soft constraint based pattern mining framework.

Our approach benefits from the recent progress on cross-fertilization between data mining and Constraint Programming [8,14,6]. The common point of all these methods is to model in a declarative way pattern mining as Constraint Satisfaction Problems (CSP), whose resolution provides the complete set of solutions satisfying all the constraints.

Our approach proceeds as follows. First, to each soft threshold constraint is associated a violation measure to determine the distance between a pattern and a threshold. Then, soft threshold constraints are transformed into equivalent hard constraints that can be directly handled by a CSP solver. We show how soft threshold constraints can be exploited for extracting the top-k patterns according to an interestingness measure. The technique fully benefits from the handling of the soft threshold constraints: contrary to the data mining methods, the top-k patterns can include patterns violating constraints on the measures given by the user. Our method offers a natural way to simultaneously combine in a same framework usual data mining measures with measures coming from the background knowledge. The relevance of our approach is highlighted through a case study in chemoinformatics for discovering toxicophores.

This paper is organized as follows. Section 2 presents the context. Section 3 describes the disjunctive relaxation framework we used to model and solve soft threshold constraints. Section 4 focusses on mining top-k patterns. Section 5 presents the case study in chemoinformatics for discovering toxicophores and reports our experimental results. Finally, we review related work in Section 6.

2 Context and Motivations

2.1 Definitions

Let $\mathcal{I}$ be a set of distinct literals called *items*. An itemset (or pattern) is a non-null subset of $\mathcal{I}$. The language of itemsets corresponds to $\mathcal{L}_{\mathcal{I}} = 2^{\mathcal{I}} \backslash \emptyset$. A transactional dataset is a multiset of patterns of $\mathcal{L}_{\mathcal{I}}$. Each pattern (or transaction) is a database entry. Table 1 (left side) presents a transactional dataset $\mathcal{T}$ where each transaction t_i gathers articles described by items denoted A,...,F. The traditional example is a supermarket database in which each transaction corresponds

Table 1. Transactional dataset $\mathcal{T}$

Trans.	Items					
t_1		B			E	F
t_2		B	C	D		
t_3	A				E	F
t_4	A	B	C	D	E	
t_5		B	C	D	E	
t_6		B	C	D	E	F
t_7	A	B	C	D	E	F

Items	A	B	C	D	E	F
Price	30	40	10	40	70	55

to a customer and every item in the transaction to a product bought by the customer. A price is associated to each product (cf. Table 1, right side).

Constraint-based pattern mining aims at extracting all patterns of $\mathcal{L}_{\mathcal{I}}$ satisfying a query (conjunction of constraints). A very usual example is the frequency measure leading to the minimal frequency constraint. The latter provides patterns X_i having a number of occurrences in the database exceeding a given minimal threshold min_{fr}: $freq(X_i) \geq min_{fr}$. Another well-known measure is the *size* of a pattern, i.e. the number of items that a pattern contains. In many applications, it appears highly appropriate to look for contrasts between subsets of transactions, such as toxic and non toxic molecules in chemoinformatics. The growth rate is a well-used contrast measure [11]. Let $\mathcal{T}$ be a database partitioned into two subsets $\mathcal{D}_1$ and $\mathcal{D}_2$:

Definition 1 (Growth rate). *The growth rate of a pattern X_i from $\mathcal{D}_2$ to $\mathcal{D}_1$ is:*

$$m_{gr}(X_i) = \frac{|\mathcal{D}_2| \times freq(X_i, \mathcal{D}_1)}{|\mathcal{D}_1| \times freq(X_i, \mathcal{D}_2)}$$

Emerging Patterns and Jumping Emerging Patterns stem from this measure. They are at the core of a useful knowledge in many applications involving classification features such as the discovery of structural alerts in chemoinformatics.

Definition 2 (Emerging Pattern). *Given a threshold $min_{gr} > 1$, a pattern X_i is said to be an Emerging Pattern (EP) from $\mathcal{D}_2$ and $\mathcal{D}_1$ if $m_{gr}(X_i) \geq min_{gr}$.*

Definition 3 (Jumping Emerging Pattern). *A pattern X_i which does not occur in $\mathcal{D}_2$ ($m_{gr}(X_i) = +\infty$) is called a Jumping Emerging Pattern (JEP).*

Moreover, the user is often interested in discovering richer patterns satisfying properties involving several local patterns. These patterns correspond to pattern sets [15] or n-ary patterns [8]. The approach that we present in this paper is able to deal with pattern sets such as the top-k patterns.

2.2 Motivating Example

Example 1. Let us consider the following query $q(X_i)$. It addresses all frequent patterns ($min_{fr} = 4$), having a size greater than or equal to 3, and an average price (*avgPrice*) greater than 45:

$q(X_i) \equiv \mathit{freq}(X_i) \geq 4 \wedge \mathit{size}(X_i) \geq 3 \wedge \mathit{avgPrice}(X_i) \geq 45$

Thereafter, we use the notation $X_i < v_1,v_2,v_3 >$, where X_i is a pattern, and v_1, v_2, v_3 denote its value for the three measures: *freq*, *size* and *avgPrice*. When considering only the frequency constraint, we get 17 solutions. With the conjunction of the three constraints, there is only one solution: $BDE < 4{,}3{,}50 >$. Let us consider the following four patterns which are missed by the mining process:

- $BEF < 3, 3, 55 >$
- $BCE < 4, 3, 40 >$
- $CDE < 4, 3, 40 >$
- $BCDE < 4, 4, 40 >$

The pattern BEF slightly violates the frequency threshold and satisfies the two other constraints. However, this pattern is clearly interesting because its value on the average price measure is largely higher than the value of BDE which satisfies the query. By slightly relaxing the frequency threshold ($\mathit{freq}(X_i) \geq 3$), BEF would be extracted.

