Threshold-free Pattern Mining Meets Multi-Objective Optimization: Application to Association Rules

Charles Vernerey[1] , **Samir Loudni**[1] , **Noureddine Aribi**[2] and **Yahia Lebbah**[2]
[1]TASC (LS2N-CNRS), IMT Atlantique, 44307 Nantes, France
[2]Université Oran1, Lab. LITIO, 31000 Oran, Algeria
{samir.loudni, charles.vernerey}@imt-atlantique.fr, {ylebbah, aribi.noureddine}@gmail.com

Abstract

Constraint-based pattern mining is at the core of numerous data mining tasks. Unfortunately, thresholds which are involved in these constraints cannot be easily chosen. This paper investigates a Multi-objective Optimization approach where several (often conflicting) functions need to be optimized at the same time. We introduce a new model for efficiently mining Pareto optimal patterns with constraint programming. Our model exploits condensed pattern representations to reduce the mining effort. To this end, we design a new global constraint for ensuring the closedness of patterns over a set of measures. We show how our approach can be applied to derive high-quality non redundant association rules without the use of thresholds whose added-value is studied on both UCI datasets and a case study related to the analysis of genes expression data integrating multiple external genes annotations.

1 Introduction

Constraint-based pattern mining is a fundamental data mining task, extracting locally interesting patterns to be either interpreted directly by domain experts, or to be used as descriptors in downstream tasks, such as classification or clustering. Since the publication of the seminal paper [Agrawal and Srikant, 1994], two problems have limited the usability of this approach: 1) how to fix thresholds which are involved in many of these constraints, and 2) how to deal with the large result sets that often number in the thousands or even millions of patterns. Translating user preferences into *top-k* constraints [Wang *et al.*, 2005] w.r.t. to some quality measures presents a serious drawback as the choice of k is not obvious. Moreover, combining several measures into a single function is difficult. Post-processing results via condensed representations [Pasquier *et al.*, 1999] still typically leaves many patterns, while pattern set mining [De Raedt and Zimmermann, 2007] just pushes the problem further down the line.

A decade ago, several approaches have promoted the discovery of more complex interactions between patterns. An example of such interactions can be found in multi-objective optimization (MOO) [Figueira *et al.*, 2005] where several (often conflicting) objectives need to be simultaneously optimized. Without preferences, it is common to assess solutions of best compromise according to Pareto dominance. A solution is said to be non-dominated (or Pareto optimal) if we cannot improve an objective without degrading another.

Regarding the association of pattern discovery and MOO, a few approaches have been proposed. In [Ghosh and Nath, 2004], a multi-objective rule mining using genetic algorithms was proposed. [van Leeuwen and Ukkonen, 2013] proposes an algorithm to mine skyline subgroups. The notion of skyline patterns is exploited in [Ugarte *et al.*, 2017] to mine high quality patterns according to multiple measures.

This paper continues a line of work that aims at connecting Constraint Programming (CP) to pattern mining to solve some data mining tasks [Guns *et al.*, 2011; Kemmar *et al.*, 2017]. We propose a new compact and flexible MOO CP model to efficiently discover Pareto optimal patterns (a.k.a skypatterns) according to a set of measures. Our model exploits condensed pattern representations to reduce the mining effort. To that end, we design a new global constraint, ADEQUATECLOSURE, for ensuring the closeness constraint over multiple measures. We show how skypatterns can be used to find high-quality non-redundant association rules without the use of thresholds. Finally, we show the relevance of our approach on both UCI datasets and a case study related to the analysis of genes expression data integrating multiple external genes annotations to discover associations between them.

2 Preliminaries

2.1 Itemset and Association Rules Mining

Let $\mathcal{I} = \{1, ..., n\}$ be a set of n *items*, an *itemset* (or pattern) P is a non-empty subset of $\mathcal{I}$. The language of itemsets corresponds to $\mathcal{L}_{\mathcal{I}} = 2^{\mathcal{I}} \backslash \emptyset$. A transactional dataset $\mathcal{D}$ is a multiset of transactions over $\mathcal{I}$, where each *transaction* $t \subseteq \mathcal{I}$; $\mathcal{T} = \{1, ..., m\}$ a set of m *transaction* indices. An itemset P *occurs* in a transaction t, iff $P \subseteq t$. The *cover* of P in $\mathcal{D}$ is the set of transactions in which it occurs: $\boldsymbol{t}(P) = \{t \in \mathcal{D} \mid P \subseteq t\}$. The *support* of P in a dataset $\mathcal{D}$ is the size of its cover: $sup(P) = |\boldsymbol{t}(P)|$. An itemset P is said to be *frequent* in $\mathcal{D}$ if $sup(P) \geq \theta$, where θ is a user-specified *minimal support threshold*. Given $T \subseteq \mathcal{D}$, $\boldsymbol{i}(T)$ is the set of items that are common to all transactions in T: $\boldsymbol{i}(T) = \{i \in \mathcal{I} \mid \forall t \in T, i \in t\}$. We define by $clos$ a *closure*

operator, given as $clos(P) = \boldsymbol{i} \circ \boldsymbol{t}(P) = \boldsymbol{i}(\boldsymbol{t}(P))$. The *closure* of an itemset P is the set of common items that belong to all transactions in $\boldsymbol{t}(P)$: $clos(P) = \{i \in \mathcal{I} \mid \forall t \in \boldsymbol{t}(P), i \in t\}$. An itemset P is said to be *closed* iff $clos(P) = P$. The closure operator is at the root of the definition of the equivalence classes, and thus the *condensed* representation of the itemsets [Pasquier *et al.*, 1999]. Interestingly, [Soulet and Crémilleux, 2008] proposed an *adequate closure operator* to measures other than the support.

