Université Montpellier II
— Sciences et Techniques du Languedoc —

T H È S E

pour obtenir le grade de
Docteur de l'Université Montpellier II

Discipline : Informatique
Spécialité Doctorale : *Informatique*
Ecole Doctorale : *Information, Structure, Systèmes*

présentée et soutenue publiquement par

Julien Rabatel

le 23 septembre 2011

Extraction de motifs contextuels : Enjeux et applications dans les données séquentielles

Jury

Rapporteurs :	Bruno Crémilleux	Professeur, Université de Caen
	Osmar R. Zaïane	Professeur, University of Alberta, Canada
Examinateurs :	François Jacquenet	Professeur, Université de Saint-Étienne
	Rosa Meo	Professeur, Università di Torino, Italie
	Maguelonne Teisseire	Directrice de Recherche, Cemagref, Montpellier
Directeurs :	Sandra Bringay	Maître de Conférence, Université Montpellier III
	Pascal Poncelet	Professeur, Université Montpellier II

Résumé : Dans de nombreux domaines d'application tels que le diagnostic médical ou la surveillance d'équipement industriel, les données peuvent être associées à des informations contextuelles décrivant les circonstances dans lesquelles les données ont été collectées. Tirer parti de ces informations peut être d'une grande utilité pour les preneurs de décision. Par exemple, l'âge ou le sexe d'un patient peut avoir une influence sur le diagnostic d'un médecin. Cette thèse aborde l'extraction de motifs dans les données séquentielles dans le but de : (1) fournir des motifs intéressants en considérant les informations contextuelles, et (2) exploiter de tels motifs pour d'autres tâches telles que la classification, la prédiction ou encore la détection d'anomalies. La première partie de cette thèse vise à considérer les informations contextuelles associées aux données pendant le processus de fouille afin de fournir aux experts des motifs représentatifs d'un contexte. Les travaux existants ne peuvent en effet pas montrer que certains motifs dépendent fortement du contexte. Nous définissons par conséquent la notion de motif fréquent contextuel. De plus, nous généralisons la notion de motif contextuel à diverses mesures d'intérêt (autres que la fréquence) : le gain d'information, le taux d'émergence, etc. Dans les deux cas, nous dévoilons et exploitons des propriétés théoriques essentielles et définissons des algorithmes efficaces. La seconde partie de ce travail concerne l'utilisation des motifs contextuels pour aborder diverses tâches de fouille de données. Nous nous intéressons particulièrement ici aux données séquentielles et abordons les problèmes de la classification, de la prédiction et de la détection d'anomalies basées sur les motifs extraits. Dans ce cadre, tenir compte des informations contextuelles a un intérêt certain. Par exemple, les motifs contextuels peuvent mettre en lumière le fait qu'un comportement considéré comme anormal en été peut être considéré comme normal en hiver. Toutes les approches proposées dans cette thèse ont été expérimentées et montrées efficaces sur des jeux de données réelles provenant de domaines variés : données de capteurs, données textuelles ou encore données médicales.

Mots-clés : Fouille de données, motifs fréquents, motifs contextuels, données séquentielles, classification, prédiction, détection d'anomalies

SUMMARY : In many application domains, such as medical diagnosis or equipment monitoring, data can be associated with contextual information describing the circumstances over which data have been collected. Taking into account this information can be of great interest for decision makers. For instance, the age or the gender of a patient can impact the diagnosis made by a medical expert. This thesis investigates the mining of sequential data in order to *(1)* provide interesting patterns by considering contextual information, and *(2)* exploit such patterns for other tasks such as classification, prediction or anomaly detection. The first part of this thesis aims at considering contextual information associated with data during the frequent pattern mining process, in order to provide the expert with patterns that are representative of a context. Existing work prior to this thesis could not reveal that some patterns strongly depend on context. We thus provide the notion of contextual frequent pattern, where a pattern is associated with a context. In addition, we generalize the notion of contextual pattern to various interestingness measures (other than frequency) : information gain, growth rate, etc. In both cases, we unveil and exploit some essential theoretical properties of contextual patterns and provide efficient algorithms. The second part of this work concerns the use of contextual patterns to address various data mining tasks. We mainly focus here on sequential data in order to perform pattern-based classification, prediction and anomaly detection. Being able to consider contextual information is here of great interest. For instance, contextual patterns can highlight the fact that a behavior that is considered as anomalous in *summer* can be considered as normal in *winter*. All our approaches have been experimented on real-world datasets coming from various application domains such as sensor numerical data, textual data or health data and have been showed to be efficient in practical applications.

KEYWORDS : Data mining, frequent patterns, contextual patterns, sequential data, classification, prediction, anomaly detection

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Ces 20 dernières années ont vu se développer une incroyable prolifération des données provoquée par les progrès des technologies de l'information. Les domaines touchés sont nombreux. Citons par exemple la grande distribution où l'entreprise *WalMart* collecte approximativement 1 million de transactions chaque heure pour un total estimé de 2,5 petabytes de données. De même, la recherche scientifique bénéficie de ces avancées. Le télescope *SDSS* (*Sloan Digital Sky Survey*) installé en 2000 a amassé plus de données dans ses premières semaines de fonctionnement que dans l'histoire de l'astronomie, puis a recueilli environ 140 terabytes de données les 10 années suivantes. Un des ses successeurs, dont l'installation est prévue pour 2016, prévoit de collecter cette même quantité de données tous les 5 jours [1]. Plus proches de la vie quotidienne, les réseaux sociaux ne sont pas en reste. On estime ainsi que *Facebook* atteindra dans l'été 2011 le nombre de 100 milliards de photos hébergées, avec 6 milliards de nouvelles photos chaque mois [2].