Similarly, relaxing the average price threshold ($\mathit{avgPrice}(X_i) \geq 40$) would enable to discover three new patterns: CDE, BCE and $BCDE$. Due to the uncertainty inherent to the determination of the thresholds, it is difficult to say that these patterns are less interesting than BDE which is produced. So, the stringent aspect of the classical constraint-based mining framework means that interesting patterns are lost as soon as at least one threshold is slightly violated. Moreover, in real life applications, all threshold constraints are not considered to be equally important, and this characteristic should be taken into account in the mining process. Overcoming these drawbacks is the motivation of our proposal.

3 Modeling and Solving Soft Threshold Constraints

This section shows how soft threshold constraints can be transformed into equivalent hard constraints that can be directly handled by a CSP solver with a method using the disjunctive relaxation framework [12].

3.1 Constraint Relaxation

Constraint relaxation is a technique to deal with over-constrained problems, i.e., problems with no solution satisfying all the constraints. Over-constrained problems are generally modeled as Constraint Optimization Problems (COP). Violation measures associate costs to constraints in order to quantify their violation. A global objective related to the whole set of costs is usually defined (for example to minimize the total sum of costs).

Definition 4. (violation measure). *μ_c is a violation measure for the constraint $\mathtt{c}(X_1, ..., X_n)$ iff μ_c is a function from $D_1 \times D_2 \times ... \times D_n$ to $\Re^+$ s.t. $\forall A \in D_1 \times D_2 \times ... \times D_n$, $\mu_c(A) = 0$ iff A satisfies $c(X_1, ..., X_n)$.*

For a given constraint, several violation measures can be defined. For the soft threshold constraints which will be studied in Section 3.3, we propose two different violation measures.

3.2 Disjunctive Relaxation

Over constrained problems can be modeled using disjunctive relaxation [12]. To each soft constraint c are associated a violation measure μ_c and a cost variable z_c that measure the violation of c. So the COP is transformed into a CSP where all constraints are hard and the cost variable $z = \sum_c z_c$ will be minimized. If the domain of a cost variable is reduced during the search, propagation will be performed on domains of other cost variables. Each soft constraint is modeled as a disjunction: either the constraint is satisfied and the cost is null, or the constraint is not satisfied and the cost is specified.

Definition 5 (disjunctive relaxation of a constraint). *Let c be a constraint, $\bar{c}$ its negation and z the associated cost variable. The disjunctive relaxation of c is $c' \equiv [c \wedge (z = 0)] \vee [\bar{c} \wedge (z > 0)]$*

Example 2. Let X_1=X_2 be a binary constraint over variables X_1 and X_2 with domains D_1=D_2=$\{1,2,3\}$. Let z be the associated cost variable and μ the violation measure defined as the distance between the two variables. The disjunctive relaxation of $c \equiv (X_1{=}X_2)$ is $c' \equiv [X_1 = X_2 \wedge z = 0] \vee [X_1 \neq X_2 \wedge z = |X_1 - X_2|]$.

We have selected the disjunctive relaxation framework for two reasons. First, as any soft threshold constraint can be transformed into an equivalent hard constraint (Section 3.4), this enables to integrate relaxation in existing CSP solvers and to benefit from progress made in this area. Moreover, we can directly include soft threshold constraints in our n-ary pattern extractor based on Constraint Programming [8].

3.3 Violation Measures for Soft Threshold Constraints

In this section, we take as an introductory example the frequency measure, then we consider any measure.

Frequency measure. Let X_i be a pattern, α a minimal threshold and the constraint $\mathit{freq}(X_i) \geq \alpha$. A first violation measure μ_1 can be defined as the absolute distance from threshold α. However, to combine violations of several threshold constraints, it is more appropriate to consider relative distances. A second violation measure μ_2 can be defined as the relative distance from α:

$$\mu_2(X_i) = \begin{cases} 0 & \text{if } \mathit{freq}(X_i) \geq \alpha \\ \frac{\alpha - \mathit{freq}(X_i)}{\alpha} & \text{otherwise} \end{cases}$$

For any measure m. Let $\mathcal{I}$ be a set of distinct items and $\mathcal{T}$ a set of transactions. Let max_m be the maximum value[1] for measure m. Violation measures are defined as follows:

[1] For the frequency measure, max_m=$|\mathcal{T}|$; for the size measure, max_m=$|\mathcal{I}|$.

$$For \quad c \equiv m(X_i) \geq \alpha \qquad \mu_2(X_i) = \begin{cases} 0 & if \;\; m(X_i) \geq \alpha \\ \frac{\alpha - m(X_i)}{\alpha} & otherwise \end{cases}$$

$$For \quad c \equiv m(X_i) \leq \alpha \qquad \mu_2(X_i) = \begin{cases} 0 & if \;\; m(X_i) \leq \alpha \\ \frac{m(X_i) - \alpha}{max_m - \alpha} & otherwise \end{cases}$$

Violation measures are normalized in order to combine violations of several threshold constraints occurring in a same query. For semantic μ_2, violation values will be real numbers ranging from 0.0 to 1.0.

3.4 From Soft Constraints to Equivalent Hard Constraints

This section shows how to transform any soft threshold constraint into an equivalent hard constraint. First, we present the CSP modeling for the n-ary pattern mining problem. Then, we describe the transformation and the resulting CSP.

