Motifs séquentiels et écarts flous

Paola Salle, Sandra Bringay, Anne Laurent, Maguelonne Teisseire

Motifs séquentiels et écarts flous

Handling Fuzzy Gaps in Sequential Patterns

Paola Salle[1], Sandra Bringay[1,2], Anne Laurent[1] et Maguelonne Teisseire[3]

[1] LIRMM-Univ. Montpellier 2-CNRS UMR 5506 161 rue Ada, 34095 Montpellier
[2] Dpt MIAP- Univ. Montpellier 3 Route de Mende, 34199 Montpellier Cedex 5
[3] Cemagref, UMR TETIS Maison de la télédétection 500, rue J.F.Breton 34093 Montpellier Cedex 5

Résumé :

Fouiller des données numériques pour découvrir de nouvelles connaissances est une tâche non triviale très étudiée depuis quelques années. Dans notre travail, nous nous intéressons à des données numériques qui peuvent être issues par exemple de biotechnologies comme les puces ADN. Elles permettent aux biologistes de découvrir de nouvelles corrélations de gènes pour mieux comprendre certaines maladies comme le cancer. Dans ce contexte, des motifs séquentiels du type $\langle(Gene1 \ \ Gene5)(Gene2)\rangle$ ont été découverts signifiant que les gènes Gene1 et Gene5 ont le même niveau d'expression alors que le gène Gene2 a une expression plus élevée. Cependant il est difficile de mesurer les écarts entre les éléments de la base de données. Nous proposons donc, dans cet article, d'utiliser une partition floue fournie par les experts pour définir des motifs pouvant servir à caractériser plus finement ces écarts.

Mots-clés :

Fouille de données, motifs séquentiels, partition floue.

Abstract:

Dealing with numerical data for mining novel knowledge is a non trivial task that has received much attention in the last years. In our work, we focus on biological data from DNA chips that biologists study in order to discover new gene correlations that could help understanding diseases like cancer. In previous work, sequential patterns like $\langle(Gene1 \ \ Gene5)(Gene2)\rangle$ have been discovered, meaning that genes Gene1 and Gene5 have the same expression level followed by gene Gene2 that has a higher expression value. However it is difficult to measure differences between elements of the database. We therefore propose in this article, using a fuzzy partition provided by the experts to identify patterns that can be used to characterize more precisely these differences.

1 Introduction

De plus en plus de données numériques sont récoltées et historisées au sein de bases de données en vue de leur exploitation à des fins d'analyse. Le traitement de ces grandes masses de données en tenant compte de la sémantique des résultats est donc un véritable défi pour les chercheurs. Dans ce contexte, les approches de fouille de données ont un rôle clé à jouer puisqu'elles permettent de découvrir des connaissances.

Dans cet article, nous nous focalisons sur la découverte de motifs à partir de gros volumes de données numériques pour aider les experts à analyser les comportements fréquents au sein de leurs bases de données. En particulier, nous travaillons sur les données issues de l'analyse de puces ADN. Grâce à ces technologies, il est possible d'acquérir des milliers de données sur un objet à un instant précis. Dans un travail précédent, nous avons proposé un algorithme appelé MDS [SBT09] pour extraire des motifs séquentiels à partir de ces données. Nous étudions l'apport de la logique floue dans le but d'obtenir des motifs séquentiels plus compréhensibles pour les experts. Pour cela, nous fournissons une information supplémentaire en apportant une description linguistique des écarts fréquemment identifiés entre les éléments de la base de données. Les écarts d'expression sont définis en utilisant une partition floue décrivant *des écarts flous*. Nous donnons ci-dessous les définitions des "écarts flous" et les algorithmes associés. Nous présentons les expérimentations en montrant plusieurs exemples de motifs obtenus grâce à notre approche et qui n'auraient pas

été découverts en utilisant des intervalles stricts.

2 Définition : Motifs séquentiels

Dans cette section, nous présentons les informations relatives aux données que nous souhaitons fouiller et les motifs séquentiels qui sont la base de notre approche.

