Mining Skypattern Cubes
Willy Ugarte Rojas, Patrice Boizumault, Samir Loudni, Bruno Crémilleux

To cite this version:
Willy Ugarte Rojas, Patrice Boizumault, Samir Loudni, Bruno Crémilleux. Mining Skypattern Cubes. 21st Europ. Conf. on Artificial Intelligence, (ECAI’14), Aug 2014, Prague, Czech Republic. p1-6. hal-01026285

HAL Id: hal-01026285
https://hal.archives-ouvertes.fr/hal-01026285
Submitted on 21 Jul 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Computing Skypattern Cubes

Willy Ugarte\(^1\) and Patrice Boizumault\(^1\) and Samir Loudni\(^1\) and Bruno Crémilleux\(^1\)

Abstract. We introduce skypattern cubes and propose an efficient bottom-up approach to compute them. Our approach relies on derivation rules collecting skypatterns of a parent node from its child nodes without any dominance test. Non-derivable skypatterns are computed on the fly thanks to Dynamic CSP. The bottom-up principle enables to provide a concise representation of the cube based on skypattern equivalence classes without any supplementary effort. Experiments show the effectiveness of our proposal.

1 Introduction

The notion of skyline queries [1] has been recently integrated into the pattern discovery paradigm to mine skyline patterns (henceforth called skypatterns) [10, 11]. Given a set of measures, skypatterns are patterns based on the Pareto-dominance relation for which no measure can be improved without degrading the others. As an example, a user may prefer a pattern with a high frequency, large size and a high confidence. In this example, a pattern \(x_i\) dominates another pattern \(x_j\) if \(\text{freq}(x_j) \geq \text{freq}(x_i), \text{size}(x_j) \geq \text{size}(x_i), \text{confidence}(x_j) \geq \text{confidence}(x_i)\) where at least one strict inequality holds.

Given a set of patterns, the skypattern set contains the patterns that are not dominated by any other pattern. Skypatterns are highly interesting because they do not require any threshold on the measures and the dominance relation gives to the skypatterns a global interest with semantics easily understood by the user.

In practice, users do not know the exact role of each measure which be used and it is difficult to beforehand select the most appropriate set of measures. Users would like to keep all the measures potentially useful, look what happens on a skypattern set by removing or adding a measure to evaluate the impact of measures and then converge to a proper skypattern set. Similarly to the notion of the skyline cube in the database [9], users would like to have available the skypattern cube. Each element of the cube is a node which associates to a subset of the measures its skypattern set. By comparing two neighboring nodes, which are differentiated by adding or removing one measure, users can observe the new skypatterns and the ones which die out. It greatly helps to better understand the role of the measures. Moreover, users can spot that different subsets of measures have the same skypattern set: such an equivalence class over subsets of measures shows useless measures (i.e., measures that can be added to a set of measures without changing the skypattern set).

To sum up, the cube is the proper structure to enable various user queries in an efficient manner and to discover the most interesting skypattern sets.

More formally, given a set \(M\) of \(n\) measures, the \(2^n-1\) possible non-empty skypattern subsets should be precomputed to efficiently handle various queries of users. A baseline method to build the skypattern cube needs the computing of the skypatterns on every measure subset and incurs a prohibitive cost. Therefore the problem of efficient computing of the skypattern cube is the focus of this paper.

For computing the skypattern cube, we propose a bottom-up approach motivated by the following observations. First, we formally give two derivation rules providing an easy way to automatically infer a large proportion of the skypatterns of a parent node from the skypattern sets of its child nodes without any dominance test (if \(k\) measures are associated to a parent node, its child nodes are the nodes defined by the \((k-1)\) subsets of \(k-1\) measures). For the new skypatterns of a parent node (i.e., skypatterns which are not skypatterns in its child nodes), we give an efficient technique based on dynamic CSP to mine them on the fly. We show that in practice the number of new skypatterns remains low. Second, we demonstrate how the bottom-up principle enables us to determine skypattern equivalence classes without any supplementary effort. This result has the advantage to provide a more concise cube, highlighting the measures giving the same skypattern set. Third, experiments conducted on real-life datasets show the practical effectiveness achieved by our formal results. To sum up, to the best of our knowledge, we designed the first method to build the skypattern cube without enumerating and mining all the possible skypatterns.

This paper is organized as follows. After introducing the background in Section 2, we present in Section 3 the formal properties to automatically infer skypatterns and build the concise representation of the cube. Section 4 describes our CSP method to mine the new skypatterns. We discuss related work in Section 5. Section 6 presents the experiments and we conclude in Section 7.

2 Skypattern Cube

2.1 Context and 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 \(\mathcal{L}_T = 2^\mathcal{I}\setminus\emptyset\). A transactional dataset \(T\) is a multiset of patterns of \(\mathcal{L}_T\). Fig. 1a depicts a transactional dataset \(T\) where each transaction (or pattern) \(t_i\) is described by items denoted \(A, \ldots, F\). The traditional example is a supermarket database in which each transaction corresponds to a customer and every item in the transaction is a product bought by the customer. An attribute (price) is associated to each product (see Fig. 1a).

