Consommation mémoire et puissance de calcul en fouille de motifs graduels basée sur les ordres flous multi-précisions

Perfecto Malaquias Quintero Flores, Federico del Razo Lopez, Nicolas Sicard, Anne Laurent

Consommation mémoire et puissance de calcul en fouille de motifs graduels basée sur les ordres flous multi-précisions

Memory Consumption and Computing Power in Gradual Pattern Mining Based on Multi-Precision Fuzzy Orderings

M. Quintero Flores[1] F. Del Razo[2] Nicolas Sicard[3] Anne Laurent[1]

[1] Université Montpellier 2 CNRS - LIRMM France

[2] I.T. Toluca Mx

[3] EFREI - AllianSTIC, Paris, France

161 rue Ada, 34095 Montpellier, France, quinterofl@lirmm.fr

Av. Instituto Tecnologico Toluca S/N, Mexico, delrazo@ittoluca.edu.mx

30-32 av. de la République, 94800 Villejuif Cedex, Paris, France, nicolas.sicard@efrei.fr

161 rue Ada, 34095 Montpellier, France, laurent@lirmm.fr

Résumé :

Dans cet article, nous proposons un cadre pour traiter deux grands problèmes lors de l'extraction de motifs graduels basée sur les ordres flous et sur le coefficient de corrélation de rang gamma flou. Les problématiques abordées sont i) la consommation mémoire et ii) la précision, la représentation, et le stockage efficace des degrés de concordance floue de chaque indice paire *(i, j)* par rapport à la perte ou le gain de puissance de calcul. Dans ce contexte, notre approche implique l'utilisation d'une technique dédiée au traitement des matrices creuses (afin de éviter le stockage des valeurs zéro) et une vaste gamme de représentations de précision variable (de 1 à 64 bits).

Mots-clés :

Ordres flous, extraction de motifs graduels, coefficient de corrélation de rang gamma flou.

Abstract:

In this paper we introduce a framework to address two major problems in gradual itemset mining based on *fuzzy orderings* and *fuzzy gamma rank correlation*. The issues addressed are : 1) the high memory consumption, 2) the precision, representation and efficient storage of the fuzzy concordance degrees of each index pair *(i,j)* versus the loss or gain of computing power. In this context, our approach involves the use of a dedicated technique for handling sparse matrices (in order to avoid the storage of zero values) and a wide range of representations of precision from 2 to 64 bits.

Keywords:

Fuzzy orderings, gradual itemsets mining, fuzzy rank correlation measure.

1 Introduction

Les ordres flous et les classements flous de données incertaines sont l'objet d'études importantes. Ceci est dû au fait qu'ils permettent de gérer naturellement l'imprécision, l'ambiguïté et le caractère flous présents dans les problèmes de décision d'alternative floue et de données incertaines [2] [5] [7].

Les classements et ordres flous présentent de nombreux avantages [2] [7] [9] [11] mais ils présentent des défis complexes tels que la représentation et la précision des données floues et une très grande consommation mémoire [5].

Dans le contexte particulier de l'extraction automatique d'itemsets graduels basée sur les ordres flous et le coefficient de corrélation de rang gamma flou, nous proposons une solution pour traiter le problème de consommation mémoire, la représentation, la précision et le stockage efficace des degrés de concordance floue de chaque paire d'indice *(i, j)*.

Cet article est organisé comme suit. Dans la Section 2, nous présentons une revue des mesures de corrélation de rang. Dans la Section 3, nous examinons le coefficient de corrélation de rang gamma flou. Dans la Section 4, nous expliquons notre algorithme d'extraction d'itemsets graduels ainsi que notre proposition en terme de stokage et de précision des données. Nous concluons dans la Section 5.