Les motifs séquentiels mettent en évidence des corrélations entre les enregistrements d'une base de données comme par exemple, une relation temporelle. Dans cet article, nous utilisons les motifs séquentiels avec des écarts flous pour donner aux experts une lecture des séquences plus précises. Les motifs séquentiels décrivent des séquences fréquentes d'itemsets. Un itemset correspond à un ensemble d'items qui apparaissent en même temps dans un enregistrement de la base de données. Par exemple, lorsque nous considérons une base de données d'un supermarché, celle-ci décrit les items achetés par des clients à des dates différentes. Les motifs fréquents extraits à partir de cette base de données sont $\langle (beurre \quad chips)(chocolat)\rangle sup$ signifiant que $sup\%$ des clients ont acheté du beurre avec des chips et *ensuite* du chocolat.

Définition 1 (Itemset) *Soit $I = \{i_1, i_2, ..., i_m\}$ un ensemble d'items, un itemset est un ensemble non vide d'items $(i_1 i_2 \dots i_k) \subseteq I$.*

Notons que dans une séquence, les itemsets sont ordonnés et qu'un itemset est un ensemble, il est donc non-ordonné.

Définition 2 (Séquence) *Une séquence s est une liste non vide d'itemsets, notée par $\langle it_1 it_2 ... it_p\rangle$. Une séquence est aussi appelée une* n-sequence *(ou une séquence de taille n) si elle contient n items.*

Définition 3 (Base de données) *Soit $\mathcal{R}$ un ensemble d'enregistrements et chaque enregistrement est composé de trois informations : l'identifiant de l'enregistrement, une date et un ensemble d'items apparaissant dans l'enregistrement. $\mathcal{R}$ est appelée la base de données.*

Puce_id	Niveaux d'expression	Gènes
1	-6.2	Gene2
	-1.8	Gene1, Gene5
	2.3	Gene3
	4.8	Gene4
2	-5.4	Gene2
	-2	Gene1, Gene5
	2.3	Gene3
	4.8	Gene4
3	-5.8	Gene2
	-3.6	Gene1, Gene5
	2.3	Gene3
	7.5	Gene4
4	-4.7	Gene2
	0	Gene4
	2.3	Gene5
	8.57	Gene3, Gene1

Tableau 1 – Exemple de base de données

Par exemple, dans le cas des données issues de l'analyse des puces ADN, l'id de l'enregistrement est une expérimentation (Puce_id), les enregistrements correspondent à des gènes et les dates à des niveaux d'expression. Le tableau 1 montre cette base de données, les gènes ayant été ordonnés selon leurs niveaux d'expressions.

Exemple 1 *En reprenant le tableau 1, $Gene1$, $Gene2$, $Gene3$, $Gene4$ et $Gene5$ sont des items. $(Gene1\ Gene5)$ est un itemset, et $s = \langle (Gene2)\ (Gene1\ Gene5)\ (Gene3)\ (Gene4)\rangle$ est une séquence.*

Les bases de données peuvent être vues comme un ensemble de séquences, comme le montre le tableau 2. Tous les enregistrements à partir du même objet sont ensuite regroupés ensemble et triés dans l'ordre croissant selon leurs

Puce_id	Séquence d'expression de gènes
1	⟨(Gene2)(Gene1 Gene5)(Gene3)(Gene4) ⟩
2	⟨(Gene2)(Gene1 Gene5)(Gene3)(Gene4) ⟩
3	⟨(Gene2)(Gene1 Gene5)(Gene3)(Gene4) ⟩
4	⟨(Gene2)(Gene4)(Gene5)(Gene3 Gene1) ⟩

Tableau 2 – Exemple d'une base de données de séquences

dates. Ils constituent les séquences de données. Une séquence est dite fréquente si la condition $freq(s) \geq minFreq$ est respectée, avec $minFreq$ une valeur de fréquence minimale spécifiée par l'utilisateur. Le problème de découverte de motifs séquentiels fréquents est décrit comme la découverte de toutes les séquences qui apparaissent pour plus de $minsup$ éléments, où $minsup$ est appelé le support minimal et est fourni par l'utilisateur. L'ensemble des motifs séquentiels contient toutes les séquences fréquentes maximales[1]. Les séquences sont découvertes lorsque les intervalles de temps apparaissent entre deux itemsets (par exemple, des achats).