Definition 1 (Adequate closure operator) *Let M be a set of measures. The closure of an itemset P w.r.t. M (called adequate closure), denoted by $clos_M(P)$, is the set of items s.t. $clos_M(P) = \{i \in \mathcal{I} \mid \forall m \in M, m(P \cup \{i\}) = m(P)\}$.*

An item i belongs to $clos_M(P)$ iff each measure $m \in M$ remains constant with the addition of i. Proposition 1 extends the closure operator $clos$ to the set M: two itemsets P and Q are in the same equivalence class iff $clos_M(P) = clos_M(Q)$. Thus, all the itemsets belonging to the same equivalence class have the same value for each $m \in M$.

Proposition 1 *Let M be a set of measures, $clos_M$ is a closure operator. P is closed w.r.t. M iff $clos_M(P) = P$.*

Many well-known interestingness measures based on support are used to evaluate the relevance of itemsets (see Table 1). All-confidence ($aconf$) is a measure of interestingness of a pattern that indicates the dependency between all of the items in the pattern [Omiecinski, 2003]. The $aconf$ of an itemset P can be defined as the ratio between its support and the maximum support of its items. This simply states that all-confidence is the smallest confidence of any rule for the set of items of P. Note that the maximum value for the denominator will occur when the subset of P consists of a single item. Hence, it is only needed to compute the support of each item. Additional information (such as numerical values associated to items) can also be used. Given a function $val : \mathcal{I} \to \mathbb{R}_+$, we extend it to an itemset P and denote by $P.val$ the multi-set $\{val(i) \mid i \in P\}$. Note that val may also be the support. This kind of function is used with the usual primitive-based measures like sum, min and max. For instance, $sum(P.val)$ is the sum of val for each item of P.

Association rules mining. An association rule is an implication $r : X \Rightarrow Y$ where X and Y are itemsets such that $X \cap Y = \emptyset$ and $Y \neq \emptyset$. X is called antecedent of the rule and Y its consequent. The support of the rule is given by $sup(r) = sup(X \cup Y)$. The confidence of the rule indicates how often the rule has been found to be true, i.e. $conf(r) = \frac{sup(r)}{sup(X)}$. Given a minimum confidence c and a minimum support θ, the goal is to find all the rules r such that $conf(r) \geq c$ and $sup(r) \geq \theta$. The lift of a rule r is defined as $lift(r) = \frac{conf(r) \times |\mathcal{D}|}{sup(Y)}$. One common use of lift is to measure the correlation between X and Y : a value greater (resp. smaller) than 1 means that X and Y are positively (resp. negatively) correlated while a value close to 1 means that X and Y are independent. To reduce the number of rules, Bastide *et al.* [Bastide *et al.*, 2000] proposed the notion of *minimal non-redundant rules* (MNR). An association rule $r : X \Rightarrow Y$ is an MNR iff: (1) $sup(r) \geq \theta$ and $conf(r) \geq c$; (2) $X \cup Y$ is closed w.r.t. the support; and (3) X is a generator. An itemset X is said to be a generator if it has no frequent subset with the same support. In the sequel, we will show how our approach avoids the hard task of setting these thresholds to get interesting patterns, thanks to multi-objective optimization.

Name	Definition
$area$	$P \mapsto sup(P) \times size(P)$
$mean$	$P \mapsto \frac{\min(P.val)+\max(P.val)}{2}$
min	$P \mapsto \min(P.val)$
$size$	$P \mapsto \lvert P \rvert$
$aconf$	$P \mapsto \frac{sup(P)}{max\{sup(P') \mid \forall P' \subseteq P \wedge P' \neq \emptyset\}}$
gr_1	$P \mapsto \frac{\lvert\mathcal{D}_2\rvert}{\lvert\mathcal{D}_1\rvert} \times \frac{sup_{\mathcal{D}_1}(P)}{sup_{\mathcal{D}_2}(P)}$

Table 1: Examples of measures where $\mathcal{D}_1$ and $\mathcal{D}_2$ are a disjoint partition of $\mathcal{D}$

.

2.2 Multi-Objective Optimization (MOO)

A MOO problem $\mathcal{P}$ [Figueira *et al.*, 2005] consists of a set of m objective functions $f_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1..m$ and a discrete set $\mathcal{X}$ of feasible solutions (a.k.a. *decision space*). For simplicity, we assume (w.l.o.g.) that the objective functions are to be maximized simultaneously. We note F as the vector-function $F : \mathbb{R}^n \to \mathbb{R}^m$, which maps each feasible solution $x \in \mathcal{X}$ into the corresponding objective vector $F(x) = (f_1(x), \ldots, f_m(x))$. We denote by $\mathcal{Y}$ the image of $\mathcal{X}$ in the objective space, s.t. $\mathcal{Y} = \{y | y = F(x),\ x \in \mathcal{X}\}$. Now, comparing solutions in $\mathcal{X}$ amounts to comparing solutions of the objective space $\mathcal{Y}$. In the absence of the decision maker's preferences, the objective vectors are commonly compared using the Pareto dominance. Let's consider $y, y' \in \mathcal{Y}$ be two solutions of $\mathcal{P}$. We say that y dominates y', denoted $y \succ y'$, iff : $\forall i \in [1..m] : y_i \geq y'_i$ and $\exists j \in [1..m] : y_j > y'_j$. A solution $y^* \in \mathcal{Y}$ is Pareto optimal (a.k.a. *efficient*) iff there is no solution $y \in \mathcal{Y}$ that dominates y^*, i.e. $\nexists y \in \mathcal{Y} : y \succ y^*$. The set of all efficient solutions is called the Pareto front.