Initial CSP. Let $\mathcal{T}$ be a set of transactions and $\mathcal{I}$ the set of all its items. The n-ary itemset mining problem can be modeled as a CSP $\mathcal{P}$=($\mathcal{X}, \mathcal{D}, \mathcal{C}$) where:

- $\mathcal{X} = \{X_1, ..., X_n\}$. Each variable X_i represents an unknown pattern.
- $\mathcal{D} = \{D_{X_1}, ..., D_{X_n}\}$. Initial domain of X_i is the set interval $[\emptyset \, .. \, \mathcal{I}]$.
- $\mathcal{C} = \mathcal{C}_{ens} \cup \mathcal{C}_{num}$ is the whole set of constraints where:
  - $\mathcal{C}_{ens}$ is a conjunction of set constraints handling set operators. Examples: $X_1 \subset X_2$, $I \in X_4$, ...
  - $\mathcal{C}_{num}$ is is a conjunction of numerical constraints. Examples: $|freq(X_1) - freq(X_2)| \leq \alpha_1$, $size(X_4) < size(X_1) + 1, \ldots$

More information on the implementation of the above constraint-based pattern mining task using Constraint Programming techniques are in [8,6].

Transformation for the frequency measure. Let X_i be a pattern, α a minimal threshold and the constraint $c \equiv freq(X_i) \geq \alpha$. Let z be the associated cost variable. The disjunctive relaxation of c for μ_2 is:

$$[(freq(X_i) \geq \alpha) \wedge z = 0] \;\; \vee [(freq(X_i) < \alpha) \wedge z = \frac{\alpha - freq(X_i)}{\alpha}]$$

This disjunction can be reformulated in an equivalent way by the following (hard) constraint:

$$z = max(0, \frac{\alpha - freq(X_i)}{\alpha})$$

Transformation for any measure *m*. By applying the previous transformation, soft threshold constraints associated to a measure m can be transformed into equivalent hard constraints:

- The relaxation of $c \equiv (m(X_i) \geq \alpha)$ is $c' \equiv [z = max(0, \frac{\alpha - m(X_i)}{\alpha})]$
- The relaxation of $c \equiv (m(X_i) \leq \alpha)$ is $c' \equiv [z = max(0, \frac{m(X_i) - \alpha}{max_m - \alpha})]$

Thus, any query containing one or more soft threshold constraints c_i can be transformed into an *equivalent* query with only hard constraints: if c_i is a hard constraint then it remains unchanged; if c_i is a soft threshold constraint then it is replaced by its transformation. Then, we define the global cost variable $z{=}\sum_{c_i} z_i$ representing the total sum of violations, where z_i is the cost variable associated to the soft threshold constraint c_i. Finally, we add the constraint $z \leq \lambda$, where λ is the maximum amount of violation that is allowed ($\lambda \in [0.0\,,1.0]$). This parameter (λ) quantifies a deviation from the measure thresholds, thus its semantics is understandable to the user.

Resulting CSP. Let $\mathcal{P}'{=}(\mathcal{X}',\mathcal{D}',\mathcal{C}')$ be the CSP obtained by the disjunctive relaxation of the initial CSP $\mathcal{P}{=}(\mathcal{X},\mathcal{D},\mathcal{C})$:

- $\mathcal{X}' = \mathcal{X} \bigcup_{1 \leq i \leq k} \{z_i\} \cup \{z\}$,
- $\mathcal{D}' = \mathcal{D} \bigcup_{1 \leq i \leq k} \{D_{z_i}\} \cup \{D_z\}$ with $D_{z_i}{=}[0.0\,,1.0]$ and $D_z{=}[0.0\,,\lambda]$,
- $\mathcal{C}' = \mathcal{C}_{ens} \cup \mathcal{C}'_{num} \cup \{z = \sum_{1 \leq i \leq k} z_i\}$ with $\mathcal{C}'_{num} = \mathcal{C}_{hard} \cup \mathcal{C}_{disj}$ where:
  - $\mathcal{C}_{hard}$ is the set of (initial) hard numerical constraints,
  - $\mathcal{C}_{disj}$ is the set of hard constraints associated to the soft threshold constraints.

The steps presented above lead to a soft constraint based pattern mining framework. The next section shows how this framework also addresses pattern sets such as the top-k patterns.

4 Mining top-*k* Patterns with an Interestingness Measure

Looking for the k patterns optimizing an interestingness measure is an attractive data mining task [7,16]. These pattern sets are called top-k patterns. The top-k pattern methods associate each pattern with a rank score and compute an ordered list of the k patterns with the highest score. Rank scores are determined by interestingness measures provided by the user [7,16]. In this section, we define an interestingness measure enabling us to exploit our method on pattern mining with soft threshold constraints. The technique fully benefits from the handling of the soft threshold constraints: the top-k patterns can include patterns violating constraints on the measures given by the user. Up to now, data mining techniques are not able to take into account softness in top-k mining.

Let us consider the constraint $\mathit{freq}(X_i) \geq \alpha$. A pattern X_i having a frequency much larger than the threshold α, will be considered as more interesting than a pattern X_j whose frequency is slightly higher than α.

Interestingness of a pattern for a threshold constraint. An interestingness measure of a pattern for a threshold constraint c may be either positive (when c is satisfied) or negative (when c is not satisfied). As for a violation measure (see Section 3.3), an interestingness measure is also normalized in order to combine interests of several threshold constraints occurring in a same query. Let $\mathcal{M}$ be a set of measures. Let $m \in \mathcal{M}$ be a measure, and max_m its maximal value.