Constraint-based pattern mining aims at extracting all patterns \(x\) of \(\mathcal{L}_T\) satisfying a query \(q(x)\) (conjunction of constraints) which is usually called theory [5]: \(\text{Th}(q) = \{x \in \mathcal{L}_T \ | \ q(x)\ \text{is true}\}\). A common example is the frequency measure leading to the minimal frequency constraint \((\text{freq}(x) \geq \theta)\). The latter provides patterns \(x\) having a number of occurrences in the dataset exceeding a given minimal threshold \(\theta\). There are other usual measures for a pattern \(x\):
Lattice associated to any skypattern. The other part is called the dominance area. The shaded area is called the forbidden area since it cannot contain \( w .r.t. x \). Figure 1a: \( M = \{ \text{freq}, \text{area} \} \).

- area \( x \) = \( \text{freq}(x) \times \text{size}(x) \).
- min (\( x.\text{att} \)) (resp. max (\( x.\text{att} \))) is the smallest (resp. highest) value of the set of item values of \( x \) for attribute \( \text{att} \).
- mean \( x \) = \(( \text{min}(x.\text{att}) + \text{max}(x.\text{att}))/2 \).

**Example 1** For the dataset in Fig. 1a, \( \text{freq}(BC) = 5 \), area \( (BC) = 10 \) and \( \text{mean}(\text{BCD.price}) = 25 \).

A pattern \( x_i \) is closed w.r.t. a measure \( m \) iff \( \forall x_j \supseteq x_i, m(x_j) \not\geq m(x_i) \). The set of closed patterns is a compact representation of the patterns (i.e. we can derive all the patterns with their exact value for \( m \) from the closed ones). This definition is straightforwardly extended to a set of measures \( M \).

### 2.2 Skypatterns

As stated above, skypatterns enable to express a user-preference point of view according to a dominance relation [10].

**Definition 1 (Pareto Dominance)** Given a set of measures \( M \), a pattern \( x \) dominates another pattern \( x_j \) w.r.t. \( M \) (denoted by \( x_i \succ_r x_j \)) iff \( \exists m \in M, m(x_i) \geq m(x_j) \) and \( \exists m \in M, m(x_i) > m(x_j) \).

**Definition 2 (Skypattern and skypattern operator)** Given a set of measures \( M \), a skypattern w.r.t. \( M \) is a pattern not dominated w.r.t. \( M \). The skypattern operator \( \text{Sky}(M) \) returns all the skypatterns w.r.t. \( M \): \( \text{Sky}(M) = \{ x_i \in L_2 \mid \exists x_j \in L_2, x_j \succ_r M x_i \} \).

**Example 2** From \( T \) and with \( M = \{ \text{freq}, \text{area} \} \), BCD dominates BC as \( \text{freq}(BCD) = \text{freq}(BC) = 5 \) and area \( (BCD) > \text{area}(BC) \) (cf. Fig. 1a). Fig. 1b provides a graphical representation of \( \text{Sky}(M) = \{ \text{BCDE}, \text{BCD}, \text{B}, \text{E} \} \). The shaded area is called the forbidden area since it cannot contain any skypattern. The other part is called the dominance area.

Let \( M \) be a set of measures. Two patterns \( x_i \) and \( x_j \) are indistinct w.r.t. \( M \) (denoted by \( x_i =_M x_j \)) iff \( \forall m \in M, m(x_i) = m(x_j) \). Two patterns \( x_i \) and \( x_j \) are incomparable w.r.t. \( M \) (denoted by \( x_i \triangleleft_M x_j \)) iff \( x_i \not> M x_j \) and \( x_j \not> M x_i \) and \( x_i \not= M x_j \).

**Definition 3 (Incomparable Skypattern)** A pattern \( x \in \text{Sky}(M) \) is incomparable w.r.t. \( M \) iff \( \forall x_i \in \text{Sky}(M) s.t. x_i \not> M x \).

**Definition 4 (Indistinct Skypattern)** A pattern \( x \in \text{Sky}(M) \) is indistinct w.r.t. \( M \) iff \( \exists x_i \in \text{Sky}(M) s.t. (x_i \not> M x \) and \( (x_i \not= M x) \).

Incomparable skypatterns and indistinct ones w.r.t. \( M \) constitute a partition of \( \text{Sky}(M) \). Moreover, \( =_M \) is an equivalence relation (i.e., the relation is reflexive, symmetric and transitive). So, indistinct skypatterns can be gathered into a group. This remark will be precious for the derivation rule on indistinct skypatterns.

**Definition 5 (Indistinct Skypattern Group (ISG))** \( S \subseteq \text{Sky}(M) \) is an indistinct skypattern group w.r.t. \( M \) iff \( |S| \geq 2 \) and \( \forall x_i, x_j \in S, (x_i =_M x_j) \) and \( \forall x_i, x_j \in \text{Sky}(M) \setminus S, (x_i \not= M x_j) \).