2 Mesures de corrélation de rang

Etant donné $n \geq 2$ paires d'observations numériques (*X, Y*) qui sont definies par (1), un coefficient de corrélation de rang mesure la force de l'association (corrélation) entre les variables ou les attributs d'une base de données en fonction de leur tendance à augmenter ou diminuer dans le même sens ou dans le sens opposé [1] [3] [5].

$$\{(x_i, y_i)_{i=1}^{n} | x_i \in X \; and \; y_i \in Y\}. \quad (1)$$

Le tau (τ_a) de Kendall (2), le gamma (γ) de Goodman et Kruskal (3) sont des mesures de corrélation de rang qui sont définies en fonction du nombre de paires concordantes (*Cp*) (4) et du nombre de paires discordantes (*Dp*) (6). Etant donnée une paire d'indices *(i, j)*, *i=*(x_i, y_i) et *j=*(x_j, y_j), tel que $_{i=1,\,2,\ldots,n}$, $_{j=1,\,2,\ldots,n}$, $_{i \neq j}$ et $n \neq Cp + Dp$ [3] [6].

$$\tau_a = \frac{Cp - Dp}{\frac{n(n-1)}{2}} \quad (2)$$

$$\gamma = \frac{Cp - Dp}{Cp + Dp} \quad (3)$$

$$Cp = \sum_{i=1}^{n} \sum_{j \neq i} cp\,(i, j) \quad (4)$$

$$cp(i, j) = \begin{cases} 1 & \text{Si } (i, j) \; est \; concordante \\ 0 & \text{sinon} \end{cases} \quad (5)$$

$$Dp = \sum_{i=1}^{n} \sum_{j \neq i} dp\,(i, j) \quad (6)$$

$$dp(i, j) = \begin{cases} 1 & \text{if } (i, j) \; est \; discordante \\ 0 & \text{sinon} \end{cases} \quad (7)$$

Une paire d'indices *cp(i, j)* est concordante si $\{(x_i > x_j)$ et $(y_i > y_j)\}$ | $\{(x_i < x_j)$ et $(y_i < y_j)\}$, une paire d'indices *cp(i, j)* est disconcordante si $\{(x_i > x_j)$ et $(y_i < y_j)\}$ | $\{(x_i < x_j)$ et $(y_i > y_j)\}$. La valeur du coefficient de corrélation de rang se trouve dans l'intervalle [-1, 1].

La corrélation de rang est positive si une augmentation de *X* survient avec une augmentation de *Y* ou si une diminution de *X* survient avec une diminution de *Y*. A l'inverse, la corrélation de rang est negative si une augmentation de *X* survient avec une diminution de *Y*, ou si une diminution de *X* survient avec une augmentation de *Y*. Le coefficient tend vers 0 si *X* et *Y* sont indépendants [5] [7] [5] [6].

3 Coefficient flou de corrélation de rang

Une extension floue du coefficient de corrélation de rangs du *gamma* de Goodman et Kruskal a été proposée par Bodenhofer et Klawonn [1] sur la base des ordres flous et des relations d'équivalences floues. (9) définit le degré auquel la paire d'indice $\tilde{cp}$*(i, j)* est une paire concordante et (10) définit le degré auquel l'indice paire $\tilde{dp}$*(i, j)* est une paire discordante [5].

$$\tilde{Cp} = \sum_{i=1}^{n} \sum_{j \neq i} \tilde{cp}\,(i, j) \quad (8)$$

$$\tilde{cp}(i, j) = \top(R_X(x_i, x_j), R_Y(y_i, y_j)) \quad (9)$$

$$\tilde{dp}(i, j) = \top(R_X(x_i, x_j), R_Y(y_j, y_i)) \quad (10)$$

$$R_X(x_i, x_j) = 1 - L_X(x_j, x_i) \quad (11)$$

$$L_X(x1, x2) = min(1, max(0, 1 - \frac{(x1 - x2)}{r})) \quad (12)$$

$$R_Y(y_i, y_j) = 1 - L_Y(y_j, y_i) \quad (13)$$

$$L_Y(y1, y2) = min(1, max(0, 1 - \frac{(y1 - y2)}{r})) \quad (14)$$

$$\top(a,b) = max(1, a+b-1) \qquad (15)$$

$R_X(x_i, x_j)$ est un strict $T_L{-}E_X{-}ordre$ sur *X* qui est défini dans (11), $R_Y(y_i, y_j)$est un strict $T_L{-}E_Y{-}ordre$ sur *Y* défini dans (13), $L_X(x_i, x_j)$ est un fortement complet $T_L{-}E_r{-}ordre$ sur *X* qui est défini dans (12), et $L_Y(y_i, y_j)$ est un fortement complet $T_L{-}E_r{-}ordre$ sur *Y* qui est défini dans (14) [5] [11].