We define the interestingness measure $\theta_m :: \mathcal{L}_{\mathcal{I}} \rightarrow$ [-1.0 , 1.0] by:

$$For \quad c \equiv m(X_i) \geq \alpha \qquad \theta_m(X_i) = \begin{cases} \frac{m(X_i)-\alpha}{max_m - \alpha} \; if \;\; m(X_i) \geq \alpha \\ \\ -\mu_2(X_i) \; otherwise \end{cases}$$

$$For \quad c \equiv m(X_i) \leq \alpha \qquad \theta_m(X_i) = \begin{cases} \frac{\alpha - m(X_i)}{\alpha} \; if \;\; m(X_i) \leq \alpha \\ \\ -\mu_2(X_i) \; otherwise \end{cases}$$

Interestingness of a pattern for a query. Let $\mathcal{M}$ be a set of measures and a query expressed as a conjunction of threshold constraints for these measures. We define the interestingness of a pattern for a query as the sum of the interests of this pattern for threshold constraints.

$$\theta(X_i) = \textstyle\sum_{m \in \mathcal{M}} \gamma_m \times \theta_m(X_i)$$

where γ_m is a coefficient reflecting the importance of the measure m.

The top-k patterns are extracted w.r.t the measure θ.

Computing the top-*k* patterns. Let $q(X_i)$ be a query involving soft threshold constraints and λ the maximal amount of violation that is allowed. Let $q'(X_i)$ be the hard query associated to both $q(X_i)$ and λ (see Section 3.4).

Computing the top-k patterns, for the query $q'(X_i)$ according to the interestingness measure θ, is performed as follows. The first k solutions $(s_1, s_2, ..., s_k)$ for the query $q'(X_i)$ are computed and ordered according to the interestingness measure θ. Then, as soon as a new solution s' is obtained, if $(\theta(s') > \theta(s_k))$ then s' is inserted in the top-k solutions and s_k is removed. Furthermore, the constraint $(\theta(X_i) > \theta(s_k))$ is dynamically posted in order to improve the pruning of the search tree.

5 Experiments

Toxicology is a scientific discipline involving the study of the toxic effects of chemicals on living organisms. A major issue in chemoinformatics is to establish relationships between chemicals and a given activity (e.g., CL50[2] in ecotoxicity). Chemical fragments[3] which cause toxicity are called *toxicophores* and their

[2] Lethal concentration of a substance required to kill half the members of a tested population after a specified test duration.

[3] A fragment denominates a connected part of a chemical structure containing at least one chemical bond

discovery is at the core of prediction models in (eco)toxicity [1,13]. The aim of this present study, which is part of a larger research collaboration with the CERMN Lab[4], a laboratory of medicinal chemistry, is to investigate the use of soft threshold constraints for discovering toxicophores.

5.1 Settings

The dataset is collected from the ECB web site[5]. For each chemical, the chemists associate it with hazard statement codes (HSC) in 3 categories: H400 (very toxic, CL50 $\leq$ 1 mg/L), H401 (toxic, 1 mg/L $<$ CL50 $\leq$ 10 mg/L), and H402 (harmful, 10 mg/L $<$ CL50 $\leq$ 100 mg/L). We focus on the H400 and H402 classes. The dataset $\mathcal{T}$ consists of 567 chemicals, 372 from the H400 class and 195 from the H402 class. The chemicals are encoded using 129 frequent subgraphs previously extracted from $\mathcal{T}$[6] with a 10% relative frequency threshold (experiments with lower thresholds did not bring significant results for the chemists).

In order to discover patterns as candidate toxicophores, we use both measures typically used in contrast mining [11] such as the growth rate since toxicophores are linked to a classification problem with respect to the HSC and measures expressing the background knowledge such as the aromaticity or density because chemists consider that this information may yield promising candidate toxicophores. Our method offers a natural way to simultaneously combine in a same framework these measures coming from various origins. We briefly sketch these measures and the associated threshold constraints.

Growth rate. When a pattern has a frequency which significantly increases from the H402 class to the H400 class, then it stands a potential structural alert related to the toxicity. In other words, if a chemical has in its structure fragments that are related to a toxic effect, then it is more likely to be toxic. Emerging patterns embody this natural idea by using the growth-rate measure (cf. Section 2.1). Let min_{gr} be a minimal growth threshold. We impose the soft threshold constraint: $m_{gr}(X_i) \geq min_{gr}$.

Frequency. Real-world datasets are often noisy and patterns with low frequency may be artefacts. The minimal frequency constraint ensures that a pattern is representative enough (i.e., the higher the frequency, the better it is). Thus we use the following soft threshold constraint: $freq(X_i) \geq min_{fr}$, where min_{fr} is a minimal frequency threshold.

Aromaticity. Chemists know that the aromaticity is a chemical property that favors toxicity since their metabolites can lead to very reactive species which can interact with biomacromolecules in a harmful way. We compute the aromaticity of a pattern as the mean of the aromaticity of its chemical fragments. Let m_a

[4] Centre d'Etudes et de Recherche sur le Médicament de Normandie, UPRES EA 4258 FR CNRS 3038, Université de Caen Basse-Normandie.

[5] European Chemicals Bureau `http://ecb.jrc.ec.europa.eu/documentation/` now `http://echa.europa.eu/`

[6] A chemical Ch contains an item A if Ch supports A, and A is a frequent subgraph of $\mathcal{T}$.

be the aromaticity measure of a pattern. We get the soft threshold constraint: $m_a(X_i) \geq min_a$, where min_a is a minimal aromaticity threshold.