Definition 2 (Archive) *An archive $\mathcal{A} \subseteq \mathcal{Y}$ is a set of solutions s.t. there is no solution in $\mathcal{A}$ that dominates another solution in $\mathcal{A}$: $\nexists y, y' \in \mathcal{A} : y \succ y'$.*

Our mining task is the discovery of the Pareto optimal patterns w.r.t. a set of measures (objectives) M. For a given pattern P, each variable of the objective vector $(obj_1, ..., obj_{|M|})$ represents the value $m(P)$ of the measure $m \in M$. A pattern associated to a Pareto optimal solution is called a *skypattern*. The skypattern mining problem consists of finding the set $\mathcal{S}ky$ of all skypatterns w.r.t. M. The main drawback is that the number of candidate skypatterns is exponential in the worst case (i.e. equals to all patterns, $|\mathcal{S}ky| = 2^{|\mathcal{I}|} - 1$). Fortunately, Soulet *et al.* [Soulet *et al.*, 2011] proposed *maximal skylineability* to reduce the mining effort. It consists to find a set of measures M' such that each skypattern w.r.t. M is either closed w.r.t. M' or is a subset of a skypattern closed to M'. In other words, the set of skypatterns w.r.t. M and closed w.r.t. M' forms a condensed representation of the full set of skypatterns. Thus, we only focus our search to the set of closed patterns w.r.t. M'. For instance, $M = \{sup(X), area(X), aconf(X)\}$ is maximally $M' = \{sup(X), max(X.sup)\}$-skylineable because $area$ strictly increases with the size (when sup remains constant), while $aconf$ remains constant (when $sup(X)$ and

$max(X.sup)$ remain constant). In other words, if a pattern X is a subset of pattern Y, $sup(X) = sup(Y)$, $max(X.sup) = max(Y.sup)$, then we have the guarantee that $Y \succ X$ w.r.t. the set of measures M. In this paper, we will not cover how to get M' from M, since it would be beyond its scope. For further details, we refer the reader to [Soulet *et al.*, 2011].

3 A CP Model for Mining Skypatterns

The CP paradigm [Hoeve and Katriel, 2006] is at the core of generic approaches for pattern mining. With the recent development of global constraints with efficient filtering algorithms, CP became competitive for solving some data-mining tasks [Belaid *et al.*, 2019; Schaus *et al.*, 2017; Kemmar *et al.*, 2017]. In this section, we present our CP model, called CLOSEDSKY, taking benefit of global constraints in order to get an efficient and compact model for mining skypatterns. To build this model, we need a transactional dataset $\mathcal{D}$, a set of measures M and an archive $\mathcal{A}$ (see definition 2), which is dynamically updated each time we find a new solution, since $\mathcal{A}$ should remain *domination-free*. The $\text{CLOSEDSKY}_{\mathcal{D},M,\mathcal{A}}(x, obj)$ model is given as follows:

$$\begin{cases} \text{PARETO}_{\mathcal{A}}(obj) & (1) \\ \text{ADEQUATECLOSURE}_{\mathcal{D},M'}(x) & (2) \\ \text{MEASURES}_{\mathcal{D},M}(x, obj) & (3) \end{cases}$$

(a) Variables. Our model has: (i) n *Boolean variables* x, where x_i represents the presence of the item $i \in \mathcal{I}$ in the pattern. (ii) *objective variables*: obj is a vector of integer variables, s.t. obj_i represents the value of a measure $m \in M$.
(b) Constraints. Our model exploits a set of constraints for which efficient pruning algorithms exist.
- **Pareto.** Constraint (1) is a global optimization constraint, that is used to extract skypatterns without any additional dynamic constraints (see [Ugarte *et al.*, 2017]). Formally, $\text{PARETO}_{\mathcal{A}}(obj) \equiv \bigwedge_{y \in \mathcal{A}} \bigvee_{i=1..|M|} obj_i > y_i$. It enforces that the next objective vector $obj = (obj_1, ..., obj_{|M|})$ is not dominated w.r.t. the archive $\mathcal{A}$, i.e. $\nexists y \in \mathcal{A} : y \succ obj$. This constraint is fully detailed in [Schaus and Hartert, 2013].
- **Closedness.** We introduce Constraint (2) as a new global constraint to ensure that x is closed w.r.t. a set of measures M'. It is proposed to genuinely discover condensed representations of patterns (without reified constraints). M' is automatically computed s.t. M is maximally M'-skylineable. The filtering rules of this constraint are given in Section 3.1.
- **Measures.** Constraint (3) is used to link each objective variable obj_i to a measure $m \in M$ such that $obj_i = m(x)$. This constraint also involves the definition of each measure $m \in M$. For instance, if $M = \{sup, area\}$, then $\text{MEASURES}_{\mathcal{D},M}$ can be defined as follows:

$$\text{MEASURES}_{\mathcal{D},M}(x, obj) \equiv \begin{cases} \text{COVERSIZE}_{\mathcal{D}}(x, obj_1) & (4) \\ obj_2 = obj_1 \times \sum_{i \in \mathcal{I}} x_i & (5) \end{cases}$$

where the COVERSIZE global constraint [Schaus *et al.*, 2017] models the $sup(x)$ measure, i.e. $obj_1 = |\boldsymbol{t}(x)|$, while the Constraint (5) models the *area* of x as the product of its support (saved in obj_1) by its size.

3.1 The Global Constraint ADEQUATECLOSURE

In the following, we introduce our global constraint ADEQUATECLOSURE for mining condensed representations of patterns w.r.t. M'. This is achieved thanks to the closure operator $clos_{M'}$ which is adequate for a set of measures. Contrary to [Ugarte *et al.*, 2017], our global constraint requires neither reified constraints nor auxiliary variables. All the proofs are given in the supplementary material [Vernerey *et al.*, 2022]. We will use the notations : $x^+ = \{i \in \mathcal{I} | dom(x_i) = \{1\}\}$, $x^- = \{i \in \mathcal{I} | dom(x_i) = \{0\}\}$, $x^* = \mathcal{I} \setminus \{x^+ \cup x^-\}$, where $dom(x_i)$ denotes the set of allowed values of the variable x_i.