Récemment, les contraintes de temps ont été introduites dans le but de gérer les interrelations d'une manière plus précise [SA96]. Par exemple, ces contraintes permettent de ne pas considérer les séquences qui ont un intervalle plus grand qu'un certain nombre de jours fixé ou au contraire, deux items sont considérés comme étant simultanés lorsqu'ils arrivent à des dates proches. Les contraintes floues ont été introduites par [FLT07] mais elles ne permettent pas de décrire l'écart qui sépare les itemsets. Dans cet article, nous proposons une approche pour extraire des motifs ordonnés flous en intégrant des intervalles flous pour décrire les écarts.

3 Motifs séquentiels et écarts flous

Dans cette section, nous détaillons notre contribution, ayant pour but de proposer une méthode efficace et pertinente pour extraire des motifs incluant des écarts flous. Pour ce faire, nous considérons une partition floue proposée par l'expert [JKZ73]. Cette partition est donnée comme un ensemble de n ensembles flous $\mathcal{A} = \{A_1, \ldots, A_n\}$ définis sur l'univers U (par exemple, les valeurs qui peuvent caractériser la différence entre deux valeurs d'expression de gènes). Cette partition floue est définie par $\forall u \in U, \sum_{i=1}^{n} \mu_i(u) = 1$ où μ_i est la fonction d'appartenance associée au sous-ensemble flou A_i. La figure 1 décrit un exemple de cette partition.

En se basant sur cette partition, il est possible de décrire linguistiquement les écarts *peu*, *moyennement* ou *beaucoup* exprimés, etc.

3.1 Ecart flou

Dans cette section, nous détaillons comment obtenir des écarts flous à partir d'une partition d'intervalles. Nous rappelons ici que les motifs séquentiels extraits sont du type $\langle(Gene2)\ (Gene1\ Gene5)\ (Gene3)\ (Gene4)\rangle$ ce qui signifie que *fréquemment, le gène Gene2 a une expression inférieure aux gènes Gene1 et Gene5 qui ont la même expression mais qui est inférieure à l'expression du gène Gene3 et du gène Gene4.*

Exemple 2 *Considérons les données du tableau 1, pour 3 puces parmi les 4 (1, 2 et 3), l'expression du gène Gene2 est plus faible que l'expression du gène Gene1 et Gene5. Avec MDS, nous aurions donc découvert le motif* $\langle(Gene2)\ (Gene1\ Gene5)\rangle$. *Ce motif permet uniquement de capturer la notion de sur expression entre deux itemsets. Or ce motif peut correspondre à différents cas d'expression de gènes. Dans le premier cas, l'écart est important alors que dans le dernier cas l'écart est très petit. Pour les biologistes, ces trois cas n'ont pas la même signification et par conséquent les motifs séquentiels ne sont pas suffisants (cf. Figure 2). C'est pour cette raison que nous propo-*

[1] en terme de nombre d'items

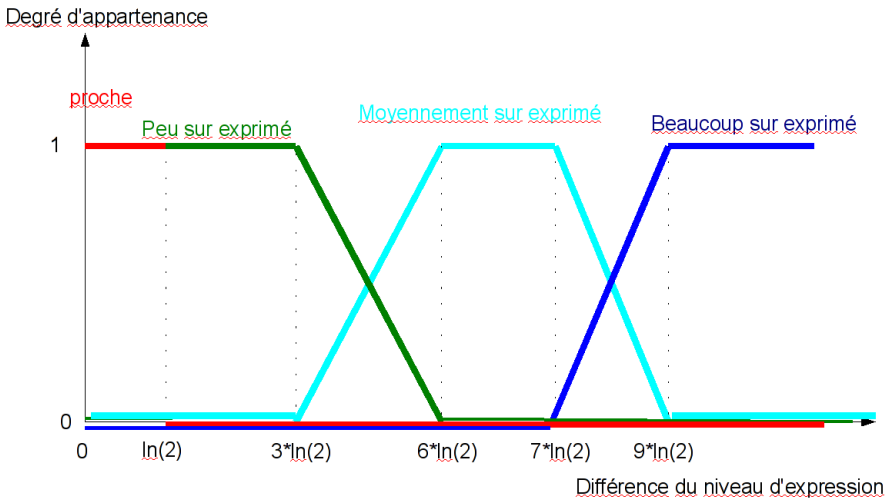

Figure 1 – Partition floue pour les différences de niveaux d'expression pour les valeurs de deux objets