Density. In addition, chemists consider that the density of chemicals may yield an interest for candidate toxicophores. A common hypothesis is that the higher the chemical density, the stronger its chemical behavior. The density of a pattern is given by the mean of density of its subgraphs[7]. Let m_d be the density measure of a pattern and min_d a minimal threshold leading to the soft threshold constraint: $m_d(X_i) \geq min_d$
Finally, we get the following query $q(X_i)$:

$$m_{gr}(X_i) \geq min_{gr} \wedge freq(X_i) \geq min_{fr} \wedge m_a(X_i) \geq min_a \wedge m_d(X_i) \geq min_d$$

5.2 Experimental Protocol

The thresholds on aromaticity and density measures were set to 2/3 of the maximal values of these measures on the dataset (min_a=60 and min_d=60). Indeed, high thresholds suggest an interest for candidate toxicophores. The minimal growth rate and the minimal frequency thresholds were fixed to 1/4 of the maximal values of these measures (min_{gr}=5 and min_{fr}=90) in order to keep only the most frequent emerging patterns (EPs) with the highest growth rates. Setting these thresholds might be subtle and it illustrates the interest of the soft constraints because the choice of the user is then downplayed.

We consider three different values for $\lambda : \{0, 20\%, 40\%\}$. For the interestingness measure θ, we set γ_{gr}, γ_{fr} and γ_d to 1 et γ_a to 2. Indeed, aromaticity is the most important chemical knowledge. The extracted EPs are made of molecular fragments and to evaluate the presence of toxicophores in their description, we identified six fragments based on well-known environmental toxicophores, namely the benzene, the phenol ring, the chloro-substituted aromatic ring (i.e., chlorobenzene), the organo-phosphorus moiety, the aniline aromatic ring, and the pyrrole.

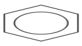
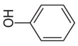
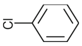
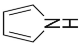
Experiments were conducted on a computer running Linux operating system with a core i3 processor at 2,13 GHz and a RAM of 4 GB. The implementation of our approach was carried out in `Gecode` by extending the n-ary patterns extractor based-CSP [8].

5.3 Extracting Frequent Emerging Patterns

Table 2 depicts the numbers of EPs containing at least one complete toxicophore compound (columns marked **T**) or sub-fragments of a toxicophore (columns marked **F**) among the six fragments previously identified in the database according to the three values of λ. Col. 2-7 provide the total number of solutions, Col. 8-13 over the top_{25} and Col. 14-19 over the top_{50}. As the two categories **T** and **F** are not disjoint, the cumul of the number of EPs in the two categories

[7] The density of a subgraph is equal to $2e/v(v-1)$, where e (resp. v) is the number of its edges (resp. vertices).

Table 2. Numbers of emerging patterns according to known toxicophores

	Total						top-25						top-50					
λ	0		20%		40%		0		20%		40%		0		20%		40%	
# Solutions	**7650**		**402204**		**4289335**		**28**		**37**		**57**		**55**		**64**		**85**	
	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F	T	F
Benzene c1ccccc1 	1912	7573	183881	396749	1565883	4210482	0	25	**2**	25	**6**	24	7	50	**7**	50	**8**	49
Phenol c1(ccccc1)O 	900	4519	93632	217195	556890	3234279	2	9	**6**	3	**2**	0	4	18	**9**	12	**5**	8
Chlorobenzene Clc1ccccc1 	0	3041	74182	184502	253429	509281	0	14	**2**	14	**2**	1	0	28	**7**	22	**2**	15
Pyrrole c1cncc1 						1						1						1

may exceed #(Solutions). The CPU time for extracting the set of all solutions is 16 s. for (λ=0), 2 min. for (λ=20%) and 2h22 min. for (λ=40%).

As shown in Table 2, 45%[8](resp. 36.5%) of EPs with λ=20% (resp. 40%) contain a benzene (fragment of category **T**), against 25% for λ=0. Thus, soft thresholds allow to better discover this toxicophore (average gain of about 16%). Regarding the category **F**, the proportion of EPs containing sub-fragments of benzene (Smiles code[9]): $\{cc, ccc, cccc, ccccc\}$) is almost the same in the hard and soft cases (about 98%). This trend is also confirmed for phenol ring, where 23% (resp. 13%) of extracted solutions with λ=20% (resp. 40%) include such a fragment, against 11% for λ=0. Once again, soft thresholds enable to better meet this toxicophore (average gain of about 7%).

For the chlorobenzene (with $\lambda = 0$), only patterns containing fragments of category **F** are extracted : $\{Clc(c)cc$, $Clc(c)ccc$, $Clc(c)cccc$, $Clc(cc)ccc$, $Clccc\ldots\}$. The soft thresholds enable to find on average 19% of toxicophores containing the chlorobenzene (i.e., fragment of category **T**). Moreover, for pyrrole, a new pattern with a novel chemical characteristic (containing the subfragment *nc*) is discovered. Indeed, this derivative, not detected with ($\lambda = 0$), is rather difficult to extract as it is associated to a chemical fragment with a low value of frequency.

EPs containing the aniline aromatic ring are not detected because of their low density (33). Indeed, with λ=40%, the minimal value allowed is 60×0.60=36. Increasing very slightly λ (λ=45%), would permit the extraction of those EPs. Finally, the organo-phosphorus fragment has the highest growth rate ($+\infty$) and thus is a JEP (cf. Definition 2). The chemists have a strong interest for such patterns. They are not listed in Table 2 and we will come back on these patterns in Section 5.5.

[8] Ratio of the number of solutions containing a toxicophore by the total number of solutions.