**Example 3** For \( M = \{ \text{freq}, \text{area} \} \), BCD, BCD and BCD are incomparable. B and E are indistinct and belong to the same ISG.

### 2.3 Skypattern Cube

Let \( M \) be a set of measures. We define the skypattern cube over \( M \) which consists of the \( 2^{|M|} - 1 \) skypatterns \( \text{Sky}(M) \) on all possible non-empty subset \( M_a \subseteq M \).

**Definition 6 (Skypattern Cube)** Let \( M \) be a set of measures. \( \text{Sky} \text{Cube}(M) = \{ (M_a, \text{Sky}(M_a)) \mid M_a \subseteq M, M_a \neq \emptyset \} \).

**Example 4** Consider the dataset in Fig. 1a. Fig. 1c depicts the lattice associated to \( M \). Fig. 1d associates to each non-empty subset of \( M \) its skypattern set.

### 3 Derivation Rules and Concise Representation

This section presents our bottom-up approach for computing the skypattern cube. The key idea is to collect the skypatterns of a parent node from the skypatterns of its child nodes. Then we compute the missing skypatterns of the node (i.e., skypatterns that are not skypatterns in its child nodes and thus not yet mined). We show how a concise representation of the cube is straightforwardly provided.
3.1 Derivation Rules

Theorems 1 states that all the incomparable skypatterns of a child node remain incomparable skypatterns in its parent nodes. Theorem 2 exhibits the indistinct skypatterns of a child node that remain skypatterns in its parent nodes. These two theorems define two derivation rules that enable to derive a subset of skypatterns of a parent node.

Theorem 1 (Incomparability Rule) Let $M_u \subseteq M$. If $x$ is an incomparable skypattern w.r.t. $M_u$, then $\forall m \in M\setminus M_u$, $x \in Sky(M_u \cup \{m\})$. Moreover $x$ is incomparable w.r.t. $M_u \cup \{m\}$.

Proof (By Contradiction) Assume that $x$ is an incomparable skypattern w.r.t. $M_u$, and $\exists m \in M\setminus M_u$, s.t. $x \notin Sky(M_u \cup \{m\})$. So, $\exists y \neq x \in L_x$, $(y \succ M_u \cup \{m\})$ i.e. (1) $\forall m_i \in M_u \cup \{m\}, m_i(y) \geq m_i(x)$ and (2) $\exists m_j \in M_u \cup \{m\}, m_j(y) > m_j(x)$. For (2), there are two cases:

1. $(m_j=m)$. As $x \in Sky(M_u), \forall m_i \in M_u, m_i(x) \geq m_i(y)$.
   From (1), we deduce: $\forall m_i \in M_u, m_i(x) = m_i(y)$. So $x$ is indistinct w.r.t. $M_u$. It contradicts that $x$ is incomparable w.r.t. $M_u$.

2. $(m_j \neq m)$. As $x \in Sky(M_u), \forall m_i \in M_u, m_i(y) \geq m_i(x)$.
   From (1), we have $\forall m_i \in M_u, m_i(y) \geq m_i(x)$.
   As, $m_j(y) > m_j(x)$, we deduce that $y \succ M_u x$. It contradicts that $x$ is a skypattern w.r.t. $M_u$.

Example 5 Let $M_u =\{m_1, m_3\}$. BCDE and BCD are incomparable w.r.t. $M_u$. Theorem 1 enables to deduce that BCDE and BCD belong to $Sky(M_u \cup \{m\})$ and $Sky(M_u \cup \{m_1\})$.

Theorem 2 (ISG Rule) Let $M_u \subseteq M$ and $S$ an ISG w.r.t. $M_u$. $\forall m \in M\setminus M_u$, each skypattern $x \in S$ s.t. $m(x) = \max_{i \in S} m_i(x_i)$ is a skypattern w.r.t. $M_u \cup \{m\}$.

Proof (By Contradiction) Assume that there exists an ISG $S$ w.r.t. $M_u$, s.t. $\exists m \in M\setminus M_u, s.t. \exists x \in S$ s.t. $m(x) = \max_{i \in S} m_i(x_i)$ and $x \notin Sky(M_u \cup \{m\})$. So, $\exists y \neq x \in L_x$, $(y \succ M_u \cup \{m\})$ i.e. (1) $\forall m_i \in M_u \cup \{m\}, m_i(y) \geq m_i(x)$ and (2) $\exists m_j \in M_u \cup \{m\}, m_j(y) > m_j(x)$. For (2), there are two cases:

1. $(m_j=m)$. As $x \in Sky(M_u), \forall m_i \in M_u, m_i(x) = m_i(y)$. i.e. $y \in S$.
   So, $m_i(y) \leq m_i(x)$ (as $m(x) = \max_{i \in S} m_i(x_i)$). It contradicts (2).