$\top(R_X(x_i, x_j), R_Y(y_i, y_j))$ est une $\top{-}relation$ *d'équivalence* et représente une *t-norm* de Lukasiewicz qui est défini dans (15) pour $a = R_X(x_i, x_j)$ et $b = R_Y(y_i, y_j)$ [1] [5].

4 Extraction des itemsets graduels en utilisant les ordres flous

Pour traiter le problème de l'extraction des itemsets graduels fréquents à partir de données imprécises et incertaines, nous proposons un algorithme basé sur les principes du coefficient de corrélation de rangs Kendall *tau*, du coefficient *gamma* de corrélation de rangs de Goodman et Kruskal, du coefficient *gamma* flou de corrélation de rangs de Bodenhofer et Klawonn, et enfin sur le concept de matrice binaire de couples concordants.

Dans le cadre de notre approche, nous proposons les définitions suivantes d'un *item graduel*, d'un *itemset graduel*, d'un *couple concordant* et du *support* d'un *itemset graduel*.

Soit db une base de données constituée d'un ensemble d'objets (*n-uplets*) **O**=$\{o_1, o_2, ..., o_n\}$, où chaque valeur o_i est définie sur un attribut A_l pris dans un ensemble d'attributs numérique (*m-attributs*) **A**=$\{A_1, A_2, ..., A_m\}$ dont les domaines sont muni d'un ordre.

Un *iG* (*item graduel*) est défini comme la variation **v**$\in\{\geq|\leq\}$ associée aux valeurs d'un attribut $A_l \in db$ et notée A_l**v**. **v** est dit en ordre ascendant ($\geq$) si les valeurs de l'attribut ont tendance à augmenter, **v** est dit en ordre descendant ($\leq$) si les valeurs de l'attribut ont tendance à diminuer, *i.e.*, $\{A_l\geq\} \simeq \{A_l(o_i)<A_l(o_j)\}$ et $\{A_l\leq\} \simeq \{A_l(o_i)>A_l(o_j)\}$ pour *i=1, 2, ..., n*, *j=i+1, ..., n*, $i\neq j$ et $l\in\{1,2, ..., k\}$ [6] [8].

Un *IG* (itemset graduel) est une combinaison de deux ou plusieurs des *items graduels* de la forme *IG*=$\{A_1\leq A_2\leq A_3\geq\}$, interprétée comme $\{l'A_1\ diminue, l'A_2\ diminue, l'A_3\ augmente\}$. La taille (*k*) d'un *IG* est définie comme le nombre d'*items graduels* contenus dans l'*IG*, tel que $k\in\{2, 3, 4,..., m\}$. Chaque *item graduel* $\in$*IG* est unique [6] [8].

Un *couple concordant (cc)* est une paire indice où les objets (o_i, o_j) satisfont toutes les variations **v** exprimées par les *items graduels* impliqués dans un *IG* donné de taille *k*. Soit *IG*=$\{A_1\leq A_2\leq\}$ de taille *k=2*, une paire indice *cc(i, j)* est un *couple concordant* si $((A_1(o_i)>A_1(o_j)\ implies\ A_2(o_i)>A_2(o_j))$, où *i*=$(A_1(o_i),A_2(o_i))$ et *j*=$(A_1(o_j),A_2(o_j))$ [5] [6].