Definition 3 (ADEQUATECLOSURE) *Let x be a vector of Boolean variables, $\mathcal{D}$ a transactional dataset and M' a set of measures. The constraint* $\text{ADEQUATECLOSURE}_{\mathcal{D},M'}(x)$ *holds iff* $clos_{M'}(x^+) = x^+$.

Now, we define the *closure inclusion* operator $\texttt{cl}_{inc}$ exploited by our global constraint filtering rules, for mining adequate condensed representations w.r.t. a set of measures M'.

Definition 4 (Closure inclusion) *Let x be a partial assignment of variables in $\{x_1, \ldots, x_{|\mathcal{I}|}\}$, M' a set of measures and i an item.* $\texttt{cl}_{inc}(x^+, i, M')$ *holds iff* $\forall m \in M', m(x^+ \cup \{i\}) = m(x^+)$.

Definition 4 provides a necessary and sufficient condition for the property of pattern condensation with respect to a set of measures M', when extending the current itemset x^+ with a free item (from x^*). In other words, $\texttt{cl}_{inc}(x^+, i, M') \Leftrightarrow i \in clos_{M'}(x^+)$. Lemma 1 characterizes a consistent partial assignment w.r.t. ADEQUATECLOSURE constraint, that is a partial assignment that can be extended to a solution which satisfies this constraint.
Lemma 1 (Consistent partial assignment) *Let x^+ be a partial assignment of variables in $\{x_1, \ldots, x_{|\mathcal{I}|}\}$ and M' a set of measures. x^+ is a consistent partial assignment iff $\nexists j \in x^-$ s.t. $\texttt{cl}_{inc}(x^+, j, M')$ holds.*

The propagator we propose for ADEQUATECLOSURE is based on two filtering rules given through Proposition 2.

Proposition 2 (Filtering rules) *Given a consistent partial assignment x, a set M' of measures, for any $i \in x^*$:* $(\mathbf{R_1})$ *if* $\texttt{cl}_{inc}(x^+, i, M') \Rightarrow 0 \notin dom(x_i)$. $(\mathbf{R_2})$ *if* $\exists j \in x^-$ *s.t.* $\texttt{cl}_{inc}(x^+ \cup \{i\}, j, M') \Rightarrow 1 \notin dom(x_i)$.

The first rule filters 0 from $dom(x_i)$ if $\{i\}$ is a closure inclusion of x^+ (see Def. 4). It allows to add all items $i \in x^*$ that are required to extend x^+ to a closed pattern w.r.t. M'. The second rule filters 1 from $dom(x_i)$ if $\texttt{cl}_{inc}(x^+ \cup \{i\}, j, M')$ holds where $j \in x^-$. Roughly speaking, this rule checks whether $x^+ \cup \{i\}$ cannot be extended to a closed itemset w.r.t. M' without adding j (see Lemma 1).

4 Mining MNRs using Skypatterns

In association rules mining, finding appropriate values for c and θ (which are involved in confidence and support constraints) without prior knowledge is very difficult, unless the end-user is able to perfectly manage thresholds selections in this paradigm. In this section, we show how skypatterns can

be used to mine high-quality MNRs **without any threshold**. The whole process consists of two solving steps:

- *Generating skypatterns*: compute skypatterns with CLOSEDSKY model detailed in Section 3, using the set of measures $M_r = \{sup, area, aconf\}$.
- *Generating MNRs*: generate all MNRs from the collection of representative skypatterns computed in the first step using a dedicated CP model.

(a) Generating skypatterns. The main idea is to take advantage on a condensed representation of $\mathcal{L}_\mathcal{I}$ to compute a **reduced but relevant** collection of representative skypatterns w.r.t. M_r for mining MNRs. Indeed, the maximal skylineability principle allows us to reduce M_r to $M'_r = \{sup(X), max(X.sup)\}$. Moreover, the set of patterns maximizing both support and All-confidence measures (a.k.a. skypatterns) might be the more interesting ones for generating relevant rules. However, as these measures are anti-monotone (i.e. they decrease with specialization), small patterns are more likely to be selected. To increase the average length of the extracted skypatterns and thus the size of the MNRs generated, we add the $area$ measure to M_r.

Pareto optimal patterns w.r.t. M_r are interesting since they lead to rules with a high support and/or high confidence. Let $P = X \cup Y$ be a skypattern and $r : X \Rightarrow Y$ be a rule. We have $conf(r) \geq aconf(P)$ and $sup(r) = sup(P)$. This can be seen directly from the definition. As we are maximizing both the $aconf$ and sup measures when computing skypatterns, all rules that can be produced from P have a confidence (resp. support) greater than or equal to our minimum $aconf(P)$ (resp. $sup(P)$) value. Thus, all the extracted rules are valid since they satisfy the implicit thresholds $c = aconf(P)$ and $\theta = sup(P)$. Moreover, our approach allows to discover rules with a low support but a high confidence (i.e. *rare* rules). This kind of rules is hard to extract with a classic rule miner since (1) the user should tune the thresholds accordingly, (2) the number of rules extracted with low thresholds may be prohibitive.

(b) Generating MNRs. Let $clos_{sup}(P)$ be the closure of P w.r.t. $\{sup\}$ and $clos^-_{sup}(P) = clos_{sup}(P) \setminus P$.