Paradoxalement, cette escalade vertigineuse de chiffres à laquelle nous assistons se traduit par une difficulté grandissante d'analyse et d'accès à la connaissance. Ainsi, l'opportunité marquée par l'accroissement des capacités de génération de données et de stockage soulève un défi majeur : « *Comment donner du sens aux masses de données amassées ?* »

Les communautés scientifique et industrielle ont redoublé d'attention pour offrir une réponse adaptée à cette interrogation. Au carrefour des divers domaines mis à contribution tels que les statistiques, l'apprentissage automatique, les bases de données ou encore l'interaction homme-machine, les efforts fournis ont rapidement donné jour à une nouvelle thématique de recherche : l'*Extraction de Connaissances dans les Données* (ou ECD). Mais alors, qu'est-ce que l'ECD ? Bien que fournir une réponse claire et définitive à cette question soit difficile en raison des différentes formes prises par les connaissances recherchées, la communauté concernée s'est progressivement accordée autour de la définition suivante :

> « *L'extraction de connaissances dans les données est un processus non-trivial visant à identifier dans les données des schémas valides, nouveaux, potentiellement utiles et finalement compréhensibles dans les données.* »[FPSSU96]

Avec cette définition, l'extraction de connaissances est présentée comme un processus composé d'étapes successives, visant à offrir à l'utilisateur des schémas porteurs de connaissance. Ce processus peut être décomposé en cinq étapes, illustrées dans la figure 1.1. Les trois premières, à savoir la sélection *(1)*, le pré-traitement *(2)* et la transformation des données *(3)* forment la

1. Sources : *The Ecocomist*, A special report on managing information, 25 février 2010.
2. Source : `www.pixable.com`, 14 février 2011.

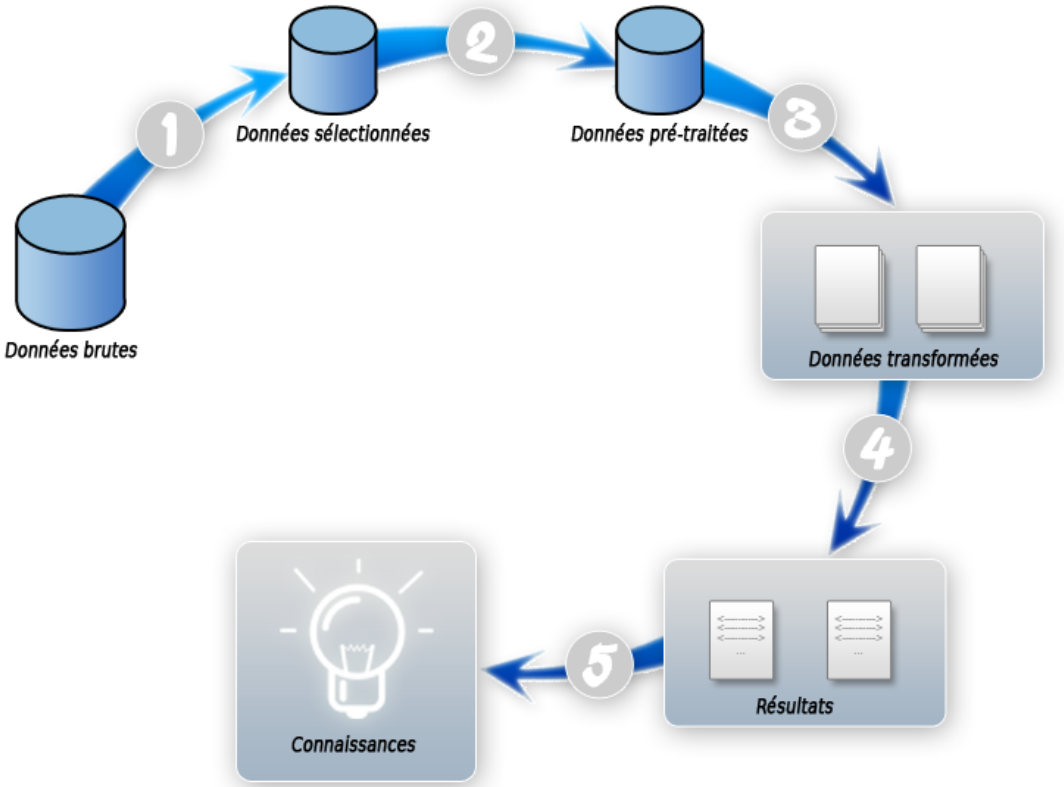


FIGURE 1.1 – Le processus d'extraction de connaissances dans les données (ECD).

préparation des données. Puis, une fois mises en forme, l'étape principale du processus concerne la fouille de données *(4)*, i.e., l'application d'un algorithme d'analyse ou de recherche pour découvrir les schémas recherchés. Ce processus n'est complet qu'au bout de l'étape de validation des résultats *(5)* où les schémas extraits sont interprétés ou évalués de manière à dégager les « *pépites de connaissances*[3] » souhaitées.

