# Une contrainte globale pour l'extraction de motifs d'intervalles fréquents fermés

Djawad Bekkoucha, Abdelkader Ouali, Patrice Boizumault, Bruno Crémilleux

▶ **To cite this version:**

Djawad Bekkoucha, Abdelkader Ouali, Patrice Boizumault, Bruno Crémilleux. Une contrainte globale pour l'extraction de motifs d'intervalles fréquents fermés. Journées Francophones de Programmation par Contraintes 2024, Lens, France, Jun 2024, Lens, France. hal-04644333

**HAL Id: hal-04644333**
**https://hal.science/hal-04644333v1**

Submitted on 11 Jul 2024

**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.

# Une contrainte globale pour l'extraction de motifs d'intervalles fréquents fermés

**Djawad Bekkoucha**[1]*  **Abdelkader Ouali**[1]  **Patrice Boizumault**[1]  **Bruno Crémilleux**[1]

[1] Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, Caen, FRANCE

prénom.nom@unicaen.fr

## 1 Introduction

L'extraction de motifs à partir de *données numériques* est une tâche cruciale en fouille de données. Cette tâche cherche à identifier des relations implicites dans de grands volumes de données, afin qu'elles puissent être interprétées par des experts ou utilisées dans des tâches ultérieures.

Bien que de nombreuses méthodes (qu'elles soient ad hoc ou bien déclaratives) aient été proposées pour extraire des motifs sur des données binaires ou séquentielles, il n'existe qu'une unique approche ad hoc MinIntChange [3] pour l'extraction de *motifs d'intervalles fermés* sur les données numériques.

Une première voie déclarative serait d'utiliser les approches "à base de contraintes" pour l'extraction de motifs ensemblistes [4, 5]. Mais, cela oblige à un prétraitement (binarisation du jeu de données numériques) et à un post-traitement (pour retrouver les motifs numériques). De plus, cette approche entraîne une perte d'informations ou une explosion combinatoire en raison de la grande taille des jeux de données binarisés.

Une toute première approche déclarative, basée sur l'usage de contraintes réifiées, a été proposée dans [2]. Mais cette approche est confrontée au passage à l'échelle. Cet article présente une contrainte globale permettant d'imposer qu'un motif d'intervalles soit fréquent et fermé. La section 2 introduit les différentes définitions. La section 3 présente notre contrainte globale gc4cip et les règles de filtrage associées. La section 4 illustre la généricité de notre approche en modélisant un problème de clustering conceptuel. Enfin, la section 5 compare les différentes approches et, montre les apports de gc4cip [1].

*Papier doctorant : Djawad Bekkoucha[1] est auteur principal.

1. La version originale de ce papier est publiée dans [1]

## 2 Préliminaires

**Un jeu de données numériques**, noté $\mathcal{N}$, est défini par un ensemble d'objets $\mathcal{G}$, où chaque objet est caractérisé par un ensemble d'attributs $\mathcal{M}$. Chaque attribut $m \in \mathcal{M}$ a un ensemble fini de valeurs noté $\mathcal{N}_m$. Un objet $g \in \mathcal{G}$ est défini par un vecteur de valeurs numériques $< v_{g,m} >_{m\in\{1,...,|\mathcal{M}|\}}$.

**Motifs d'intervalles fermés.** Dans ce papier, une régularité dans les données numériques est représentée par un motif d'intervalles [3] qui est un vecteur d'intervalles $\mathcal{V} = \langle[\underline{w_m}, \overline{w_m}]\rangle_{\forall m\in\mathcal{M}}$, où $\underline{w_m}, \overline{w_m} \in \mathcal{N}_m$. Chaque dimension correspond à un attribut selon un ordre canonique sur l'ensemble d'attributs $\mathcal{M}$. Notons $\mathcal{B}[g] = \langle[v_{g,m}, v_{g,m}]\rangle_{m\in\{1,...,|\mathcal{M}|\}}$ comme le vecteur d'intervalles correspondant à un objet $g$. Un objet $g$ est une occurrence de $\mathcal{V}$ si chaque intervalle dans $\mathcal{B}[g]$ est inclus dans l'intervalle de $\mathcal{V}$, noté comme $\mathcal{B}[g] \sqsubseteq \mathcal{V} \iff [v_{g,m}, v_{g,m}] \subseteq [\underline{w}_m, \overline{w}_m], \forall m \in \{1, ..., |\mathcal{M}|\}$. **La couverture** d'un motif d'intervalles $\mathcal{V}$ dans $\mathcal{N}$ est l'ensemble d'objets $g$ inclus dans les intervalles de $\mathcal{V}$, i.e. $cover(\mathcal{V}) = \{g \in \mathcal{G} |\ \ \mathcal{B}[g] \sqsubseteq \mathcal{V}\}$. **La fréquence** de $\mathcal{V}$ correspond au cardinal de sa couverture, soit $freq(\mathcal{V}) = |cover(\mathcal{V})|$. Pour un seuil de fréquence minimum $\theta$ donné, le motif $\mathcal{V}$ est considéré comme fréquent ssi $freq(\mathcal{V}) \geq \theta$. Pour un sous-ensemble d'objets $G \subseteq \mathcal{G}$, la description de $G$ est un motif d'intervalles $\mathcal{V}$ où, pour chaque $g \in G$, $g$ est une occurrence de $\mathcal{V}$, i.e. $desc(G) =< [a_m, b_m] >_{m\in\{1,...,|\mathcal{M}|\}}$ tel que $a_m = min(\{v_{g,m} |\ g\ \in G\})$ et $b_m = max(\{v_{g,m} | g\ \in G\})$.