2. $(m_j \neq m)$. As $x \in Sky(M_u), \forall m_i \in M_u, m_i(y) \geq m_i(x)$.
   From (1), we have $\forall m_i \in M_u, m_i(y) \geq m_i(x)$.
   As, $m_j(y) > m_j(x)$, we deduce that $y \succ M_u x$. It contradicts that $x$ is a skypattern w.r.t. $M_u$.

Example 6 $S = \{B, E\}$ is an ISG w.r.t. $\{m_1\}$. Theorem 2 enables to deduce that:

- $E \in Sky(\{m_1, m_2\})$ since $m_2(E) = \max_{x \in S} m_2(x_i)$
- $E \in Sky(\{m_1, m_4\})$ since $m_4(E) = \max_{x \in S} m_4(x_i)$
- $B \in Sky(\{m_1, m_3\})$ since $m_3(\{B\}=m_3(E)=\max_{x \in S} \{m_3(x_i)\}$

Corollary 1 Let $S$ an ISG w.r.t. $M_u$, and $m \in M\setminus M_u$. Let $S' = \{x | x \in S \setminus m(x) = \max_{x \in S} m(x_i)\}$. If $S'$ is a singleton then the unique skypattern is incomparable w.r.t. $M_u \cup \{m\}$ else $S'$ is an ISG w.r.t. $M_u \cup \{m\}$. Finally $x \in S \setminus S'$ are not skypatterns for $M_u \cup \{m\}$.

Then the missing skypatterns (i.e., non-derivable skypatterns) are computed. We start by defining the derivable skypatterns.

Let $M_u \subseteq M$ and $m \in M\setminus M_u$. We define $inc(M_u)$ the set of incomparable skypatterns w.r.t $M_u$ and $ind(M_u, m)$ the set of maximal indistinct skypatterns w.r.t a measure $m$:

- $inc(M_u) = \{x \in Sky(M_u) | x \text{ is incomparable w.r.t. } M_u\}$
- $ind(M_u, m) = \bigcup_{S \subseteq Sky(M_u)} \{x | x \in S \in \{m(x) = \max_{x \in S} m(x_i)\}\}$

$derived(M_u)$ is the set of skypatterns of the node associated to $M_u$ which can be derived from the skypatterns of its child nodes:

- $derived(M_u) = \bigcup_{m \in M_u} (inc(M_u \setminus \{m\}) \cup ind(M_u \setminus \{m\}, m))$.

First, it is obvious that: $derived(M_u) \subseteq Sky(M_u)$ (see Theorem 1 and 2). Experiments show that a large proportion of skypatterns are obtained by this way. Moreover, if a skypattern $x$ in a parent node is also a skypattern in at least one of its child nodes, then $x$ will be necessarily collected by one of these rules. It is expressed by:

$derived(M_u) = \bigcup_{m \in M_u} Sky(M_u \setminus \{m\}) \cap Sky(M_u)$.

This property shows the power of our derivation rules. (The proof is immediate since, for each node, incomparable and indistinct skypatterns constitute a partition.) However, $derived(M_u)$ can be strictly included in $Sky(M_u)$, i.e., some skypatterns are missing. It happens when a skypattern of a node is not a skypattern in any of its child nodes as illustrated by the following example.

Example 7 As $BCDE$ is incomparable w.r.t. $\{m_3\}$, $BCDE \in Sky(\{m_1, m_3\})$. As $B$ and $E$ constitute an ISG w.r.t. $\{m_3\}$ and $m_3(B)=m_3(E)$, then $B, E \in Sky(\{m_1, m_3\})$. But, the derivation rules cannot deduce that $BCDE \in Sky(\{m_1, m_3\})$.

We compute on the fly the non-derivable skypatterns thanks to a dynamic CSP method described in Section 4. Moreover, we can go further by detecting a priori that $derived(M_u) = Sky(M_u)$ for some $M_u$ and thus avoiding useless computation. Theorem 3 states a sufficient condition ensuring that $derived(M_u) = Sky(M_u)$. Experiments show that this condition is effective in practice.

Theorem 3 (Non-Computing Sufficient Condition) Let $M_u \subseteq M$. If $\forall m \in M_u s.t. \min_{x \in derived(M_u)} m(x) = \max_{x \in derived(M_u)} m(x)$ then $Sky(M_u) = derived(M_u)$.

Proof (By Contradiction) Let $m \in M_u$ a measure s.t. $\min_{x \in derived(M_u)} m(x) = \max_{x \in derived(M_u)} m(x)$. Assume that $\exists p \in Sky(M_u)$ \ derived(M_u), so $m(p) = \min_{x \in derived(M_u)} m(x)$. Henceforth $m(p) = \max_{x \in derived(M_u)} m(x)$. Thus, $p \notin Sky(M_u)$, and:

(i) either $p$ is incomparable w.r.t. $\{m\}$, and
(ii) or $p$ is indistinct w.r.t. $\{m\}$ with maximal value for $m$.