First, it is interesting to stress that the first MNR condition (i.e. $sup(r) \geq \theta$ and $conf(r) \geq c$) is satisfied since $c = aconf(P)$ and $\theta = sup(P)$. Thus, our MNRs generation step is **threshold-free**. Second, to be an MNR (2^{nd} condition), we need to ensure that, for any rule $r : X \Rightarrow Y$, $P = X \cup Y$ must be *closed* w.r.t. $\{sup\}$. However, P can be closed w.r.t. M'_r without being closed w.r.t. $sup \in M'_r$. To satisfy the second MNR condition, we impose that $X \cup Y = clos_{sup}(P)$. However, we have $aconf(P) \geq aconf(clos_{sup}(P))$ which means that the confidence of the generated rules might be smaller than $aconf(P)$. That is why we add the constraint $clos^-_{sup}(P) \subseteq Y$. Finally ($3^{rd}$ condition), we enforce that X must be a generator using the global constraint GENERATOR introduced by [Belaid *et al.*, 2019]. As a summary, given the set $\mathcal{S}ky$ of skypatterns w.r.t. M_r, we generate all the MNRs $r : X \Rightarrow Y$ s.t. $\exists P \in \mathcal{S}ky : X \cup Y = clos_{sup}(P) \wedge clos^-(P) \subseteq Y \wedge$ GENERATOR(X). We note that all the rules generated from the closure of the skypatterns have the same support and confidence as those that could be generated using skypatterns directly. More interestingly, these rules are MNRs. Proposition 3 summarises this important result.

Proposition 3 *Let P be a skypattern, $r : X \Rightarrow Y$ a rule s.t. $X \cup Y = P$, and X a generator. Let $r' : X \Rightarrow Y'$ be a rule s.t. $X \cup Y' = clos_{sup}(P) \wedge Y' = Y \cup clos^-(P)$. r' is an MNR with $sup(r) = sup(r') \wedge conf(r') = conf(r) \geq aconf(P)$.*

Hence, extracting all MNRs from $\mathcal{S}ky$ can be modeled using three vectors x, y and z of n Boolean variables, where x_i, y_i and z_i respectively represent the presence of item $i \in \mathcal{I}$ in the antecedent of the rule, in the consequence of the rule, and in the rule as a whole z. A forth vector sky of $|\mathcal{S}ky|$ Boolean variables is also introduced. Our model involves different types of constraints : a channeling constraint ensuring that $z = (x \cup y)$; one channeling constraint per $s \in \mathcal{S}ky$ ensuring that $z = clos_{sup}(s)$ and $clos^-(s) \subseteq y$ and GENERATOR(x). Other constraints ensure that both the antecedent and the consequence of the rule are not empty and that an item cannot simultaneously appear in both of them.

5 Related Work

Computing skypatterns. [Soulet *et al.*, 2011] have proposed AETHERIS, which proceeds in two steps. First, condensed representations of patterns w.r.t. the set M' are extracted. Then, a post-processing step is applied to remove all patterns that are not skypatterns. [Ugarte *et al.*, 2017] also proposed a two-step approach (called CP+SKY), but contrary to AETHERIS, CP+SKY dynamically builds a more concise representation thanks to constraints on the dominance relation, which are dynamically added during the mining. In [Négrevergne *et al.*, 2013], the concept of *dominance programming* is used to propose a generic method extending relational algebras towards pattern mining. The solving step is similar to the technique used in CP+SKY, but does not exploit the skylineability principle. Our approach also benefits from theoretical relationships between pattern condensed representations and skypatterns but requires no post processing step and exploits global constraints to efficiently extract condensed representations of patterns and their Pareto fronts.

Computing MNRs. Several algorithms have been proposed to extract ARs and MNRs. ECLAT-Z [Szathmary *et al.*, 2008] is the most efficient one. Omiecinski [Omiecinski, 2003] proposed an algorithm for mining ARs with a minimum All-confidence. Belaid *et al.* [Belaid *et al.*, 2019] proposed a CP model for mining ARs and MNRs. SAT-based approaches have also been proposed [Izza *et al.*, 2020]. Unlike our proposition, all these methods require thresholds.

6 Experimental Evaluation

We report an experimental study on UCI datasets and on a case study from gene expression analysis. Experiments were conducted on Intel(R) Xeon(R) CPU E7-8870, 2.10GHz with a RAM of 1.48 TB and a time limit of one hour. Our approach is implemented in Java top-on the `Choco` solver [Prud'homme *et al.*, 2016] version 4.10.8, a Java

library for constraint programming, and is publicly available [Vernerey *et al.*, 2022]. We implemented a variant of CLOSEDSKY (denoted CLOSEDSKY-WC) ensuring a weaker consistency by disabling the filtering rule ($\mathbf{R_2}$). First, we carry out a comparison to enumerate skypatterns. For baseline comparison, we retain the CP-based method CP+SKY implemented using `Choco` library and the C++ implementation of the specialized method AETHERIS. Second, we perform an evaluation to mining MNRs, where we retain the CP-based method CP4MNR [Belaid *et al.*, 2019] and the specialized method ECLAT-Z [Szathmary *et al.*, 2008]. For all the skypattern mining experiments, we use the following heuristic: we select the free item i such that $sup(x^+ \cup \{i\})$ is minimal and we first instantiate it to 0.

6.1 UCI Datasets

We evaluate our approach on six datasets of the FIMI repository[1]. We consider two sets of measures $M_s = \{sup, area, gr, mean, \max\}$ and $M_r = \{sup, aconf, area\}$. Measures using numeric values, like $mean$ or max, were applied to randomly generated attribute values (within the range $[0, 1]$) because the used repositories have no variables with numeric values. We stress that all measures having continuous values are encoded as integer decision variables.

Mining skypatterns. Table 2a reports comparative results of CLOSEDSKY against CP+SKY and AETHERIS. We report the number of skypatterns $|\mathcal{S}ky|$, which is the same for all methods, and the total CPU times (in seconds) for each dataset. Since AETHERIS does not implement the $aconf$ measure, we do not include it for the set M_r. As one can observe, the number of skypatterns generated is very low. Regarding the runtime, CLOSEDSKY outperforms CP+SKY on almost all tested datasets. Moreover, for some datasets e.g., Chess, Heart-cleveland and Mushroom, the gain is impressive. This is due to the huge number of reified constraints required by CP+SKY to model the closeness constraint. Interestingly enough, our approach overpasses the specialized algorithm AETHERIS. The huge number of closed patterns w.r.t. M' computed by AETHERIS in the first step makes the whole extraction process intractable for large datasets. Finally, we can notice that CLOSEDSKY-WC greatly improves the performance of CLOSEDSKY, e.g., Connect dataset.