Afin de réduire le nombre de motifs extraits nous nous concentrons sur des représentations condensées de motifs, à travers les *motifs fermés*. Un motif d'intervalles fermé est défini comme étant le vecteur d'intervalles le plus restreint sur sa couverture (i.e. $close(\mathcal{V}) \iff desc(cover(\mathcal{V})) = \mathcal{V}$).

## 3 Contrainte globale $GC4CIP_{\mathcal{N},\theta}(\underline{X},\ \overline{X})$

Les bornes inférieures et supérieures d'un motif d'intervalles sont représentées par deux ensembles de variables $\underline{X} = \{\underline{x}_1, ..., \underline{x}_{|\mathcal{M}|}\}$ et $\overline{X} = \{\overline{x}_1, ..., \overline{x}_{|\mathcal{M}|}\}$, chaque variable ayant comme domaine initial les valeurs associées dans la base de données. La contrainte globale $GC4CIP_{\mathcal{N},\theta}(\underline{X},\ \overline{X})$ est satisfaite ssi le motif courant d'intervalles $\mathcal{V}$ est fermé, i.e. $close(\mathcal{V})$ et $\mathcal{V}$ est fréquent, i.e. $freq(\mathcal{V})) \geq \theta$. Cette contrainte globale possède 3 règles de filtrage :

**Règle 1 (fermeture)** Elle supprime les valeurs apparaissant uniquement dans des objets non couverts par le plus grand motif d'intervalle possible noté $\mathcal{V}^* = \langle[\min(\mathcal{D}(\underline{x}_1)), \max(\mathcal{D}(\overline{x}_1))], \ldots, [\min(\mathcal{D}(\underline{x}_{|\mathcal{M}|})), \max(\mathcal{D}(\ \overline{x}_{|\mathcal{M}|}))]\rangle$. Soit $m \in \mathcal{M}$, $g \in \mathcal{G}$. $v_{g,m} \notin \mathcal{D}(\underline{x}_m)$ et $v_{g,m} \notin \mathcal{D}(\overline{x}_m)$ si :

$$\begin{cases} \exists\ m' \in \mathcal{M}, m \neq m', v_{g,m'} < \min(\mathcal{D}(\underline{x}_{m'})) \vee \\ \qquad\qquad\qquad\qquad v_{g,m'} > \max(\mathcal{D}(\overline{x}_{m'})) \\ \wedge \\ \forall g' \in \mathcal{G},\ g \neq g' \text{ tel que } g' \in cover(\mathcal{V}^*), v_{g,m} \neq v_{g',m} \end{cases}$$

**Règle 2 (fermeture)** Considérons tout d'abord la jointure entre les domaines de deux attributs $m$ et $m'$. Pour chaque valeur $v_{g,m}$ dans $\mathcal{D}(x_m)$ [2], nous considérons les objets $g$ ayant une telle valeur pour l'attribut $m$, i.e. $I_m = \{g \mid g \in \mathcal{G}, v_{g,m} \in D(x_m)\}$. Ainsi, pour chaque objet $g$ dans $I_m$, nous déterminons sa valeur pour l'attribut $m'$. On obtient donc $join(x_{m'}, x_m) = \{v_{g,m'} \mid v_{g,m'} \in D(x_{m'}), g \in I_m\}$. Dès lors, toute valeur située en dehors des limites de cet ensemble est supprimée.