From (i) and (ii), $p \notin derived(M_u)$ leading to a contradiction.

3.3 Concise Representation of a Cube

Different subsets of measures may lead to a same set of skypatterns. This observation can be used to provide a concise representation of the cube without loss of information. We define an equivalence relation over subsets of measures having the same skypattern set:

3.3.1 Computing a Skypattern Cube

The skypatterns of a node are computed in two steps. First, we collect all the skypatterns which can be derived from its child nodes.
Definition 7 (Equivalence between sets of measures) Let $M_a$ and $M_b$ two sets of measures. $M_a$ and $M_b$ are said to be equivalent iff $Sky(M_a) = Sky(M_b)$.

Example 8 Equivalence classes on our running example are illustrated in Figure 1c. There are 8 classes: 4 have a cardinality of 1, 3 have a cardinality of 2 and 1 has a cardinality of 5.

Theorem 4 (Equivalence class) Let $M_a \subseteq M$ and $m \in M \setminus M_a$. $Sky(M_a \cup \{m\}) = Sky(M_a)$ iff (1) all indistinct skypatterns w.r.t. $M_a$ are indistinct skypatterns w.r.t. $M_a \cup \{m\}$ and (2) $Sky(M_a \cup \{m\}) = derived(M_a \cup \{m\})$.

Proof (Double inclusion) All incomparable w.r.t. $M_a$ are incomparable w.r.t. $M_a \cup \{m\}$ (Theorem 1). All indistinct w.r.t. $M_a$ are indistinct w.r.t. $M_a \cup \{m\}$ according to (1). Since incomparable w.r.t. $M_a$ and indistinct w.r.t. $M_a$ form a partition of $Sky(M_a)$, $Sky(M_a) \cup \{m\} = derived(M_a \cup \{m\})$. As derived skypatterns w.r.t. $M_a \cup \{m\}$ can only come from $Sky(M_a)$, $Sky(M_a \cup \{m\}) \subseteq Sky(M_a)$.

4 Mining non-derivable Skypatterns using DCSP

This section describes how the non-derivable skypatterns can be mined using Dynamic CSP [12]. The main idea of our approach [11], taking benefit from cross-fertilization between CSP and data mining [3, 4], is to improve the mining step during the process thanks to constraints dynamically posted and stemming from the current set of the candidate skypatterns. The process stops when the forbidden area cannot be enlarged. Finally, the completeness of our approach is insured by the completeness of the CP solver.

4.1 Mining Skypatterns

A Constraint Satisfaction Problem (CSP) $P=(X, D, C)$ is defined by a set of variables $X$, a domain $D$, which maps every variable $x_i \in X$ to a finite set of values $D(x_i)$, and a set of constraints $C$.

A Dynamic CSP [12] is a sequence $P_1, P_2, ..., P_n$ of CSP, each one resulting from some changes in the definition of the previous one. These changes may affect every component in the problem definition: variables, domains and constraints. For our approach, changes are only performed by adding new constraints. Solving such dynamic CSP involves solving a single CSP with additional constraints posted during search. Each time a new solution is found, new constraints are imposed. Such constraints will survive backtracking and state that next solutions should verify both the current set of constraints and the added ones.

Constraints on the dominance relation are dynamically posted during the mining process. Variable $x$ will denote the (unknown) skypattern we are looking for. Changes are only performed by adding new constraints. So, we consider the sequence $P_1, P_2, ..., P_n$ of CSP where $M$ is a set of measures, each $P_i = ((x), E_T, q_i(x))$ and:

- $q_1(x) = closed_M(x)$
- $q_{i+1}(x) = q_i(x) \land \phi_i(x)$ where $s_i$ is the first solution to $q_i(x)$

First, the constraint $closed_M(x)$ states that $x$ must be a closed pattern w.r.t. $M$, it allows to reduce the number of redundant patterns (see Section 2.1). Then, the constraint $\phi_i(x) \equiv \neg(s_i \succ_M x)$ states that the next solution (which is searched) will not be dominated by $s_i$. Using a short induction proof, we can easily argue that query $q_{i+1}(x)$ looks for a pattern $x$ that will not be dominated by any of the patterns $s_1, s_2, ..., s_i$. Each time the first solution $s_i$ to query $q_i(x)$ is found, a new constraint $\phi_i(x)$ is dynamically posted, leading to reduce the search space. This process stops when the forbidden area cannot be further extended (i.e. there exists $n$ s.t. query $q_{i+1}(x)$ has no solution). For skypatterns, $\phi_i(x)$ states that $\neg(s_i \succ_M x)$.