Un *IG* est un motif intéressant si *support(IG)* est supérieur ou égal au support minimal défini par l'utilisateur et nommé *seuil minimal* [4].
Dans la littérature, il existe différentes méthodes pour calculer le *support* des *itemsets graduels*. Celles-ci diffèrent dans l'interprétation du concept de dépendance graduelle implicite dans les itemsets graduels. Nous avons opté pour l'interprétation basée sur le *classement induit par la corrélation* et le concept de *couple concordant* [5] [6], où le *support* d'un *IG* est calculé comme :

$$support(IG) = \frac{\sum_{i=1}^{n}\sum_{j\neq i} cc\,(i,j)}{n(n-1)} \qquad (16)$$

Nous proposons de calculer chaque *couple concordant* $cc(i, j)$ comme le niveau auquel la paire indice $\tilde{cp}(i, j)$ est considérée comme une paire concordante, défini par (8) dans le cadre du *coefficient* ***gamma*** *flou de corrélation de*

rangs. Nous proposons également d'utiliser le concept de *matrice de degrés de concordance floue* $\tilde{cp}(i,j)$ pour stocker les degrés de concordance floue de chaque paire indice $\tilde{cp}(i,j)$) des itemsets graduels.

Plus précisément, l'algorithme 1 illustre notre approche de l'extraction d'itemsets graduels qui consiste en trois phases principales :

- **Phase 1** - pour chaque attribut $A_l \in db$, construire ses items graduels : $\{A_l\geq, A_l\leq\}$,
- **Phase 2** - générer les itemsets graduels fréquents au niveau $k = 2$ à partir de l'ensemble des items graduels *iG*,
- **Phase 3** - générer les itemsets graduels fréquents au niveau $k > 2$ à partir des itemsets graduels fréquents au niveau $(k-1)$,

Dans les phases 2 et 3, le concept de matrice de degrés de concordance floue joue un rôle important car le support de chaque *itemset* est calculé à partir des informations contenues dans chaque matrice. Dans le reste de l'article, nous présentons la mise en œuvre de telles matrices.

4.1 Mise en œuvre des matrices de degrés de concordance floue

Dans le cadre de la mise en œuvre des matrices de degrés de concordance floue $\tilde{cp}(i,j)$, nous avons pris en compte deux aspects importants qui sont i) le problème de consommation de mémoire et ii) la représentation et la précision des degrés de concordance floue $\tilde{cp}(i,j)$.

Afin de réduire la consommation de mémoire, nous avons représenté et stocké chaque matrice de degrés de concordance floue selon le format Yale de matrices creuses *(Yale Sparse Matrix Format)* afin de ne retenir que les coefficients non nuls [10].
Comme les candidats itemsets sont générés à partir des $k-$*itemsets* graduals fréquents, seules les matrices d'itemsets graduels fréquents de niveau (*k-1*) sont conservées en mémoire et utilisées pour générer les matrices des candidats itemsets graduels du niveau *k*. Si le *sup-*

Algorithm 1: Fuzzy Orderings-based Gradual Itemset Mining

Data: db (Database), # m (Attributes),
Data: # n (Records), $minimum\_threshold$.
Result: Frequent Gradual Itemsets $\mathcal{F}_k$
$\mathcal{F}_k \leftarrow \emptyset$; $gI \leftarrow \emptyset$;
foreach $attributeA_l \in db$ **do**
  /*build their respective gradual items :*/;
  $gI = gI + \{A_l\geq, A_l\leq\}$;
/* Candidate gradual itemset generation at level $k = 2$, of the form :*/;
$\mathcal{C} = \{A_1\geq A_2\geq, A_1\geq A_2\leq, ..., A_{m-1}\leq A_m\leq\}$;
foreach $gradual\ itemset\ candidate \in \mathcal{C}$ **do**
  $Compute\ their\ matrices\ of\ concordance\ degrees\ \tilde{cp}(i,j)\ as\ in\ (9)$
  $Compute\ their\ support,\ as\ the\ sum\ of\ their\ matrices\ of\ concordance\ degrees\ as\ in\ (8)\ divided\ by\ n(n-1)$;
  if $support(candidate) \geq minimum\_threshold$ **then**
    $\mathcal{F}_k \leftarrow \mathcal{F}_k \cup \{candidate\}$;
  Else $delet(candidate\ and\ matrix)$;
$k++$;
repeat
  /*Frequent gradual itemset mining,$k > 2$*/;
  $\mathcal{I} = GenItemset1From\{\mathcal{F}_{k-1}\}$;
  $\mathcal{J} = GenItemset2From\{\mathcal{F}_{k-1}\}$;
  $q = 1$;
  foreach $\{\mathcal{I}, \mathcal{J}\} \in \mathcal{F}_{k-1}$ **do**
    $Compute\ matrix\ of\ concordance\ degrees\ of\ candidate\ \mathcal{C}_{k,q}.M\ as:$
    $\mathcal{C}_{k,q}.M = \top - norm(\mathcal{I}.M, \mathcal{J}.M)$;
    /* $\mathcal{I}.M, \mathcal{J}.M$ are matrices of concordance degrees of itemsets $\mathcal{I}, \mathcal{J}$*/
    $support(\mathcal{C}_{k,q}) = sumMatrix(\mathcal{C}_{k,q}.M)/n(n-1)$;
    if $support(\mathcal{C}_{k,q}) \geq minimum\_threshold$ **then**
      $\mathcal{F}_k \leftarrow \mathcal{F}_k \cup \{\mathcal{C}_{k,q}\}$;
    Else $delet(candidate\ and\ matrix)$;
  $Delet(\mathcal{F}_{k-1}\ and\ Matrices)$;
  k++ ; $q++$;
until $\mathcal{F}_{FGP}$ *does not grow any more*;