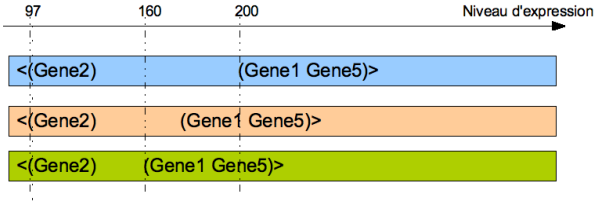

Figure 2 – Différence d'expression entre deux itemsets

sons d'extraire des motifs à écarts flous, qui expriment l'importance de l'écart entre les itemsets.

Chaque écart entre les itemsets d'une séquence est associé à un degré pour lequel la relation floue est respectée (e.g. peu, moyennement, beaucoup). Cette description des écarts est agrégée au niveau de la séquence et est utilisée pour calculer son support.

Nous présentons ci-dessous l'approche formelle proposée et les définitions associées. Nous rappelons que notre approche a pour but de fouiller des bases de données définies dans la Définition 3. Elles contiennent des objets, et chaque objet est une séquence d'itemsets. D'abord, nous définissons comment peuvent être obtenus les intervalles stricts.

Définition 4 (Différence d'itemsets) *Soient $\mathcal{R}$ une base de données, o un enregistrement de $\mathcal{R}$, et it_1 et it_2 deux itemsets de la séquence associée à o. La différence $\delta(it_2, it_1)$ entre it_1 et it_2 est définie comme la valeur absolue de la différence entre les expressions du premier item de it_2 et du dernier item de it_1.*

Exemple 3 *Dans le tableau 1, la différence $\delta(it_2, it_1)$ entre les itemsets $it1 = (Gene2)$ et $it2 = (Gene1)$ est égale à et $|-1.8-(-6.2)| = 4.4$ pour l'objet 1. En revanche pour l'objet 2, la différence entre les itemsets avec $it2 = (Gene1\ Gene5)$ est égale à $|-2-(-5.4)| = 3.4$.*

Dans notre exemple lié aux puces ADN, le calcul de la différence revient donc à calculer la différence entre les deux niveaux d'expression.

Définition 5 (Validité d'une séquence) *Une séquence est valide pour un objet si elle est vérifiée par cet objet.*

Exemple 4 *La séquence* $\langle(Gene1)(Gene3)\rangle$ *est valide pour les puces 1, 2 et 3 mais pas pour la puce 4.*

Définition 6 (Séquence à écarts flous) *Soit* $\mathcal{A}$ *une partition floue, une séquence à écart flou* s_{FG} *est définie comme une liste d'itemsets et des ensembles flous tels que* $s_{FG} = \langle it_1(A_{i_1})it_2(A_{i_2})...(A_{i_{p-1}})it_p\rangle$ *où* $\forall i \in [1, p-1]$, $A_i \in \mathcal{A}$.

Définition 7 (Degré d'appartenance flou) *Soient* s_o *une séquence de données associée à un objet* o, it_1 *et* it_2 *deux itemsets et* A *un ensemble flou associé à une fonction d'appartenance* μ. *Un degré d'intervalle flou entre* it_1 *et* it_2 *est une paire* (A, d) *et nous avons* $\langle it_1(A, d)it_2\rangle$ *où* $d = \mu(\delta(it_2, it_1))$