$\phi_i(x) \equiv (\bigvee_{m \in M} m(s_i) < m(x)) \lor (\bigwedge_{m \in M} m(s_i) = m(x))$

But, the $n$ extracted patterns $s_1, s_2, ..., s_n$ are not necessarily all skypatterns. Some of them can be only “intermediate” patterns simply used to enlarge the forbidden area. A post processing step must be performed to filter all candidate patterns $s_i$ that are not skypatterns, i.e. for which there exists $s_j (1 \leq i < j \leq n)$ s.t. $s_j$ dominates $s_i$. So mining skypatterns is achieved in a two-steps approach:

1. Compute the set $S = \{s_1, s_2, ..., s_n\}$ of candidates using Dynamic CSP.
2. Remove all patterns $s_i \in S$ that are not skypatterns.

While the number of candidates ($n$) could be very large, it remains reasonably-sized in practice (see [11]).

4.2 Mining the non-derivable Skypatterns

In order to find the non-derivable skypatterns, we proceed in the same way as Section 4.1, by stating that any non-derivable skypattern could not be dominated by any derived skypattern.

Let $M_a \subseteq M$ and $derived(M_a)$ the subset of $Sky(M_a)$ obtained using the two derivation rules. Consider the sequence $P_1, P_2, ..., P_n$ of CSP where each $P_i = ((x), E, q_i(x))$ and:

- $q_1(x) = closed_{M_a}(x) \land \Psi_{M_a}(x)$
- $q_{i+1}(x) = q_i(x) \land \neg(s_i \succ_{M_a} x)$ where $s_i$ is the first solution to query $q_i(x)$
- $\Psi_{M_a}(x)$ states that $x$ cannot be dominated w.r.t. $M_a$ by any derived skypattern:

$$\Psi_{M_a}(x) = \bigwedge_{x \in derived(M_a)} \neg(x \succ_{M_a} x)$$

Example 9 Consider example 7. For $M_a=\{m_1, m_3\}$, $derived(M_a) = \{B, E, BCD, DE\}$, so $\Psi_{M_a}(x) = \neg(B \succ_{M_a} x) \land \neg(E \succ_{M_a} x) \land \neg(BCD \succ_{M_a} x)$. The associated Dynamic CSP has a unique solution: $x = BCD$. 

\begin{algorithm}
\caption{Bottom-up approach for computing the cube}
\begin{algorithmic}[1]
\State Input: $M$: a set of measures, $T$: a dataset.
\State Output: The skypattern cube of dataset $T$ w.r.t. $M$.
\State\textbf{cube} $\leftarrow \emptyset$;
\For{$m \in M$}
\State $\textbf{cube} \leftarrow \textbf{cube} \cup \{\{m\}, Sky\{\{m\}\}\}$;
\EndFor
\For{$i \leftarrow 2$ to $|M|$}
\For{$M_a \in M$ s.t. $|M_a| = i$}
\State $\textbf{cube} \leftarrow \textbf{cube} \cup \{\{M_a\}, complete(derived(M_a))\}$;
\EndFor
\EndFor
\State\textbf{return cube}
\end{algorithmic}
\end{algorithm}
5 Related Work

Mining skypatterns is different from mining skylines [1]. Skyline queries focus on the extraction of tuples of the dataset and assume that all skylines are in the dataset. The skypattern mining task consists in extracting patterns which are elements of the frontier defined by the given measures. The skypattern problem is clearly harder because the search space for skypatterns is much larger than the search space for skylines: \(O(2^{|T|})\) instead of \(O(|T|)\) for skylines.

Two methods have been designed for mining skypatterns. Aetheris [10] takes benefit of theoretical relationships between pattern condensed representations and skypatterns. Aetheris proceeds in two steps: first, condensed representations of the whole set of patterns (i.e. closed patterns according to the considered set of measures) are extracted; then, the sky operator (see Definition 2) is applied. CP+SKY [11] mines skypatterns using Dynamic CSP (see Section 4.1). Both methods have the same efficiency, but CP+SKY also allows to mine soft skypatterns [11].

Computing the skyline cube efficiently, [7, 8, 13] proposed several strategies to share skyline computation in different subspaces, but they have to cope with the problem of enumerating skylines over all possible subspaces. [6] proposed Stellar, which computes seed skylines groups in the full space, then extend them to build the final set of skyline groups and thus avoiding the computation of skylines in all the subspaces. But Stellar does not take profit from any parent-child relationships in the lattice. [9] is able to decrease the number of domination tests by reducing the number of measure subspaces that needs to be searched. However, its complex strategy makes impossible the success of the full use of the parent-child relationships. Moreover, all of these techniques address skylines.

6 Experimental Evaluation

6.1 Skypattern Cubes for Mutagenicity Dataset

In this section, we report an experimental evaluation on a real-life dataset of large size extracted from mutagenicity data [2] (a major problem in risk assessment of chemicals). This dataset has \(|T|=6,512\) transactions encoding chemicals and \(|I|=1,073\) items encoding frequent closed subgraphs previously extracted from \(T\) with a 2% relative frequency threshold. Chemists use up to \(|M|=11\) measures, five of them are typically used in contrast mining (frequency and growth rate) and enable to express different and express the parent-child relationships. The other six measures are related to topological, geometrical and chemical properties of the chemicals.