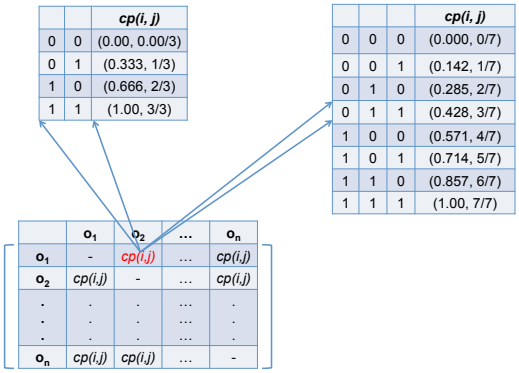

Figure 1 – Illustration de la matrice des degrés de concordance floue représentée avec une précision de 2 et 3 bits.

port d'un *itemset graduel candidat* ($\mathcal{C}_{k,q}$) est inférieur au *seuil minimal* le $\mathcal{C}_{k,q}$ est élagué et sa matrice de degrés de concordance floue $\tilde{cp}(i,j)$ est éliminée.

4.2 Analyse de la précision des matrices de degrés de concordance floue

Pour résoudre le problème de la représentation, de la précision et du stockage efficace des degrés de concordance floue $\tilde{cp}(i,j)$, on s'intéresse aux besoins de stockage dans le cas binaire et dans le cas flou.

Dans le cas binaire, nous utilisons une précision d'un bit, **i.e.** pour chaque $\tilde{cp}(i,j) \in \{0,1\}$, un bit suffit pour représenter et stocker la valeur $\{0|1\}$.

Dans le cas flou, chaque $\tilde{cp}(i,j) \rightarrow [0,1]$ est représenté avec une précision de 2, 3, 4, *etc.* jusqu'a 64 bits. Ainsi nous pouvons représenter 4 valeurs floues avec une précision de 2 bits, 8 valeurs floues avec une précision de 3 bits, 2^8 valeurs floues avec 8 bits et ainsi de suite.

La Figure 1 illustre la structure de la matrice de degrés de concordance floue, où chaque $\tilde{cp}(i,j) \rightarrow [0,1]$ est représenté avec une précision de 2 ou 3 bits.

La Figure 2 montre la représentation d'une

				cp(i, j)
0	0	0	0	(0.00, 0/15)
0	0	0	1	(0.066, 1/15)
0	0	1	0	(0.133, 2/15)
0	0	1	1	(0.200, 3/15)
0	1	0	0	(0.266, 4/15)
0	1	0	1	(0.333, 5/15)
0	1	1	0	(0.400, 6/15)
0	1	1	1	(0.466, 7/15)
1	0	0	0	(0.533, 8/15)
1	0	0	1	(0.600, 9/15)
1	0	1	0	(0.666, 10/15)
1	0	1	1	(0.733, 11/15)
1	1	0	0	(0.800, 12/15)
1	1	0	1	(0.866, 13/15)
1	1	1	0	(0.933, 14/15)
1	1	1	1	(0.00, 15/15)

Figure 2 – Illustration de degrés de concordance floue $\tilde{cp}(i,j) \rightarrow [0,1]$, représenté avec une précision de 4-bits.

précision de 4 bits, où nous pouvons représenter jusqu'à 16 valeurs floues possibles.