[9] Smiles code is a line notation for describing the structure of chemical molecules : `http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html`

Table 3. *top*$_{25}$ emerging patterns

(a) *top*$_{25}$ EP with λ=20%.

N	Interest	Pattern	Growth rate	Frequency	Aromaticity	Density	SMILES
1	193	24 35 69	7	101	95	66	cc ccc c1(ccccc1)O
2	191	13 24 35	8	89	95	66	Clc1ccccc1 cc ccc
3	189	24 35 47 69	7	101	96	62	cc ccc cccc c1(ccccc1)O
4	187	13 24 35 47	8	89	96	62	Clc1ccccc1 cc ccc cccc
5	185	12 24 35 47	8	90	96	61	Clc(c)cccc cc ccc cccc
6	185	14 24 35 47	8	90	96	61	Clcccccc cc ccc cccc
7	185	24 35 47 68	6	103	96	61	cc ccc cccc ccccc(c)O
8	185	24 35 47 80	6	103	96	61	cc ccc cccc ccc(ccc)O
9	185	24 35 38	5	118	90	72	cc ccc cccO
10	184	24 35 47 78	8	89	96	61	cc ccc cccc Clc(cc)ccc
11	184	6 24 35	9	93	90	72	Clc(c)c cc ccc
12	184	8 24 35 47	9	93	94	64	Clc(c)cc cc ccc cccc
13	184	8 24 35	9	93	92	68	Clc(c)cc cc ccc
14	184	7 24 35	9	94	90	72	Clccc cc ccc
15	184	9 24 35 47	9	94	94	64	Clcccc cc ccc cccc
16	184	9 24 35	9	94	92	68	Clcccc cc ccc
17	183	24 35 47 59 69	7	101	97	57	cc ccc cccc ccccc c1(ccccc1)O
18	183	24 35 47 69 77	7	101	97	57	cc ccc cccc c1(ccccc1)O c1ccccc1
19	183	24 35 59 69	7	101	96	59	cc ccc ccccc c1(ccccc1)O
20	183	24 35 69 77	7	101	96	59	cc ccc c1(ccccc1)O c1ccccc1
21	183	12 24 35	8	90	94	64	Clc(c)cccc cc ccc
22	183	14 24 35	8	90	94	64	Clcccccc cc ccc
23	183	11 24 35 47	9	92	95	62	Clccccc cc ccc cccc
24	183	11 24 35	9	92	93	66	Clccccc cc ccc
25	183	10 24 35 47	9	93	95	62	Clc(c)ccc cc ccc cccc

(b) *top*$_{25}$ EP with λ=40%.

N	Interest	Pattern	Growth rate	Frequency	Aromaticity	Density	SMILES
26	301	24	3	289	100	100	cc
27	275	15	7	65	100	100	nc
28	258	24 35	3	288	100	83	cc ccc
29	237	24 47	3	281	100	75	cc cccc
30	230	24 35 47	3	281	100	72	cc ccc cccc
31	224	24 59	3	279	100	70	cc ccccc
32	223	24 77	3	274	100	70	cc c1ccccc1
33	219	24 35 59	3	279	100	68	cc ccc ccccc
34	218	24 35 77	3	274	100	68	cc ccc c1ccccc1
35	216	35	3	288	100	65	ccc
36	213	24 35 76	3	274	100	66	cc ccc cccccc
37	213	24 76	3	274	100	66	cc cccccc
38	209	24 35 47 59	3	279	100	64	cc ccc cccc ccccc
39	208	24 35 47 77	3	274	100	64	cc ccc cccc c1ccccc1
40	206	24 47 59	3	279	100	63	cc cccc ccccc
41	205	24 47 77	3	274	100	63	cc cccc c1ccccc1
42	203	24 35 47 76	3	274	100	62	cc ccc cccc cccccc
43	200	24 47 76	3	274	100	61	cc cccc cccccc
44	200	24 35 59 77	3	274	100	61	cc ccc ccccc c1ccccc1
45	198	24 59 77	3	274	100	60	cc ccccc c1ccccc1
46	193	24 35 69	7	101	95	66	cc ccc c1(ccccc1)O
47	191	13 24 35	8	89	95	66	Clc1ccccc1 cc ccc
48	189	24 35 47 69	7	101	96	62	cc ccc cccc c1(ccccc1)O
49	187	13 24 35 47	8	89	96	62	Clc1ccccc1 cc ccc cccc
50	185	12 24 35 47	8	90	96	61	Clc(c)cccc cc ccc cccc

5.4 Mining the top-*k* Patterns

Results from Table 2 show that among the *top*$_{25}$ (resp. *top*$_{50}$) EPs mined with λ=0, only 2 (resp. 4) patterns contain the phenol ring. Moreover, the *top*$_k$ EPs are constituted solely of subfragments of benzene or chlorobenzene.

Table 3a gives the *top*$_{25}$ EPs extracted with λ=20%. Yellow lines correspond to patterns obtained with λ=0 and having at least one complete phenol ring, while gray lines correspond to the new patterns mined with soft thresholds constraints (the violated constraints are highlighted in black).

The soft thresholds enable us to find 4 new EPs containing the phenol ring among the *top*$_{25}$ patterns (lines $17-20$), that represents a ratio of 3 (λ=20% detects 3 times more useful EPs compared to λ=0). Let us note that two of these patterns also contain benzene (lines 18 and 20). Moreover, these patterns, which violate slightly the density constraint, are highly aromatic and from a biodegradability point of view, aromatic compounds are among the most recalcitrant of the pollutants. These patterns have a high growth rate and this result strengthens our hypothesis that the growth rate measure captures toxic behavior. Furthermore, λ=20% enables to extract two new EPs containing the chlorobenzene (lines 2, 4) and one pattern containing the fragment *Clc*(*cc*)*ccc* (line 10). These patterns are of a great interest and they reinforce our previous hypothesis of toxicophore.