![Figure 2: Comparing CPU-times for the 3 methods.](Image)

| \(|M|\) | \(|E|\) | \(|K|\) | \(|T|\) | \(|I|\) | \(|M|\) | Speed-up |
|-------|-------|-------|-------|-------|-------|---------|
| 1     | 1     | 100   | 338   | 338   | 1.00  | 1.00    |
| 2     | 2.07  | 680   | 753   | 530   | 1.18  | 1.19    |
| 3     | 3.75  | 1,036 | 1,280 | 699   | 2.00  | 2.00    |
| 4     | 0.64  | 1,421 | 1,993 | 1,621 | 0.80  | 0.80    |
| 5     | 0.53  | 1,865 | 2,982 | 1,062 | 0.62  | 0.62    |
| 6     | 0.44  | 2,424 | 4,526 | 2,532 | 0.56  | 0.56    |
| 7     | 0.36  | 3,200 | 7,146 | 3,184 | 0.45  | 0.45    |
| 8     | 0.29  | 4,690 | 12,015| 4,150 | 0.36  | 0.36    |
| 9     | 0.23  | 6,327 | 21,773| 6,124 | 0.30  | 0.30    |
| 10    | 0.20  | 9,619 | 42,386| 7,778 | 0.25  | 0.25    |
| 11    | 0.20  | 15,263| 87,374|11,578 |0.17  |0.17    |

(2) # of equivalence classes
(3) \(|M|\)
(4) \(|K|\)
(5) \(|T|\)
(6) \(|I|\)
(7) \(|M|\)
(8) \(|T|\)
(9) \(|I|\)
(10) \(|M|\)
(11) \(|T|\)

Table 1: Space analysis (left part) and CPU-time analysis (right part).

Experimental protocol. The implementation of CP+SKY+CUBE (bottom-up approach proposed in this paper) was carried out in

A chemical \(Ch\) contains an item \(A\) if \(Ch\) supports \(A\), and \(A\) is a frequent subgraph of \(T\).
skypatterns) at each level of a cube. Reported values in Fig. 3 are average values over all 11 possible cubes of 10 measures. For each level \( i \) (\( 2 \leq i \leq 10 \)), the proportion of incomparable and indistinct skypatterns is also depicted.

Our derivation rules are very efficient since they are able to deduce about 80 – 90% of the skypatterns, except for the first levels. Moreover, when the number of measures increases, the number of indistinct skypatterns decreases (in percentage), while the number of incomparable skypatterns increases. Indeed, incomparable skypatterns of child nodes remain incomparable for a parent node (see Theorem 1) whereas indistinct skypatterns may become any kind of skypatterns or dominated patterns (see Corollary 1).

(d) **Effectiveness of our sufficient condition.** Theorem 3 gives our sufficient condition stating if, for a subset of measures \( M_u \), \( \text{derived}(M_u) = \text{Sky}(M_u) \) without requiring any Dynamic CSP.

To asses the effectiveness of this condition, we measured the percentage of success at each level of a cube. Reported values in Table 2 are average values over all 11 possible cubes of 10 measures. Line 1 provides the level in the lattice. Line 2 indicates the number of nodes where our condition applies, while Line 3 reports the number of nodes where our condition should apply. Line 4 depicts their ratio (percentage of success). The more the number of measures increases, the more our sufficient condition becomes effective. From level 5 to level 10, the percentage of success increases from 73% to 100%.

![Figure 3: Effectiveness of our derivation rules.](http://www.ics.uci.edu/~mlearn/MLRepository.html)

### 6.2 Skypattern Cubes for UCI Datasets

Experiments were carried out on 14 various datasets from UCI benchmarks. We considered 5 measures \( M = \{ \text{freq}, \text{max}, \text{area}, \text{mean}, \text{growth-rate} \} \). Measures using numeric values, like \( \text{mean} \), were applied on attribute values that were randomly generated within the range \([0,1]\). Table 3 summarizes the results we obtained.

(a) **CPU-time analysis.** Columns 2-4 compare the CPU-times of the three methods. \( \text{CP+SKY+CUBE} \) clearly dominates the two base-line methods. On half of the datasets, the speed-up is at least 10.43 (see column 5).

(b) **Effectiveness of our derivation rules.** Column 7 reports, for each dataset, the percentage of derived skypatterns. Reported values are the average values over levels \( i \) (\( 2 \leq i \leq 5 \)) of the cube. Our derivation rules are able to deduce about 79 –97% of the skypatterns, except for two datasets (lymph and vehicle).

### 7 Conclusion

We have designed an efficient bottom-up method to compute skypattern cubes. Our derivation rules are able to collect a large part of the skypatterns of a parent node. Non-derivable skypatterns are computed on the fly thanks to Dynamic CSP. The bottom-up strategy makes easy the building of a concise representation of the cube according to skypattern equivalence classes. Experiments show the effectiveness of our proposal. Navigation through the cube is a highly promising perspective.