$$\begin{cases} v_{g,m} \notin \mathcal{D}(\underline{x}_m) \text{ si} & v_{g,m} > max(join(x_{m'}, \underline{x}_m)). \\ v_{g,m} \notin \mathcal{D}(\overline{x}_m) \text{ si} & v_{g,m} < min(join(x_{m'}, \overline{x}_m)). \end{cases}$$

**Règle 3 (fréquence)** Elle supprime les valeurs n'apparaissant jamais dans des motifs d'intervalles fréquents. Soit $m \in \mathcal{M}$, et $\mathcal{V}^p = \langle[\min(\mathcal{D}(\underline{x}_i)), \max(\mathcal{D}(\overline{x}_i))]\rangle_{\ 1 \leq i \neq m \leq |\mathcal{M}|}$ un motif d'intervalles partiel.

$$\begin{cases} a_m \notin \mathcal{D}(\underline{x}_m) \text{ si} & freq(\mathcal{V}^p +\!\!+ [a_m, \max(\mathcal{D}(\overline{x}_m))]) < \theta. \\ b_m \notin \mathcal{D}(\overline{x}_m) \text{ si} & freq(\mathcal{V}^p +\!\!+ [\min(\mathcal{D}(\underline{x}_m)), b_m]) < \theta. \end{cases}$$

## 4 Application au clustering conceptuel

Pour illustrer la déclarativité et la généricité de notre approche, nous considérons le problème du k-clustering conceptuel. L'objectif est de trouver un clustering qui forme une k-partition des objets du jeu de données. Chaque cluster est représenté par un motif d'intervalles fermé, d'où le nom de clustering conceptuel. La recherche de $k$ motifs d'intervalles fermés $\{\mathcal{V}^1, ..., \mathcal{V}^k\}$ est modélisée par la conjonction de contraintes suivantes :

$$\begin{cases} \textbf{Closure} & GC4CIP_{\mathcal{N},\theta}(\mathcal{V}^i) \quad ,\forall\ 1 \leq i \leq k \\ \textbf{No overlapping} & cover(\mathcal{V}^i) \cap cover(\mathcal{V}^j) = \emptyset, \\ & \forall\ 1 \leq i < j \leq k \\ \textbf{Total coverage} & \bigcup_{1 \leq i \leq k} cover(\mathcal{V}^i) = \mathcal{G} \end{cases}$$

## 5 Résultats expérimentaux

Nous comparons les performances de GC4CIP avec l'unique approche ad hoc MININTCHANGE [3] pour l'extraction de motifs d'intervalles fermés, ainsi que trois approches déclaratives : CP4CIP [2] , CP4IM [5] et CLOSEDPATTERN [4]. CP4CIP extrait directement les motifs d'intervalles fermés des données numériques, tandis que CP4IM et CLOSEDPATTERN nécessitent une étape de pré-traitement et de post-traitement pour être appliquées aux données numériques.

Nos expérimentations montrent que : (1) GC4CIP offre un meilleur passage à l'échelle avec de meilleures temps CPU que toutes les autres approches déclaratives, (ii) MININTCHANGE demeure plus efficace en termes de temps CPU, si l'on se limite à l'extraction de motifs fréquents fermés, (iii) le modèle de clustering conceptuel parvient à former des clusterings de meilleure qualité que l'approche heuristique K-MEANS.

## Références


[1] Djawad BEKKOUCHA, Abdelkader OUALI, Patrice BOIZUMAULT et Bruno CRÉMILLEUX : Efficiently mining closed interval patterns with constraint programming. *In CPAIOR*, 2024.

[2] Djawad BEKKOUCHA, Abdelkader OUALI, Justine REYNAUD, Bruno CRÉMILLEUX, Patrice BOIZUMAULT et Aymeric BEAUCHAMP : Extraction de motifs d'intervalles fermés en utilisant la programmation par contraintes. *In JFPC*, 2023.

[3] Mehdi KAYTOUE, Sergei KUZNETSOV et Amedeo NAPOLI : Revisiting numerical pattern mining with formal concept analysis. *IJCAI*, 2011.

[4] N. LAZAAR, Y. LEBBAH, S. LOUDNI, M. MAAMAR, V. LEMIÈRE, C. BESSIERE et P. BOIZUMAULT : A global constraint for closed frequent pattern mining. *In CP*, 2010.

[5] L. De RAEDT, T. GUNS et S. NIJSSEN : Constraint programming for data mining and machine learning. *In AAAI*, 2010.


2. $x_m$ désigne à la fois $\underline{x}_m$ et $\overline{x}_m$
3. $+\!\!+$ désigne un opérateur de concaténation