Le processus itératif d'extraction de connaissances dans les données décrit ici est un processus général qui peut varier dans sa forme en fonction de multiples critères. Tout d'abord, les données en entrée de ce processus sont diverses par nature : numériques, symboliques, booléennes, multi-dimensionnelles, multi-sources, etc. Elles peuvent également se distinguer par leur structure : données ensemblistes, arborescentes, sous la forme de graphes, ou encore, comme dans les objectifs couverts par ce manuscrit, séquentielles. Mais la nature des données n'est pas la seule source de variété. En effet, les objectifs du processus peuvent également prendre des formes diverses. L'utilisateur souhaitera, par exemple, faire face à un problème de classification, en associant, dans un cadre médical, une pathologie à un patient en fonction de ses symptômes. Dans le do-

3. De l'anglais « *knowledge nuggets* ».

maine de la grande distribution, il pourra également chercher à prédire le nombre de ventes d'un produit donné pour la semaine suivante afin de prendre des mesures adaptées. Dans un autre domaine, il pourra souhaiter détecter des anomalies dans le fonctionnement d'un équipement industriel afin de déceler au plus vite les signes annonciateurs d'une panne.

Les objectifs de l'extraction de connaissances peuvent également être d'ordre descriptif. Il pourra alors s'agir de mettre en lumière des corrélations cachées dans les données ou des tendances générales. Une thématique particulièrement importante dans ce cadre concerne la découverte de motifs. Ces motifs doivent donc vérifier certains critères d'intérêt : ils peuvent par exemple apparaître fréquemment, être surprenants ou rares, ou encore être discriminants d'un sous-ensemble de données. C'est dans ce cadre que s'inscrit ce mémoire.

1.1 Motivations

Ce mémoire, d'un point de vue général, s'intéresse à la découverte de motifs dans les données séquentielles. Il s'inscrit dans le cadre d'une thèse CIFRE [4] réalisée en collaboration avec le centre de recherche *Tecnalia* [5], afin d'intégrer les techniques de découverte de motifs dans le cadre d'applications industrielles (maintenance d'équipements, consommation énergétique, etc.). De manière à mieux comprendre nos motivations, nous présentons un cas d'étude qui sera également utilisé tout au long du manuscrit.

Considérons la base de données présentée dans le tableau 1.1 et qui décrit différentes activités (symbolisées par des lettres) effectuées au cours du temps par des habitants entre 7h et 11h dans un immeuble. Les activités peuvent correspondre dans la vie réelle à : *prendre un petit-déjeuner*, *allumer la radio*, *prendre une douche*, etc. De telles données sont qualifiées de séquentielles car elles présentent des événements (les activités) disposés suivant un ordre naturel (le temps). Par exemple, nous constatons que l'habitant H_1 a dans un premier temps réalisé les activités a et d entre 8h et 9h, puis l'activité b entre 10h et 11h.

En examinant le tableau 1.2 nous constatons également que le motif « *a suivi plus tard par* b » est vérifié par plus de 50% des habitants (i.e., 8 sur 14). En supposant que le décideur précise qu'il est intéressé par des comportements qui apparaissent dans au moins 50% des cas de la base alors le motif, que nous noterons par la suite $\langle(a)(b)\rangle$, est appelé un motif fréquent.

L'exemple précédent considère la base comme un ensemble indivisible pour rechercher les motifs. Pourtant, les circonstances liées aux données implique l'existence de sous-ensembles de données rassemblant des propriétés similaires. Pour notre cas d'étude, par exemple, nous pouvons supposer que différentes informations supplémentaires peuvent être obtenues : le tableau 1.3 associe à chaque habitant son âge (*jeune* ou *âgé*) et la saison durant laquelle les données ont été collectées (*été* ou *hiver*). Le constat que nous faisons tout au long de ce manuscrit est le suivant : de telles informations contextuelles peuvent avoir une influence non négligeable sur ce qui se produit dans les données et l'extraction de motifs devrait rendre cette influence perceptible pour l'utilisateur afin de lui offrir une vue contextualisée des données.

4. Conventions Industrielles de Formation par la REcherche.
5. `http://www.tecnalia.info`

Habitants	7h-8h	8h-9h	9h-10h	10h-11h
H_1		a,d		b
H_2	a,b		b	
H_3	a	a		b
H_4	c	a		b,c
H_5	d	a,b	b,c,d	
H_6		b		a
H_7		a	b	a
H_8	d	a		b,c
H_9		a,b	a	b,d
H_{10}			b,c,d	
H_{11}			b,d	a
H_{12}	e	b,c,d		a
H_{13}		b,d,e		
H_{14}	b		a	e

TABLE 1.1 – Exemple de base de données d'activités au cours du temps.

Habitant	7h-8h	8h-9h	9h-10h	10h-11h
H_1		**a**,d		**b**
H_2	**a**,b		**b**	
H_3	**a**	a		**b**
H_4	c	**a**		**b**,c
H_5	d	**a**,b	**b**,c,d	
H_6		b		a
H_7		**a**	**b**	a
H_8	d	**a**		**b**,c
H_9		**a**,b	a	**b**,d
H_{10}			b,c,d	
H_{11}			b,d	a
H_{12}	e	b,c,d		a
H_{13}		b,d,e		
H_{14}	b		a	e

TABLE 1.2 – Mise en valeur d'un motif.