Exemple 5 *Nous considérons l'ensemble flou* $beaucoup$. *Pour la puce* 1, *nous avons le degré d'intervalle flou* $\langle(Gene1\ \ Gene5)(beaucoup, 1)(Gene3)\rangle$ *puisque* $\delta(it_2, it_1) = 4.1$ *et* $\mu_{beaucoup}(4.1) = 1$.
Pour la puce 2, *nous avons* $\langle(Gene1\ \ Gene5)(beaucoup, 1)(Gene3)\rangle$ *puisque* $\delta(it_2, it_1) = 4.3$ *et* $\mu_{beaucoup}(4.3) = 1$.
Pour la puce 3, *nous avons* $\langle(Gene1\ \ Gene5)(beaucoup, 0.66)(Gene3)\rangle$ *puisque* $\delta(it_2, it_1) = 5.3$ *et* $\mu_{beaucoup}(5.3) = 0.66$.
Pour la puce 4, *la séquence* $\langle(Gene1\ \ Gene5)(Gene3)\rangle$ *n'est pas valide puisque* $Gene1$ *apparaît à un niveau d'expression plus élevé que* $Gene5$.

Afin de rendre compte du fait que tous les intervalles doivent être valides pour qu'une séquence le soit, le degré d'une séquence de données est calculé en utilisant une t-norme pour fusionner les degrés d'appartenance des écarts flous. Notons que la t-norme $\overline{\top}$ utilisée est généralisée en une fonction n-aire.

Définition 8 (Degré d'une séquence à écarts flous) *Soit* s_o *une séquence de données associée à un objet* o, *une séquence de degré à écarts flous* s_{FGD} *est définie par* $s_{FGD} = \langle it_1(A_{i_1}, d_1)it_2(A_{i_2}, d_2)...(A_{i_{p-1}}, d_{p-1})it_p\rangle$
où (A_i, d_i) *sont des degrés d'appartenance des écarts flous. Le degré de* s_o *noté par* $F_{s_{FG}}(s_o)$ *est calculé de la manière suivante :* $F_{s_{FG}}(s_o) = \overline{\top}\,(d_1, ..., d_{p-1})$

Remarque. Dans notre approche, la t-norme utilisée est le min car nous ne voulons pas pénaliser les longues séquences qui pourraient voir leurs degrés décroître fortement avec une t-norme non idempotente.

Finalement, le support d'une séquence à intervalle flou est calculé en sommant les degrés d'appartenance.

Définition 9 (Support d'une séquence à écarts flous) *Soit* s_{FG} *une séquence à écarts flous, soit* $\mathcal{O}$ *l'ensemble des objets de la base de données, le support* s_{FG} *défini par :*

$$Freq(s_{FG}) = \frac{\sum_{o \in \mathcal{O}} \left[F_{s_{FG}}(s_o)\right]}{|\mathcal{O}|}$$

Exemple 6 *Dans notre exemple, la séquence* $\langle(Gene1\ Gene5)_{beaucoup}(Gene3)\rangle$ *a un support de* $\frac{1+1+0.66+0}{4} = \frac{2.66}{4} = 0.66$.

Considérant une base de données avec des enregistrements d'objets et les motifs séquentiels associés, le problème de recherche de séquences à écarts flous est de trouver toutes les séquences à écarts flous maximales pour lesquelles le support est supérieur à un seuil spécifié par l'utilisateur ($MinFGFreq$).

Entrée : $\mathcal{R}$, une base d'enregistrements
$\mathcal{MS}$, l'ensemble des motifs séquentiels extraits à partir de $\mathcal{R}$
$MinFGFreq$, le seuil de support minimal défini par l'utilisateur
$\mathcal{A}$ l'ensemble des ensembles flous construits à partir de la partition floue
Sortie : $\mathcal{S}_{\mathcal{FG}}$, l'ensemble des motifs séquentiels flous

$\mathcal{S}_{\mathcal{FG}} \leftarrow \varnothing$
FuzzyGapSeq $\leftarrow FuzzyGapGen(\mathcal{MS}, \mathcal{A})$