Mining MNRs. We apply our two-step approach SKY4MNR (see Sect. 4) to extract relevant MNRs. We use CLOSEDSKY-WC to perform the skypatterns generating step. CP4MNR and ECLAT-Z require to fix the values for thresholds c and θ which are involved in confidence and support constraints. We set c to $min_{s \in \mathcal{S}ky} aconf(s)$ and θ to $min_{s \in \mathcal{S}ky} sup(s)$. This allows to generate a superset of all the MNRs extracted by SKY4MNR. Table 2b reports comparative results of SKY4MNR against CP4MNR and ECLAT-Z. It reports the number of MNRs (obviously, the same for both CP4MNR and ECLAT-Z) and the total CPU time (in seconds) for each dataset. The table shows that the reduction in the number of generated MNRs is huge. Remarkably, for the dataset Mushroom, this reduction is still drastic (from 10^5 to 24). It is noteworthy that SKY4MNR finds these MNR sets in less than 2 seconds on almost all tested datasets. For the Pumsb dataset, SKY4MNR takes 141 seconds while CP4MNR and ECLAT-Z reaches TO.

Qualitative analysis of MNRs. To analyse the set of all MNRs $\mathcal{R}$ generated by each approach (ECLAT-Z and CP4MNR result in the same set of MNRs), we apply to the result of each approach a postprocessing step to discard all MNRs that are not on the Pareto front w.r.t. three rule assessment measures sup, $conf$ and $lift$ as follows: Let f be a mapping from $\mathcal{R}$ to $\mathbb{R}^3$ that associates, to each rule $r \in \mathcal{R}$, a data point $f(r) \in \mathbb{R}^3$ with coordinates $(sup(r), conf(r), lift(r))$. Let $\mathcal{F} = \{f(r) \mid r \in \mathcal{R}\}$. Applying a skyline extractor method on $\mathcal{F}$ ensures to provide the Pareto front points. Figs. 1a-1c plot their values according to measures sup and $conf$. We can observe that most of the rules given by SKY4MNR are highly confident; the vast majority of the rules have confidences between 90 and 100 %. Moreover, their relative support ranges from 50 to 100 %, suggesting a trade-off between support and confidence. Interestingly, for the Mushroom dataset, SKY4MNR is able to identify rules with the lowest relative support and with the highest confidence and then lift (see suppl. material) without having to set any thresholds. This is a particularly distressing result for the traditional rule generation framework, unless the end-user is able to perfectly manage thresholds selection, which may seem unrealistic. Fig.1 (in suppl. material) plots the values of sup and $lift$ for all skyline points. We have similar observations as for support vs. confidence: SKY4MNR is able to find rules that are close to the skyline points.

6.2 Genes Expression Data

The aim of this study is to investigate the use of skypatterns to discover relationships between gene expression profiles and biological knowledge describing known gene properties represented as annotations in biological databases. Experiments were carried out on Eisen et al. genomic dataset[2] containing expression measures of 2465 Yeast genes for 79 biological conditions. Each yeast gene was annotated with the GO IDs of its associated terms in Yeast Gene Ontology, the PubMed IDs representing its associations with research papers, the IDs of the KEGG pathways in which it is involved, its phenotypes annotations and the names of the transcriptional regulator genes. All annotations were transformed into Boolean data, indicating if an annotation pertains or not to a given gene. The resulting datset is a matrix of 2465 transactions representing yeast genes and 9634 items representing discretized gene expression measures and gene annotations.

Mining skypatterns. CLOSEDSKY-WC again obtains the best performance: it takes 372 seconds to complete the extraction, which results in a total of 13 skypatterns, while the two approaches CP+SKY and AETHERIS run into Time Out. Interestingly enough, CLOSEDSKY also runs into Time Out.

Mining MNRs. We report comparative results of SKY4MNR against CP4MNR and ECLAT-Z. We also

[1] fimi.ua.ac.be/data

[2] i3s.unice.fr/~pasquier/web/

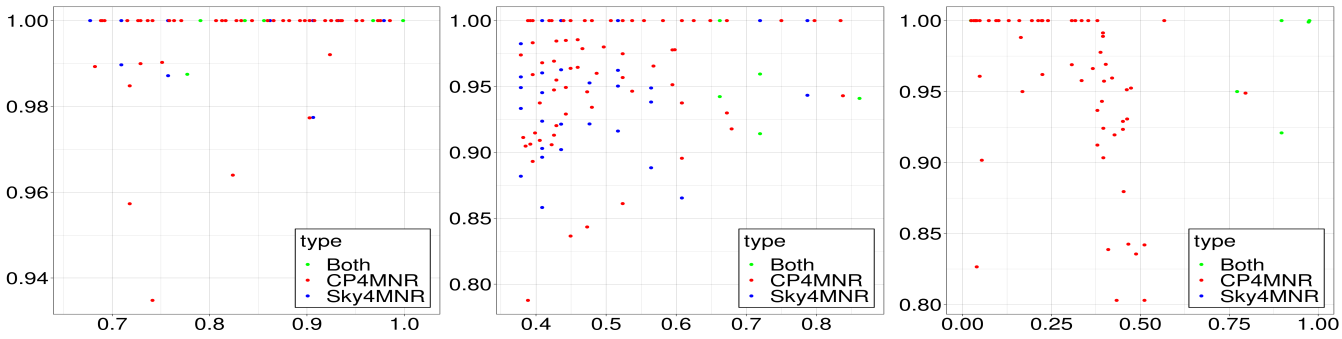
Dataset	M_s					M_r			
	$\|Sky\|$	(1)	(2)	(3)	(4)	$\|Sky\|$	(1)	(2)	(3)
Anneal	366	3	**2**	8	2468	21	**1**	**1**	2
Chess	3014	307	**199**	3517	TO	12	2	**1**	22
Connect	3015	TO	**1210**	TO	MO	16	47	**1**	1013
Heart-cleveland	795	183	**130**	3224	TO	13	6	**2**	11
Mushroom	551	7	**2**	213	16	6	2	**1**	54

(a) Mining skypatterns. (1): CLOSEDSKY (2): CLOSED-SKY-WC (3): CP+SKY (4): AETHERIS TO: Time Out MO: Memory Out.