Habitant	Age	Saison	7h-8h	8h-9h	9h-10h	10h-11h
H_1	jeune	été		**a**,d		**b**
H_2	jeune	été	**a**,b		**b**	
H_3	jeune	été	**a**	a		**b**
H_4	jeune	été	c	**a**		**b**,c
H_5	jeune	été	d	**a**,b	**b**,c,d	
H_6	jeune	hiver		b		a
H_7	jeune	hiver		**a**	**b**	a
H_8	jeune	hiver	d	**a**		**b**,c
H_9	âgé	été		**a**,b	a	**b**,d
H_{10}	âgé	été			b,c,d	
H_{11}	âgé	été			b,d	a
H_{12}	âgé	hiver	e	b,c,d		a
H_{13}	âgé	hiver		b,d,e		
H_{14}	âgé	hiver	b		a	e

TABLE 1.3 – Mise en valeur d'un motif avec les informations contextuelles.

Nombreuses sont les observations qui démontrent que l'influence du contexte se révèle parfois surprenante ou inattendue et que la connaître peut se révéler bénéfique pour le décideur. Par exemple, l'économiste Steven D. Levitt et le journaliste Stephen J. Dubner observent dans [LD05] que dans le contexte des annonces immobilières, des termes tels que *fantastique*, *charmant* ou *spacieux* doivent en réalité être interprétés négativement, puisqu'ils sont corrélés avec un prix de vente bas. Dans un tout autre domaine, les compagnies d'assurances ont remarqué que les réclamations frauduleuses sont plus susceptibles d'être déposées un lundi qu'un mardi [6]. Ces deux exemples ont un point commun : ils montrent une corrélation inattendue et difficile à mettre à jour entre le contenu des données et le contexte correspondant (celui des annonces immobilières dans le premier cas et celui du jour de la semaine dans le second).

De plus, les informations contextuelles prolifèrent avec l'évolution de l'informatique dans la vie quotidienne. En effet, l'ubiquité numérique offre de nombreux moyens d'obtenir de telles informations. Les téléphones mobiles de nouvelle génération, par exemple, fournissent des renseignements toujours plus précis sur l'environnement de leurs utilisateurs (géolocalisation, capteurs de mouvement, capteurs de lumière, etc.). De plus, dans un domaine plus industriel, le faible coût des capteurs et des moyens de stockage encourage les entreprises à collecter toute information susceptible de se révéler nécessaire. Un autre argument notable provient de l'initiative du web des données, ou *Linked Data* [BL06], visant à favoriser la publication de données dans un réseau global. On peut en effet imaginer que si cet élan se développe, il sera de plus en plus facile de relier des informations contextuelles à des jeux de données existants.

Enfin, de nombreux autres domaines de l'informatique ont récemment pris acte de l'importance du contexte. C'est par exemple le cas des systèmes de recommandation personnalisés

6. Selon l'interprétation avancée, les assurés élaborent ce type de fraude pendant le week-end, afin de préparer entre amis d'éventuels faux-témoignages. Source : The Economist, *The Data Deluge*, 25 février 2010.

[AT11], ou de suggestion de requêtes dans les moteurs de recherche [CJP+08], ou encore du domaine de l'informatique mobile où la prise en compte du contexte dans les applications se révèle de plus en plus nécessaire [Kaa03].

Pour ces différentes raisons, nous proposons au fil de ce manuscrit d'extraire des motifs en tenant pleinement compte des informations contextuelles associées aux données. En effet, nous affirmons que l'extraction de motifs, par sa capacité à permettre aux données de s'exprimer sans poser d'hypothèses *a priori*, peut naturellement faire émerger l'impact du contexte dans les données et ainsi offrir une connaissance plus fine et collant mieux à la réalité. Nous proposons de plus d'étudier l'utilité des motifs extraits dans d'autres tâches de fouille, telles que la classification, ou encore la prédiction et la détection d'anomalies dans les données séquentielles.

1.2 Contributions

Cette section décrit brièvement chacune des contributions présentées dans ce mémoire.

1.2.1 Vers de nouveaux types de motifs : les motifs fréquents contextuels

Dans l'exemple précédent, nous avons vu qu'il était possible d'extraire des comportements qui apparaissent fréquemment, i.e., en fonction d'un certain nombre d'occurrences dans la base de données. Supposons désormais que de nouvelles informations, concernant l'âge de l'habitant et la saison à laquelle les relevés ont été effectués, soient ajoutées à notre base de données. Traditionnellement, les approches d'extraction de motifs fréquents ne permettent pas, comme nous le verrons dans le chapitre 3, de prendre en compte un tel contexte. Cependant, pour le décideur, la prise en compte d'une telle information s'avère très utile. En effet, considérons à nouveau le motif $\langle(a)(b)\rangle$ de notre exemple précédent. Les approches traditionnelles vont permettre de faire apparaître que plus de 50% des personnes de la base vérifient ce comportement. En revanche, si nous regardons plus attentivement (*Cf.* tableau 1.3), nous constatons que :

- ce comportement est fréquent [7] dans la population jeune (7 jeunes sur 8), mais pas dans la population âgée (seulement 1 personne sur 6) ;
- ce comportement demeure fréquent chez les jeunes quelle que soit la saison (5 jeunes sur 5 en été, et 2 jeunes sur 3 en hiver).

Savoir que la fréquence d'un comportement est spécifique à une catégorie de personnes est alors utile pour prendre une décision la plus appropriée possible. C'est dans ce cadre que s'inscrit notre première proposition. Nous présentons une approche d'extraction de motifs fréquents contextuels, i.e., dont la propriété « *d'être fréquent* » dépend d'un contexte donné. De manière générale, l'extraction de motifs est un problème difficile qui nécessite de naviguer dans un très grand espace de recherche que la prise en compte des contextes étend encore. Via des propriétés théoriques intéressantes basées sur la fréquence, nous montrons que l'extraction des motifs peut, cependant, se faire de manière très efficace.