/* Génération de toutes les séquences à écarts flous associées aux sous-ensembles flous */
pour chaque $s_{FG} \in FuzzyGapSeq$ **faire**
  $Freq_{s_{FG}} \leftarrow \varnothing$
fin

pour chaque $o \in \mathcal{R}$ **faire**
  pour chaque $s_{FG} \in FuzzyGapSeq$ **faire**
    $Freq_{s_{FG}} \leftarrow Freq_{s_{FG}} + F_{s_{FG}}(s_o)$
    /* La fréquence est calculé en additionnant le degré de chaque objet de la séquence de données.*/
  fin
fin

pour chaque $s_{FG} \in FuzzyGapSeq$ **faire**
  si $(Freq_{s_{FG}}/|\mathcal{R}|) \geq MinFGFreq$ **alors**
    $\mathcal{S}_{\mathcal{FG}} \leftarrow \mathcal{S}_{\mathcal{FG}} \cup \{s_{FG}\}$
  fin
fin

Algorithme 1: Extraction de séquences d'écarts flous (ESEF)

3.2 Algorithmes

Dans cette section, nous détaillons les algorithmes que nous proposons pour rechercher des séquences à écarts flous.

Nous extrayons les motifs séquentiels avec un algorithme tel que MDS. Les motifs séquentiels sont calculés en utilisant une approche par niveau, les séquences de taille k étant alors calculées à partir des séquences fréquentes de taille $k-1$. Cette approche nous permet de passer à l'échelle lorsque les bases de données sont très grandes.

A partir de ces motifs séquentiels, la fonction FuzzyGapGen (Algorithme 2) génère l'ensemble de séquences FuzzyGapSeq avec toutes les combinaisons possibles des ensembles flous pour décrire les écarts entre les itemsets. Pour chaque séquence de FuzzyGapSeq, l'algorithme ESEF (Algorithme 1), calcule le support dans la base et ne conserve que les séquences ayant un support supérieur au seuil MinFGFreq. On obtient ainsi un ensemble de motifs séquentiels fréquents dont les écarts entre les itemsets sont associés à des descriptions linguistiques.
La complexité de notre approche est la suivante : soient o le nombre d'objets, s le nombre de séquences, l la longueur maximale des séquences et e le nombre de catégories d'écarts. Avec l'algorithme 2, nous générons l'ensemble $\mathcal{S}_{\mathcal{FG}}$ correspondant à s*l*e séquences à écarts flous. Ensuite, nous vérifions la fréquence en reparcourant tous les objets pour chaque séquence contenue dans $\mathcal{S}_{\mathcal{FG}}$. Finalement, la complexité est de o*s*l*e.

4 Résultats expérimentaux

Nous avons appliqué notre approche sur une base de données [WKZ+05] disponibles sur le web[2].

[2]http ://www.ihes.fr/ zinovyev/princmanif2006/

Entrée : $\mathcal{MS}$ un ensemble de motifs séquentiels
$\mathcal{A}$ un ensemble d'ensemble flous
Sortie : $\mathcal{S}_{\mathcal{FG}}$, l'ensemble de toutes les séquences à écarts flous associées à $\mathcal{MS}$ et $\mathcal{A}$
$\mathcal{S}_{\mathcal{FG}} \leftarrow \varnothing$
pour chaque $seq \in \mathcal{MS}$ **faire**
  $n \leftarrow sizeof(seq)$;
  /* Il y a $n-1$ écarts, ce qui revient à $|card(\mathcal{A})|^{n-1}$ séquences à écarts flous */
  /* loc_s est un ensemble de séquences pour seq */
  $loc_s \leftarrow \{\langle it_1 \rangle\}$
  pour *(i=1 ; $i < n$; i++)* **faire**
    pour chaque $s \in loc_s$ **faire**
      $loc_s \leftarrow loc_s - \{s\}$
      /* Nous concaténons chaque séquence à écarts flous à partir de tous les ensembles flous possibles */
      pour chaque $a \in \mathcal{A}$ **faire**
        $loc_s \leftarrow loc_s \cup \{\langle s \cdot a \cdot it_{i+1} \rangle\}$
      fin
    fin
  fin
  $\mathcal{S}_{\mathcal{FG}} \leftarrow \mathcal{S}_{\mathcal{FG}} \cup \{loc_s\}$
fin

Algorithme 2: FuzzyGapGen

4.1 Base de données considérée

Les données considérées sont les expressions de 17816 gènes pour des puces associées aux prélèvements réalisés sur 286 personnes.