Dataset	$\|\mathcal{T}\| \times \|\mathcal{I}\|$	$\|\mathcal{R}\|$		Time (s.)		
		(1)	(2) (3)	(1)	(2)	(3)
Anneal	812×89	**4806**	10^6	**1**	16	460
Chess	$3,196 \times 75$	**9450**	10^6	**1**	56	63
Connect	$67,557 \times 129$	**17882**	10^7	**2**	1423	2094
Heart-cleveland	296×95	**1104**	10^6	**2**	10	486
Mushroom	$8,124 \times 112$	**24**	10^5	**1**	7	243
Pumsb	$49,046 \times 2,113$	$\mathbf{10^5}$	$> 10^7$	**140**	TO	TO

(b) Mining MNRs. (1): SKY4MNR (2): CP4MNR (3): ECLAT-Z

Table 2: Performance analysis on UCI datasets. TO (resp. MO) is shown when the time limit (resp. the memory limit) is reached.



(a) Anneal (b) Heart-cleveland (c) Mushroom

Figure 1: Qualitative analysis of MNRs: support(x-axis) vs. confidence(y-axis).

Rule	Antecedent	Consequent	Supp.(#)	Conf. (%)
1	pmid:14576278	go:0005737, go:0005739	430	100
2	pr:FHL1	go:0005737, go:0005840, go:0006412	105	79
3	path:00970	go:0005737, go:0016874, go:0006412, pmid:1108023	32	100
4	go:0005739	go:0005737	532	100
5	go:0005737, go:0006412	heat3↓	109	38
6	path:03010	heat3↓	97	74
7	path:03010	heat3↓, ndt80-1↓	66	50
8	heat3↓, ndt80-1↓	path:03010	66	90
9	heat3↓	go:0005737, go:0006412	109	83
10	heat3↓	go:0005737, go:0005840, go:0006412, pr:FHL1	85	64

Table 3: Examples of MNRs generated by SKY4MNR. heat3 and ndt80 refer to time points of the heat shock and sporulation experiments, respectively. ↓ denotes an under-expression. The prefixes go, path, pmid, pr allow to identify GO terms, KEGG pathways, PubMed ids and names of transcriptional regulators, respectively.

compared to GENMINER [Martinez *et al.*, 2008], a specialized tool for mining MNRs from gene expression data. It uses the NORDI algorithm for discretizing continuous values and the CLOSE algorithm [Pasquier *et al.*, 1999] for MNRs extraction. Recall that SKY4MNR is threshold-free. For other approaches, we consider the same thresholds θ and c used in GENMINER, i.e. $\theta = 0.3\%$ (at least 7 genes) and $c = 50\%$. GENMINER found more than 1.33×10^6 MNRs within the time limit of 2 hours, while CP4MNR extracted 364119 MNRs in 364 seconds. Applying SKY4MNR results in a total of 63 MNRs in 375 seconds. Finally, ECLAT-Z runs into MO.

Rules analysis. Table 3 shows three forms of rules extracted by SKY4MNR. Rules with the form *annotations* $\Rightarrow$ *expressions* (see rules 5-7) mean that a group of gene associated with a specific set of annotations is likely to be over-expressed or under-expressed. Rules with the form *expressions* $\Rightarrow$ *annotations* mean that when a group of genes is over-expressed or under-expressed in a set of biological conditions, these genes are likely to have the corresponding gene annotations (see rules 8-10). For instance, rule 8 highlights a group of genes involved in ribosomal organization (path:03010) that are under-expressed after both a heat shock and a sporulation experiments [Carmona-Saez *et al.*, 2006]. Finally, rules 1-4 reveal possible links between annotations from different sources like the relationship between KEGG pathways and Gene Ontology terms. Almost all the rules of Table 3 have been highlighted in [Martinez *et al.*, 2007].

The expressiveness of our CP approach allows to ask for rules with some specific items in the antecedent and/or consequent or for rules not involving a particular set of items. For rules of the form 1 and 2, we imposed the presence of at least one expression item in the skypattern extraction step. We also forbidden the presence of the pimd items from the patterns because too many of them contain only pimd items. However, their closure may contain such items. Such constraints are not directly handled by GENMINER, a post-processing step to filter out the undesirable patterns is required.

7 Conclusions

We have proposed a new compact and flexible MOO CP model to efficiently mine Pareto optimal patterns according to a set of measures. Our model exploits condensed pattern representations that allows us to reduce the mining effort. We designed a new global constraint for ensuring the closeness constraint over multiple measures. We showed how skypatterns can be used to find a small but interesting set of MNRs without the use of thresholds. An extensive campaign of experiments conducted over UCI datasets and gene expression data has shown the effectiveness of our approach for both skypattern and association rules mining compared to CP based and specialized approaches.

References

[Agrawal and Srikant, 1994] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th VLDB*, pages 487–499, San Francisco, CA, USA, 1994.

[Bastide *et al.*, 2000] Yves Bastide, Nicolas Pasquier, Rafik Taouil, Gerd Stumme, and Lofti Lakhal. Mining Minimal Non-redundant Association Rules Using Frequent Closed Itemsets. In *Computational Logic — CL 2000*, pages 972–986. Springer, 2000.