7. Ici, comme dans l'exemple précédent, nous considérons qu'un comportement est fréquent s'il est suivi par au moins 50% des personnes concernées.

1.2.2 Généralisation des motifs contextuels

La contribution précédente utilise des propriétés basées sur la fréquence d'apparition d'un motif dans la base pour extraire l'ensemble des motifs fréquents contextuels. Déterminer si un motif est intéressant, pour un décideur, uniquement par rapport à son nombre d'apparitions dans la base n'est cependant pas toujours suffisant. Il existe pour cela d'autres mesures d'intérêt tels que le gain d'information, le taux d'émergence, la confiance, etc. Leur objectif est de s'appuyer sur les caractéristiques statistiques des motifs pour mieux isoler les plus intéressants au sens de critères précis. En particulier s'appuyer sur ces mesures peut permettre de trouver les motifs qui différencient un sous-ensemble de données d'un autre (par exemple, le sous-ensemble de données correspondant aux jeunes à celui correspondant aux habitants âgés). Ces mesures offrent un regard nouveau sur les motifs extraits en fournissant des informations nouvelles au décideur. Alors que nous sommes à présent capables d'extraire des motifs contextuels en nous appuyant sur la seule notion de fréquence, différentes questions se posent : « *Est-il possible de généraliser les concepts et l'approche liés aux motifs contextuels pour prendre en compte différentes mesures d'intérêt ?* » ou encore « *Existe-t-il des propriétés qui permettent de rechercher efficacement ces motifs ?* ». Au cours de ce mémoire, nous répondons positivement à ces deux questions en proposant d'une part une nouvelle définition des motifs contextuels généralisée aux différentes mesures d'intérêt et en montrant, d'autre part, que pour certaines mesures il est possible d'extraire efficacement de tels motifs contextuels.

1.2.3 Motifs contextuels et classification

Le problème de la découverte de motifs dans les données a d'abord été introduit dans le but de fournir des connaissances compréhensibles et interprétables pour assister la prise de décision. Il s'agit ainsi d'une tâche descriptive de fouille de données. Plus récemment, leur champ d'application s'est élargi à d'autres tâches de fouille de données. Plus particulièrement, de nombreux travaux ont cherché à exploiter les motifs extraits pour la classification. Ces approches ont en effet l'avantage d'être interprétables et compréhensibles par les décideurs, réduisant ainsi l'effet de « *boîte noire* » fréquemment rencontré dans les approches de classification. Néanmoins, les travaux visant à classer des données en s'appuyant sur les motifs extraits ne peuvent pas, là encore, tirer parti d'informations contextuelles lorsque celles-ci sont disponibles. Intégrer les informations contextuelles à un processus de classification peut pourtant considérablement améliorer ses résultats. Par exemple, supposons que nous cherchons à classer un habitant dans une catégorie d'âge, i.e., à lui associer le label *jeune* ou *âgé* en fonction des activités qu'il suit. Alors, la classification classique basée sur les motifs s'appuiera uniquement sur les motifs extraits pour différencier les habitants *jeunes* et *âgés*. Admettons désormais que cette séquence d'activités ait été enregistrée en *été*. Les motifs contextuels nous permettent alors de tirer parti de cette information afin de considérer les différences spécifiques de l'été qui existent entre les habitants *jeunes* et *âgés*. Dans ce mémoire, nous étudions la question suivante : « *En quoi l'exploitation des motifs contextuels dans une tâche de classification peut-elle enrichir la qualité de classifieurs basés sur les motifs ?* ». En adaptant le problème de l'extraction de motifs contextuels au cadre de la classification, nous montrons que cela permet d'obtenir de meilleurs résultats. De plus, dans le cadre des données séquentielles, nous observons que le problème de classification peut

facilement s'appliquer à un problème de prédiction. Par exemple, la question « *La consommation d'électricité sera-t-elle élevée ou basse dans une heure ?* » peut se traduire par « *Quel label de classe associer à la consommation d'électricité dans une heure : basse ou élevée ?* ». En suivant ce constat, nous montrons que les motifs contextuels dans les données séquentielles se révèlent alors très utiles.

1.2.4 Motifs contextuels et détection d'anomalies

La détection d'anomalies dans les données séquentielles, se rapportant à la découverte de fragments de séquences qui ne correspondent pas au comportement attendu, constitue un défi porteur d'enjeux considérables dans de multiples domaines. Par exemple, pour la surveillance d'équipements industriels une telle anomalie peut annoncer un dysfonctionnement grave. Dans le domaine médical, une anomalie relevée dans le battement cardiaque d'un patient peut mettre en évidence une maladie. D'autre part, une séquence anormale d'actions enregistrée sur un système informatique peut traduire une tentative d'attaque. De même, dans notre cas d'étude, les décideurs seront par exemple intéressés par une sur-consommation d'électricité. Une telle anomalie pourra être annonciatrice du dysfonctionnement d'un matériel électrique et la détecter pourra prévenir des dégâts plus sérieux.

La détection d'anomalies dans les données séquentielles est une tâche difficile qui doit bien souvent faire face au manque de connaissances et de données utiles pour caractériser les différentes anomalies possibles. De plus, là encore, les informations contextuelles disponibles sont extrêmement importantes puisqu'un même comportement peut être considéré comme normal ou anormal en fonction des circonstances dans lesquelles il se manifeste. Par exemple, une approche pertinente de détection d'anomalies devra prendre en compte le fait que la consommation électrique attendue ne sera pas identique en été ou en hiver.