La base de données étudiée a été constituée afin d'étudier deux types de cancer : agressif et non agressif. Les biologistes s'intéressent surtout aux gènes qui ont une différence significative entre ces deux classes. Pour obtenir ces gènes, nous utilisons une méthode statistique appelée SAM (Significant Analysis MicroArray) qui est souvent employée dans ce contexte. Avec cette méthode, nous avons obtenu 555 gènes. A partir de ces données, nous considérons les ensembles flous trapezoïdaux proposés par les biologistes (cf. Figure 1).

4.2 Quelques motifs séquentiels extraits

Nous présentons quelques motifs séquentiels extraits en utilisant notre approche et qui sont considérés comme intéressants par les experts.

Notons qu'il est vraiment difficile pour les biologistes de définir des partitions strictes. Les biologistes ont donc facilement défini une partition floue et nous avons pu affiner la valeur du support en fonction du degré de vérité de l'intervalle flou "peu".

Finalement, l'un des résultats les plus notables vient du fait que les experts cherchent à mettre en évidence des motifs pour discriminer les cas bénin et malin. Pour cette raison, nous comparons les motifs trouvés à partir de deux sous-ensembles. Dans nos expérimentations, plusieurs motifs (e.g. $\langle(5)(41)(51)\rangle$) étaient retrouvés pour ces deux conditions expérimentales et étaient donc considérés comme non-discriminants. En utilisant notre approche, nous avons trouvé que ces motifs apparaissaient en fait avec des écarts flous différents. En effet, pour des tumeurs bénignes, le motif $\langle(5)_{moyennement}(41)_{peu}(51)\rangle$ apparaît alors que le motifs $\langle(5)_{moyennement}(41)_{beaucoup}(51)\rangle$ ap-

paraît pour des tumeurs malignes. Notre approche peut donc aider à mieux discriminer les deux formes de tumeurs.

5 Conclusion

Dans cet article, nous avons montré l'intérêt de l'utilisation de la logique floue pour rechercher de motifs pouvant servir à caractériser finement les écarts entre les éléments d'une base de données. Nous avons appliqué cette approche dans le cas de l'étude du cancer du sein. Le principal avantage de notre approche est d'offrir facilement aux utilisateurs différentes lectures des motifs séquentiels sans avoir à refaire la fouille de données pour chaque partition souhaitée. Les résultats expérimentaux décrits soulignent les résultats prometteurs. Nous envisageons maintenant d'inclure l'étude des propriétés des contraintes floues introduites par une partition floue dans le but d'améliorer les performances de notre algorithme en terme de mémoire et de temps de calculs. De plus, nous aimerions étudier l'influence du calcul du support sur les séquences obtenues par notre approche (sigma-comptage seuillé, comptage seuillé). Finalement, nous étudions la possibilité d'appliquer notre approche sur d'autres domaines d'application comme par exemple, le domaine de la domotique. Dans ce contexte, notre approche permettrait d'apprécier les écarts de temps entre des événements, par exemple, "réveil sonne" peu de temps après "allumer cafetière" moyennement après "éteindre les lumières".

Références

[FLT07] C. Fiot, A. Laurent, and M. Teisseire. Extended time constraints for sequence mining. In *14th IEEE International Symposium on Temporal Representation and Reasoning (TIME'07)*, 2007.

[JKZ73] A. Jones, A. Kaufmann, and HJ Zimmermann. *Fuzzy Sets : Theory and Applications*. Masson, 1973.

[SA96] R. Srikant and R. Agrawal. Mining sequential patterns : Generalizations and performance improvements. In *Proc. of the Fifth Int. Conference on Extending Database Technology (EDBT)*, pages 3–17, 1996.

[SBT09] P. Salle, S. Bringay, and M. Teisseire. Mining discriminant sequential patterns for aging brain. In *AIME '09 : Proceedings of the 12th conference on Artificial Intelligence in Medicine*, 2009.

[WKZ+05] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365 :671–679, 2005.