Table 3b depicts the *top*$_{25}$ EPs with λ=40%. As before, soft thresholds allow to discover 6 new EPs containing benzene (cf. lines 7, 9, 14, 16, 19 and 20). These patterns, which slightly violate the growth rate constraint, are highly aromatic and relatively dense and thus reinforce the hypothesis that the higher the chemical density is, the stronger its chemical behavior. A new EP of particular

Table 4. *top*$_{25}$ Jumping Emerging Patterns (λ=50% and λ=60%)

(a) *top*$_{25}$ JEP (first 11).

N	Interest	Pattern	Growth Rate	Frequency	Aromaticity	Density	SMILES
λ = 50%							# Solutions=3
1	253	24 35 87	∞	47	66	88	cc ccc OP
2	222	24 35 90	∞	45	66	77	cc ccc OPO
3	222	24 35 105	∞	45	66	77	cc ccc COP
λ = 60%							# Solutions=457
1	174	24 35 47 59 77 87	∞	40	83	66	cc ccc cccc ccccc c1ccccc1 OP
2	174	24 35 47 59 87	∞	42	80	71	cc ccc cccc ccccc OP
3	172	24 35 47 77 87	∞	40	80	71	cc ccc cccc c1ccccc1 OP
4	171	24 35 47 59 76 77 87	∞	40	85	61	cc ccc cccc ccccc cccccc c1ccccc1 OP
5	169	24 35 47 59 76 87	∞	40	83	64	cc ccc cccc ccccc cccccc OP
6	169	24 35 47 76 77 87	∞	40	83	64	cc ccc cccc cccccc c1ccccc1 OP
7	168	24 35 47 87	∞	42	75	79	cc ccc cccc OP
8	167	24 35 47 76 87	∞	40	80	69	cc ccc cccc cccccc OP
9	167	24 35 59 77 87	∞	40	80	69	cc ccc ccccc c1ccccc1 OP
10	166	24 35 59 76 77 87	∞	40	83	63	cc ccc ccccc cccccc c1ccccc1 OP
11	162	24 35 59 76 87	∞	40	80	67	cc ccc ccccc cccccc OP

(b) *top*$_{25}$ JEP (last 14).

N	Interest	Pattern	Growth Rate	Frequency	Aromaticity	Density	SMILES
12	162	24 35 76 77 87	∞	40	80	67	cc ccc cccccc c1ccccc1 OP
13	161	24 35 59 87	∞	42	75	76	cc ccc ccccc OP
14	160	24 47 59 77 87	∞	40	80	66	cc cccc ccccc c1ccccc1 OP
15	159	24 35 77 87	∞	40	83	60	cc ccc c1ccccc1
16	159	24 47 59 76 77 87	∞	40	75	76	cc cccc ccccc cccccc c1ccccc1 OP
17	157	24 35 59 76 77 87	∞	38	83	60	cc ccc ccccc cccccc c1ccccc1 OP
18	157	24 35 47 59 77 90	∞	38	83	60	cc ccc ccccc cccccc c1ccccc1 OPO
19	156	24 35 47 76 77 105	∞	38	83	59	cc ccc cccc cccccc c1ccccc1 COP
20	156	24 35 47 59 76 105	∞	38	83	59	cc ccc cccc ccccc cccccc COP
21	156	24 35 47 76 77 90	∞	38	83	59	cc ccc cccc cccccc c1ccccc1 OPO
22	156	24 35 47 59 76 90	∞	38	83	59	cc ccc cccc ccccc cccccc OPO
23	155	24 47 76 77 87	∞	40	80	64	cc cccc cccccc c1ccccc1 OP
24	155	24 47 59 76 87	∞	40	80	64	cc cccc ccccc OP
25	155	24 35 47 59 105	∞	40	80	64	cc ccc cccc ccccc COP

interest to chemists is obtained: $\{nc\}$. This pattern is environmentally hazardous since it is very toxic to aquatic species.

For the *top*$_{50}$ EPs, soft thresholds with λ=20% (resp. 40%) allow to detect 2.25 (resp. 1.25) times more solutions containing the phenol ring. Furthermore, λ=40% enables to extract 8 (resp. 3) new EPs containing benzene (resp. the chlorobenzene). All these results confirm the benefit of using soft thresholds in order to obtain novel chemical knowledge of a great interest.

5.5 Extracting Jumping Emerging Patterns

Our third experiment evaluates the character of toxicity carried by the chemical fragments which occur only in chemicals classified H400 (i.e. high toxicity), the so-called Jumping Emerging Patterns (JEPs) (cf. Definition 2). Table 4 shows the *top*$_{25}$ JEPs according to different values of λ. One can draw the following remarks: (i) Without soft threshold constraints, JEPs are not detected; (ii) With λ=50% (resp. 60%), we get 3 (resp. 457) JEPs. Indeed, this kind of patterns are less frequent, thus it is necessary to have a relatively high threshold violation; (iii) All patterns containing organo-phosphorus fragments have a growth rate equal to $+\infty$. It appears that the organo-phosphorus fragment is a generalization of several Jumping Emerging Fragments (JEFs) and can be seen as a kind of *maximum common structure* of these fragments; (iv) Among the *top*$_{25}$ JEPs extracted with λ=60%, the most interesting patterns are those including a *benzene ring* (c1ccccc1). Actually, the benzene ring is one of the most aromatic molecular fragments. With λ=50%, the extracted JEPs contain subfragments of benzene without complete rings. Thus, these JEPS are less relevant from a chemical point of view compared to those mined with λ=60%. Again, these results demonstrate the effectiveness and the contribution of soft threshold constraints to highlight relevant chemical structures, such as benzene rings compared to its subfragments.