En partant de l'idée générale qu'un motif fréquent représente un comportement attendu, nous nous posons les questions suivantes : « *Comment utiliser les motifs fréquents pour répondre au problème de la détection d'anomalies ?* » et « *Comment les motifs fréquents contextuels peuvent être employés pour tenir compte des circonstances dans lesquelles une anomalie est rencontrée ?* ». Nous proposons donc une approche de détection d'anomalies dans les séquences basée sur les motifs contextuels et montrons que ceux-ci peuvent répondre aux problèmes posés.

1.3 Organisation du mémoire

La suite de ce mémoire s'organise comme suit.

Tout d'abord, nous revenons dans le chapitre 2 sur la découverte de motifs fréquents dans les données, puis nous nous focalisons plus particulièrement sur les données séquentielles en dressant un panorama des travaux existants.

Le chapitre 3 présente les lacunes des motifs fréquents « *traditionnels* » pour intégrer les informations contextuelles disponibles avec les données à analyser, puis s'intéresse à la définition d'un nouveau type de motifs fréquents tirant parti de ces informations : les motifs fréquents

contextuels. Ce chapitre introduit également une approche d'extraction efficace de tels motifs dans les données.

Le chapitre 4 étend le concept de motif contextuel fréquent en s'intéressant à sa généralisation pour des mesures d'intérêt diverses. En particulier, les propriétés théoriques de certaines mesures d'intérêt sont exploitées pour extraire efficacement de tels motifs contextuels.

Le chapitre 5 étudie l'apport des motifs contextuels définis et extraits dans le chapitre précédent dans une tâche de classification. Plus précisément, l'exploitation des motifs contextuels dans ce cadre offre la possibilité d'intégrer les informations contextuelles et ainsi d'améliorer les performances des classifieurs classiques. De plus, nous montrons que le concept de motif contextuel peut être étendu pour résoudre un problème de prédiction dans les données séquentielles.

Puis, le chapitre 6 s'intéresse à l'apport des motifs contextuels pour détecter des anomalies dans les données séquentielles. Là encore, nous montrons que la prise en compte du contexte offre une qualité de détection améliorée.

Enfin, le chapitre 7 dresse un bilan général des contributions développées tout au long de ce manuscrit et propose les perspectives de recherche associées.

Chapitre 2

Motifs fréquents et données séquentielles

Sommaire

Introduction

Comme nous l'avons vu en introduction, l'extraction de motifs fréquents, suite à son introduction dans [AIS93], demeure une des tâches clés de la fouille de données. Elle permet de dégager des motifs intéressants parmi une très grande quantité de données disponibles. L'évolution des méthodes basées sur les motifs fréquents a suivi ces dernières années différents développements liés aux différents types de données rencontrés (ensembles d'items [AIS93, AS94], séquences [AS95, MTIV97], arbres [MSU+01] ou graphes [KK01, YH02]) et à la variété des besoins dans les applications réelles.

Dans ce chapitre, nous présentons les principales approches d'extraction de motifs fréquents. Dans la section 2.1, nous définissons formellement la notion de motif fréquent en nous inspirant de manière originale de l'analyse de concepts formels. Nous utilisons dans la section 2.2 cette formalisation dans le cas des données séquentielles visées dans ce mémoire en l'illustrant sur trois types de motifs : les itemsets fréquents, les motifs séquentiels et les itemsets inter-transactionnels. Enfin, nous discutons l'apport de ces différents motifs dans la section 2.3.

2.1 Motifs fréquents

Depuis l'apparition de la problématique de l'extraction des motifs fréquents dans les données, de nombreuses définitions de motifs ont été proposées dans la littérature. Ces motifs diffèrent selon : *(1)* le type de données dans lesquelles les motifs sont extraits (séquences, ensembles d'items, graphes, structures arborescentes, etc.) ou encore *(2)* la définition de fréquence qui leur est associée.

Dans ce manuscrit, nous nous intéresserons, par la suite, aux motifs extraits dans les données séquentielles. Toutefois, afin de généraliser notre problématique, nous proposons, dans un premier temps, une formalisation générale pour manipuler les motifs fréquents avant de nous focaliser sur des motifs plus spécifiques.

Cette formalisation est en partie inspirée de l'analyse de concepts formels [Wil82] développée pour analyser de manière formelle la notion de concept dans le cadre de la théorie des treillis.

Intuitivement, la recherche de motifs fréquents dans un ensemble de données vise à identifier, parmi un ensemble d'*objets*, les *caractéristiques* communes à un nombre suffisant d'objets (ce nombre étant généralement fixé par l'utilisateur).

Définition 1 (Environnement d'extraction) **:** Soit le triplet $\mathcal{K} = (\mathcal{B}, \mathcal{M}, \mathcal{R})$ un **environnement d'extraction**, où $\mathcal{B} = \{o_1, o_2, ..., o_n\}$ est un ensemble d'**objets** appelé **base d'objets**, $\mathcal{M} = \{m_1, m_2, ..., m_m\}$ est un ensemble de **motifs** [1] et $\mathcal{R}$ une **relation de support** binaire entre $\mathcal{B}$ et $\mathcal{M}$ vérifiant $\mathcal{R} \subseteq \mathcal{B} \times \mathcal{M}$.

Afin d'identifier les motifs présents dans un certain nombre d'objets, nous nous intéressons

1. Le terme *motif* correspond dans d'autres domaines, en particulier dans l'analyse de concepts formels, au terme *attribut*.