[Belaid *et al.*, 2019] Mohamed-Bachir Belaid, Christian Bessiere, and Nadjib Lazaar. Constraint Programming for Association Rules. In *Proceedings of the 2019 SIAM International Conference on Data Mining (SDM)*, pages 127–135, 2019.

[Carmona-Saez *et al.*, 2006] Pedro Carmona-Saez, Monica Chagoyen, Andres Rodríguez, Oswaldo Trelles, Jose María Carazo, and Alberto Pascual-Montano. Integrated analysis of gene expression by association rules discovery. *BMC Bioinform.*, 7:54, 2006.

[De Raedt and Zimmermann, 2007] Luc De Raedt and Albrecht Zimmermann. Constraint-based pattern set mining. In *7th SIAM SDM*, pages 237–248. SIAM, 2007.

[Figueira *et al.*, 2005] José Figueira, Salvatore Greco, and Matthias Ehrgott. *Multiple Criteria Decision Analysis: State of the Art Surveys*. Springer, 2005.

[Ghosh and Nath, 2004] Ashish Ghosh and Bhabesh Nath. Multi-objective rule mining using genetic algorithms. *Inf. Sci.*, 36(1-3):123–133, 2004.

[Guns *et al.*, 2011] Tias Guns, Siegfried Nijssen, and Luc De Raedt. Itemset mining: A constraint programming perspective. *Artificial Intelligence*, 175(12):1951–1983, 2011.

[Hoeve and Katriel, 2006] Willem-Jan van Hoeve and Irit Katriel. Global constraints. In *Handbook of Constraint Programming*, pages 169–208. Elsevier, 2006.

[Izza *et al.*, 2020] Yacine Izza, Said Jabbour, Badran Raddaoui, and Abdelhamid Boudane. On the enumeration of association rules: A decomposition-based approach. In *Proc. of IJCAI 2020*, pages 1265–1271, 2020.

[Kemmar *et al.*, 2017] Amina Kemmar, Yahia Lebbah, Samir Loudni, Patrice Boizumault, and Thierry Charnois. Prefix-projection global constraint and top-k approach for sequential pattern mining. *Constraints An Int. J.*, 22(2):265–306, 2017.

[Martinez *et al.*, 2007] Ricardo Martinez, Claude Pasquier, and Nicolas Pasquier. Genminer: Mining informative association rules from genomic data. In *Proc. of BIBM 2007*, pages 15–22, 2007.

[Martinez *et al.*, 2008] Ricardo Martinez, Nicolas Pasquier, and Claude Pasquier. GenMiner: mining non-redundant association rules from integrated gene expression data and annotations. *Bioinformatics*, 24(22):2643–2644, 2008.

[Négrevergne *et al.*, 2013] Benjamin Négrevergne, Anton Dries, Tias Guns, and Siegfried Nijssen. Dominance programming for itemset mining. In *Proceedings of the 13th ICDM*, pages 557–566. IEEE Computer Society, 2013.

[Omiecinski, 2003] Edward Robert Omiecinski. Alternative interest measures for mining associations in databases. *IEEE TKDE*, 15(1):57–69, 2003.

[Pasquier *et al.*, 1999] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lofti Lakhal. Discovering frequent closed itemsets for association rules. In *Proceedings of the 7th ICDT*, pages 398–416, 1999.

[Prud'homme *et al.*, 2016] Charles Prud'homme, Jean-Guillaume Fages, and Xavier Lorca. Choco Solver Documentation, 2016.

[Schaus and Hartert, 2013] Pierre Schaus and Renaud Hartert. Multi-Objective Large Neighborhood Search. In *Proceedings of CP 2013*, 2013.

[Schaus *et al.*, 2017] Pierre Schaus, John Oscar Raoul Aoga, and Tias Guns. Coversize: A global constraint for frequency-based itemset mining. In *Proceedings of the 23rd CP 2017*, pages 529–546, 2017.

[Soulet and Crémilleux, 2008] Arnaud Soulet and Bruno Crémilleux. Adequate condensed representations of patterns. *Data Min. Knowl. Discov.*, 17(1):94–110, 2008.

[Soulet *et al.*, 2011] Arnaud Soulet, Chedy Raïssi, Marc Plantevit, and Bruno Crémilleux. Mining dominant patterns in the sky. In *Proceedings of the ICDM 2011*, pages 655–664. IEEE Computer Society, 2011.

[Szathmary *et al.*, 2008] Laszlo Szathmary, Petko Valtchev, Amedeo Napoli, and Robert Godin. An Efficient Hybrid Algorithm for Mining Frequent Closures and Generators. In *Proc. of CLA '08*, pages 47–58, 2008.

[Ugarte *et al.*, 2017] Willy Ugarte, Patrice Boizumault, Bruno Crémilleux, Alban Lepailleur, Samir Loudni, Marc Plantevit, Chedy Raïssi, and Arnaud Soulet. Skypattern mining: From pattern condensed representations to dynamic constraint satisfaction problems. *Artif. Intell.*, 244:48–69, 2017.

[van Leeuwen and Ukkonen, 2013] Matthijs van Leeuwen and Antti Ukkonen. Discovering skylines of subgroup sets. In *ECML PKDD 2013*, pages 272–287, 2013.

[Vernerey *et al.*, 2022] Charles Vernerey, Samir Loudni, Yahia Lebbah, and Noureddine Aribi. Code and supplementary material. https://gitlab.com/chaver/data-mining, 2022. Accessed: 2022-05-16.

[Wang *et al.*, 2005] Jianyong Wang, Jiawei Han, Ying Lu, and Petre Tzvetkov. TFP: an efficient algorithm for mining top-k frequent closed itemsets. *IEEE Trans. Knowl. Data Eng.*, 17(5):652–664, 2005.