**Acknowledgments.** This work is partly supported by the ANR (French Research National Agency) funded projects FiCoLoFo ANR-10-BLA-0214 and Hybride ANR-11-BS002-002.

**REFERENCES**


### Table 3: Results on UCI datasets with \( |M| = 5 \)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CPU-Time (1)</th>
<th>Speed-Up (2)</th>
<th>(3) Succ. %</th>
<th>(4) Base-line-CP+SKY</th>
<th>(5) CP+SKY+CUBE</th>
<th>(6) CP+SKY+CUBE/CP+SKY</th>
</tr>
</thead>
<tbody>
<tr>
<td>austral</td>
<td>6m04s</td>
<td>4m15s</td>
<td>1m31s</td>
<td>3.98</td>
<td>2.79</td>
<td>0.82</td>
</tr>
<tr>
<td>cleve</td>
<td>1m55s</td>
<td>1m21s</td>
<td>21s</td>
<td>5.27</td>
<td>3.76</td>
<td>0.97</td>
</tr>
<tr>
<td>cmc</td>
<td>26s</td>
<td>2m23s</td>
<td>22s</td>
<td>1.20</td>
<td>6.41</td>
<td>0.90</td>
</tr>
<tr>
<td>ecrx</td>
<td>8m40s</td>
<td>5m37s</td>
<td>1m13s</td>
<td>7.12</td>
<td>4.63</td>
<td>0.89</td>
</tr>
<tr>
<td>german</td>
<td>2h34m18s</td>
<td>5m29s</td>
<td>1m40s</td>
<td>10.98</td>
<td>3.80</td>
<td>0.88</td>
</tr>
<tr>
<td>heart</td>
<td>1m46s</td>
<td>58s</td>
<td>19s</td>
<td>5.49</td>
<td>3.01</td>
<td>0.86</td>
</tr>
<tr>
<td>hepatic</td>
<td>6m12s</td>
<td>58s</td>
<td>19s</td>
<td>18.91</td>
<td>2.97</td>
<td>0.71</td>
</tr>
<tr>
<td>horse</td>
<td>10m34s</td>
<td>3m32s</td>
<td>58s</td>
<td>10.93</td>
<td>3.67</td>
<td>0.82</td>
</tr>
<tr>
<td>hypo</td>
<td>6h13m57s</td>
<td>51m46s</td>
<td>4m41s</td>
<td>79.75</td>
<td>11.04</td>
<td>0.79</td>
</tr>
<tr>
<td>lymph</td>
<td>4m32s</td>
<td>49s</td>
<td>11s</td>
<td>23.87</td>
<td>4.38</td>
<td>0.65</td>
</tr>
<tr>
<td>tic-tac-toe</td>
<td>1m10s</td>
<td>2m48s</td>
<td>41s</td>
<td>1.68</td>
<td>4.03</td>
<td>0.84</td>
</tr>
<tr>
<td>vehicle</td>
<td>34m01s</td>
<td>16m41s</td>
<td>2m55s</td>
<td>11.64</td>
<td>5.71</td>
<td>0.66</td>
</tr>
<tr>
<td>wine</td>
<td>1m00s</td>
<td>31s</td>
<td>13s</td>
<td>4.63</td>
<td>2.43</td>
<td>0.94</td>
</tr>
<tr>
<td>zoo</td>
<td>19s</td>
<td>8s</td>
<td>1s</td>
<td>10.43</td>
<td>4.82</td>
<td>0.87</td>
</tr>
</tbody>
</table>

### Table 2: Effectiveness of our sufficient condition (Theorem 3)

<table>
<thead>
<tr>
<th>Level</th>
<th># of nodes where Th 3 applies</th>
<th># of nodes where Th 3 should apply</th>
<th>Succ. Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.85</td>
<td>10.62</td>
<td>0.82</td>
</tr>
<tr>
<td>2</td>
<td>9.85</td>
<td>10.62</td>
<td>0.90</td>
</tr>
<tr>
<td>3</td>
<td>10.00</td>
<td>10.62</td>
<td>1.00</td>
</tr>
<tr>
<td>4</td>
<td>10.00</td>
<td>10.62</td>
<td>1.00</td>
</tr>
<tr>
<td>5</td>
<td>10.00</td>
<td>10.62</td>
<td>1.00</td>
</tr>
<tr>
<td>6</td>
<td>10.00</td>
<td>10.62</td>
<td>1.00</td>
</tr>
<tr>
<td>7</td>
<td>10.00</td>
<td>10.62</td>
<td>1.00</td>
</tr>
<tr>
<td>8</td>
<td>10.00</td>
<td>10.62</td>
<td>1.00</td>
</tr>
<tr>
<td>9</td>
<td>10.00</td>
<td>10.62</td>
<td>1.00</td>
</tr>
<tr>
<td>10</td>
<td>10.00</td>
<td>10.62</td>
<td>1.00</td>
</tr>
</tbody>
</table>

3 http://www.ics.uci.edu/~mlearn/MLRepository.html