à la relation qui existe entre les objets et les motifs d'un environnement d'extraction donné.

Définition 2 (Relation de support) **:** Soit $o \in \mathcal{B}$ et $m \in \mathcal{M}$. L'objet o **supporte** le motif m (et m est supporté par o) s'il existe un couple (o, m) tel que $(o, m) \subseteq \mathcal{R}$.

Exemple 1 : Soit $\mathcal{B} = \{o_1, o_2, o_3, o_4, o_5, o_6\}$ une base d'objets et $\mathcal{M} = \{m_1, m_2, m_3, m_4, m_5\}$ un ensemble de motifs. Afin de décrire la relation de support $\mathcal{R}$ entre $\mathcal{B}$ et $\mathcal{M}$, considérons le tableau 2.1 ci-dessous qui offre une représentation communément utilisée dans l'analyse de concepts formels.

	m_1	m_2	m_3	m_4	m_5
o_1	×	×			
o_2			×		
o_3	×			×	×
o_4	×	×	×		
o_5	×	×	×		
o_6	×			×	

TABLE 2.1 – Représentation d'un environnement d'extraction $(\mathcal{B}, \mathcal{M}, \mathcal{R})$.

Dans ce tableau, les objets de $\mathcal{B}$ correspondent aux lignes et les motifs de $\mathcal{M}$ correspondent aux colonnes. Une croix ($\times$) est présente dans une cellule (o, m) où $o \in \mathcal{B}$ et $m \in \mathcal{M}$ si et seulement si $(o, m) \in \mathcal{R}$. Ainsi, l'objet o_1 supporte les motifs m_1 et m_2. En revanche, o_1 ne supporte pas le motif m_3.

La relation de support nous permet désormais de définir le support absolu d'un motif dans une base d'objets.

Définition 3 (Support absolu d'un motif) **:** Le **support absolu** du motif $m \in \mathcal{M}$ dans $\mathcal{B}$, noté $Supp_{\mathcal{B}}(m)$, est défini comme le nombre d'objets dans $\mathcal{B}$ qui supportent m :

$$Supp_{\mathcal{B}}(m) = |\{o \in |(o, m) \in \mathcal{R}\}|.$$

Exemple 2 : D'après l'exemple précédent, le support absolu du motif m_1 est $Supp_{\mathcal{B}}(m_1) = |\{o_1, o_3, o_4, o_5, o_6\}| = 5$ et le support absolu du motif m_4 est $Supp_{\mathcal{B}}(m_4) = |\{o_3, o_6\}| = 2$.

Le support absolu d'un motif, appelé par la suite support, permet de manipuler le nombre d'objets le supportant. Néanmoins, lorsque l'on s'intéresse à la proportion d'objets qui supportent un motif, il est courant d'utiliser la notion de fréquence d'un motif.

Définition 4 (Fréquence d'un motif) **:** La **fréquence** du motif $m \in \mathcal{M}$ dans $\mathcal{B}$, notée $Freq_{\mathcal{B}}(m)$, est définie telle que :

$$Freq_{\mathcal{B}}(m) = \frac{Supp_{\mathcal{B}}(m)}{|\mathcal{B}|}.$$

Exemple 3 : La fréquence du motif m_1 est $Freq_{\mathcal{B}}(m_1) = \frac{Supp_{\mathcal{B}}(m_1)}{|\mathcal{B}|} = \frac{5}{6}$ et la fréquence du motif m_4 est $Freq_{\mathcal{B}}(m_4) = \frac{Supp_{\mathcal{B}}(m_4)}{|\mathcal{B}|} = \frac{2}{6}$.

La définition d'un motif fréquent découle directement des concepts préliminaires définis plus tôt.

Définition 5 (Motif fréquent) **:** Soit σ un réel tel que $0 < minFreq \leq 1$, appelé **seuil minimum de fréquence**. Un motif $m \in \mathcal{M}$ dont la fréquence dans $\mathcal{B}$ est supérieure ou égale à σ est appelé un **motif fréquent** dans $\mathcal{B}$. En d'autres termes, m est fréquent dans $\mathcal{B}$ si :

$$Freq_{\mathcal{B}}(m) \geq \sigma.$$

Au contraire, si le motif m vérifie $Freq_{\mathcal{B}}(m) < \sigma$, alors il est dit infréquent.

L'ensemble des motifs fréquents dans $\mathcal{B}$ pour un seuil minimum de fréquence σ, noté $MFreq(\mathcal{B}, \sigma)$, est défini comme :

$$MFreq(\mathcal{B}, \sigma) = \{m \in \mathcal{M} | Freq_{\mathcal{B}}(m) \geq \sigma\}.$$

Exemple 4 : Soit un seuil minimum de fréquence $\sigma = 0,5$, alors le motif m_1 est un motif fréquent dans $\mathcal{B}$. En effet, $Freq_{\mathcal{B}}(m_1) = \frac{5}{6} \geq \sigma$. En revanche, le motif m_4 n'est pas fréquent dans $\mathcal{B}$ car $Freq_{\mathcal{B}}(m_4) = \frac{2}{6} < \sigma$.

L'ensemble $MFreq(\mathcal{B}, \sigma)$ de tous les motifs fréquents dans $\mathcal{B}$ est $MFreq(\mathcal{B}, \sigma) = \{m_1, m_2, m_3\}$.