5 Conclusion

Dans cet article, nous avons introduit un cadre pour répondre à deux problèmes majeurs dans l'extraction d'itemsets graduels, basée sur les *ordres flous* et sur la *corrélation de rang gamma floue*. Les questions abordées ont été la consommation mémoire, la représentation, la précision et le stockage efficace des degrés de concordance floue *(i,j)* par rapport à la perte ou le gain de puissance de calcul. Nous avons proposé l'utilisation d'une technique spécifique à la manipulation de matrices creuses (afin d'éviter le stockage de valeurs nulles), nous avons présenté un cadre pour aborder la précision, la représentation et stockage efficient des degrés de concordance floue de chaque paire d'indice *(i,j)* basée sur une large gamme de représentations de précision à partir de 2 bits jusqu'à 64 bits.

Nous préparons aujourd'hui une étude

expérimentale complète afin i) d'étudier les avantages et les inconvénients d'une augmentation ou d'une diminution du nombre de bits de précision et ii) de déterminer les critères à prendre en compte pour obtenir la combinaison optimale mêlant précision des valeurs floues, consommation mémoire et consommation CPU maîtrisées.

Références

[1] U. Bodenhofer and F. Klawonn. *Roboust rank correlation coefficients on the basis of fuzzy orderings : Initial steps,* in Mathware & Soft Computing 15, 5-20, 2008.

[2] U. Bodenhofer, *Fuzzy orderings of fuzzy sets,* in Proc. 10th IFSA World Congress, Istanbul, pp. 500-5007, 2003.

[3] T. Calders, B. Goethais, and S. Jarszewicz. *Mining Rank-Correlated Sets of Numerical Attributes,* in Proc. of the KDD'06, August 20-23, 2006, ACM, 2006.

[4] E. Hüllermeier, *Association rules for expressing gradual dependencies,* in Proceedings PKDD 2002 Lecture Notes in Computer Science 2431, pp. 200-211, 2002.

[5] H-W. Koh and E. Hullermeier, *Mining gradual dependencies based on fuzzy rank correlation,* in Combining Soft Computing and Statistical Methods in Data Analysis. Advances in Intelligent and Soft Computing, Vol. 77, Springer Heidelberg, pp. 379-386, 2010.

[6] A. Laurent, M.-J. Lesot, and M. Rifqi, *GRAANK : Exploiting rank correlations for extracting gradual itemsets,* in Proc. of the Eighth International Conference on Flexible Query Answering Systems (FQAS'09), LNAI 5822, Springer-Verlang Berlin Heidelberg, pp. 382-393, 2009.

[7] P. L. Nancy, and C. Hao-en, *Fuzzy Correlation Rules Mining,* in Proceedings of the 6th WSEAS International Conference on Applied Computer Science, Hangzhou, China, pp 13-18, April 15-17, 2007.

[8] M. Quintero, A. Laurent, P. Poncelet, *Fuzzy Ordering for Fuzzy Gradual Patterns,* in FQAS 2011, LNAI 7022, Springer-Verlag Berlin Heidelberg, pp. 330-341, 2011.

[9] M.D. Ruiz and E. Hüllermeier. *A formal empirical analysis of the fuzzy gamma rank correlation coefficient,* article in press in Information Sciences Elsevier xxx, xxx-xxx, 2012.

[10] D. Wilhelm, *Sparse Matrix Formats,* in ECE 250 Algorithms and Data Structures, Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada, 2006-2011.

[11] L. A. Zadeh, *Similarity Relations and Fuzzy Orderings,* in Information Sciences. Volume 3, Issue 2, pp. 177 - 200, April 1971.