6 Related Work

There are few works in data mining to cope with the stringent aspect of the usual constraint-based mining framework. Relaxation has been studied to provide soft constraints with specific properties in order to be able to manage them by usual constraint mining algorithms. In [5], regular expression constraints have been relaxed into anti-monotonic constraints for mining significant sequences.

In the context of local patterns, [3] have proposed a generic framework using semirings to express preferences between solutions. Each constraint has its own measure of interest and the interest of a query is the aggregation of the interests of all constraints composing the query. Given a query and a threshold value, the goal is to find all local patterns whose interest satisfies this threshold value. However, this approach relies on the following strong hypothesis: the interest of a given query satisfies the threshold, if and only if, the interest of *each* constraint satisfies the same threshold [3]. If the aggregation operator is performed using the *min* operator (*fuzzy semiring*), the equivalence holds. However, for the *sum* operator (*weighted semiring*) and the × operator (*probabilistic semiring*), it is no longer the case. That is why the authors need to perform a post-processing step to filter the set of effective solutions.

So, unlike [3], our approach preserves the equivalence without requiring a post-processing step. Moreover, it can be applied on pattern sets and therefore to local patterns.

7 Conclusion

In this paper, we have proposed a method to integrate soft threshold constraints into the pattern discovery process by using works on constraints relaxation. Then, by defining an interestingness measure on patterns, we have shown how soft threshold constraints can be exploited for extracting the top-k patterns. Finally, the relevance and the efficiency of our approach is highlighted through a case study in chemoinformatics for discovering toxicophores. Experimental results demonstrate the benefit of using soft threshold constraints in order to obtain novel chemical knowledge of great interest such as the top-k patterns or JEPs. As future work, we want to study the benefit of our approach on the clustering task [2] and skylines that return points of interest not dominated by other points with respect to a given set of criteria [4].

Acknowledgements. This work is partly supported by the ANR (French Research National Agency) funded project FiCOLOFO ANR-10-BLA-0214. The authors would like to thank the CERMN Laboratory (University of Caen, France) for providing the chemical data and for the highly valuable comments.

References

1. Bajorath, J., Auer, J.: Emerging chemical patterns: A new methodology for molecular classification and compound selection. J. of Chemical Information and Modeling 46, 2502–2514 (2006)
2. Basu, S., Davidson, I., Wagstaff, K.L.: Constrained Clustering: Advances in Algorithms, Theory, and Applications. Chapman & Hall (2008)
3. Bistarelli, S., Bonchi, F.: Soft constraint based pattern mining. Data Knowl. Eng. 62(1), 118–137 (2007)
4. Borzsonyi, S., Kossmann, D., Stocker, K.: The skyline operator. In: Proceedings of the 17th International Conference on Data Engineering (ICDE 2001), pp. 421–430. IEEE Computer Science, Springer (2001)
5. Garofalakis, M.N., Rastogi, R., Shim, K.: SPIRIT: Sequential pattern mining with regular expression constraints. The VLDB Journal, 223–234 (1999)
6. Guns, T., Nijssen, S., De Raedt, L.: Itemset mining: A constraint programming perspective. Artif. Intell. 175(12-13), 1951–1983 (2011)
7. Ke, Y., Cheng, J., Xu Yu, J.: Top-k correlative graph mining. In: SDM, pp. 1038–1049 (2009)
8. Khiari, M., Boizumault, P., Crémilleux, B.: Constraint Programming for Mining n-ary Patterns. In: Cohen, D. (ed.) CP 2010. LNCS, vol. 6308, pp. 552–567. Springer, Heidelberg (2010)
9. Mannila, H., Toivonen, H.: Levelwise search and borders of theories in knowledge discovery. Data Mining and Knowledge Discovery 1(3), 241–258 (1997)
10. Ng, R.T., Lakshmanan, V.S., Han, J., Pang, A.: Exploratory mining and pruning optimizations of constrained associations rules. In: Proceedings of ACM SIGMOD 1998, pp. 13–24. ACM (1998)
11. Kralj Novak, P., Lavrac, N., Webb, G.I.: Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. Journal of Machine Learning Research 10, 377–403 (2009)
12. Régin, J.-C., Petit, T., Bessière, C., Puget, J.-F.: An Original Constraint Based Approach for Solving over Constrained Problems. In: Dechter, R. (ed.) CP 2000. LNCS, vol. 1894, pp. 543–548. Springer, Heidelberg (2000)
13. Poezevara, G., Cuissart, B., Crémilleux, B.: Extracting and summarizing the frequent emerging graph patterns from a dataset of graphs. J. Intell. Inf. Syst. 37(3), 333–353 (2011)
14. De Raedt, L., Guns, T., Nijssen, S.: Constraint programming for itemset mining. In: KDD 2008, pp. 204–212. ACM (2008)
15. De Raedt, L., Zimmermann, A.: Constraint-based pattern set mining. In: Proceedings of the Seventh SIAM International Conference on Data Mining, Minneapolis, Minnesota, USA. SIAM (April 2007)
16. Wang, J., Han, J., Lu, Y., Tzvetkov, P.: Tfp: An efficient algorithm for mining top-k frequent closed itemsets. IEEE Trans. Knowl. Data Eng. 17(5), 652–664 (2005)