La notion de motif fréquent définie ici s'appuie sur la fréquence d'un motif. Toutefois, la littérature propose également des définitions liées au support du motif. Dans ce cas, un motif est fréquent si son support est supérieur ou égal à un seuil minimum défini comme un entier δ, tel que $0 < \delta \leq |\mathcal{B}|$.

2.1.1 Ordre partiel et propriétés des motifs fréquents

Les concepts présentés ci-dessus permettent de généraliser une grande partie des motifs fréquents présents dans la littérature, indépendamment de la structure des objets et des motifs considérés (séquences, itemsets, arbres, graphes, etc.). Cependant, la grande majorité des motifs existants (et en particulier ceux que nous utiliserons tout au long de ce manuscrit) possèdent des propriétés supplémentaires, sur lesquelles reposent les algorithmes d'extraction. Notamment, les motifs ne sont généralement pas indépendants les uns des autres mais sont liés par un ordre partiel sur l'ensemble des motifs.

Définition 6 (Ordre partiel sur $\mathcal{M}$) **:** L'ensemble des motifs $\mathcal{M}$ est associé à un ordre partiel $\prec$ tel que $\forall (m, m') \in \mathcal{M} \times \mathcal{M}$:

1. si $m \prec m'$, alors m est un **sous-motif** de m' et m' un **super-motif** de m ;

2. si $m \not\prec m'$ et $m' \not\prec m$, alors m et m' sont dits **incomparables** et notés $m \prec\succ m'$;
3. $\forall o \in \mathcal{B}$, si $m \prec m'$ et o supporte m', alors o supporte également m.

Exemple 5 : Afin d'illustrer notre propos, considérons que l'ensemble des objets $\mathcal{B}$ décrit des personnes et que l'ensemble des motifs $\mathcal{M}$ décrit des caractéristiques sur ces personnes. Le motif m_4 pourra par exemple correspondre à la caractéristique « *porter des lunettes* » et m_5 à « *porter des lunettes et un chapeau* ». Jusqu'ici, nous avons considéré les motifs comme indépendants les uns des autres. Cependant, par cet exemple, nous constatons que certains motifs sont liés. Le motif m_5 peut ainsi être considéré comme une composition du motif « *porter des lunettes* » et d'un autre motif « *porter un chapeau* ». Sous ce nouvel angle, un ordre est naturellement conçu sur ces motifs : m_5 est un super-motif de m_4. De plus, la troisième partie de la définition précédente est vérifiée par cet ordre : si une personne supporte le motif « *porter des lunettes et un chapeau* » alors elle supporte nécessairement le motif « *porter des lunettes* ».

L'ordre partiel ainsi défini implique directement une propriété centrale pour la construction d'algorithmes efficaces d'extraction de motifs fréquents.

Propriété 1 (Anti-monotonie) **:** Soit m et m' deux motifs. Si $m \prec m'$, alors $Supp_{\mathcal{B}}(m) \geq Supp_{\mathcal{B}}(m')$ et $Freq_{\mathcal{B}}(m) \geq Freq_{\mathcal{B}}(m')$.

La fréquence et le support des motifs sont anti-monotones, relativement à l'ordre partiel défini sur ceux-ci. L'application de la propriété d'anti-monotonie implique directement que tout super-motif m' d'un motif m est nécessairement infréquent si m est infréquent. En effet, d'après la propriété 1, $Freq_{\mathcal{B}}(m') \leq Freq_{\mathcal{B}}(m) < \sigma$, donc m' n'est pas fréquent.

Exemple 6 : Observons l'environnement d'extraction présenté dans le tableau 2.1 et revenons à l'ordre sur les motifs défini dans l'exemple précédent (tel que $m_4 \prec m_5$). Le motif m_4 étant infréquent (*cf.* exemple précédent), la propriété 1 nous permet de déduire que $Freq_{\mathcal{B}}(m_4) \geq Freq_{\mathcal{B}}(m_5)$ et donc que m_5 est également infréquent. En d'autres termes, s'il n'est pas fréquent de trouver une personne portant des lunettes alors il n'est également pas fréquent de trouver une personne portant à la fois des lunettes et un chapeau.

À notre connaissance, toutes les catégories de motifs fréquents dans une base d'objets peuvent être définis suivant la formalisation proposée ci-dessus. En effet, les travaux existants diffèrent généralement uniquement sur la définition des éléments présents dans l'environnement d'extraction.

2.1.2 Extraction de motifs fréquents

De manière générale, l'extraction de motifs fréquents dans une base d'objets peut être vue comme l'énumération des motifs respectant une contrainte de fréquence (ou de support). Ainsi, une méthode naïve d'extraction consiste en l'énumération de tous les motifs possibles, avant de compter la fréquence de chacun d'eux dans la base d'objets. Une telle méthode est dans la

pratique inenvisageable. En effet, le nombre de ces motifs est généralement extrêmement grand. Compter le support de chacun d'eux demeure par conséquent irréalisable en raison des coûts engendrés sur l'espace mémoire et du temps nécessaire à l'exécution des processus.

Pour cette raison, la propriété 1 d'anti-monotonie tient un rôle essentiel dans l'extraction de motifs fréquents en offrant la possibilité de réduire considérablement l'espace de recherche.

Dans la section suivante, nous étudions les motifs fréquents dans le cadre des données séquentielles. Ainsi, nous illustrons les différents concepts présentés jusqu'ici pour des environnements d'extraction spécifiques.

2.2 Motifs fréquents et données séquentielles

Dans cette section, nous dressons un panorama des différents types de motifs fréquents extraits dans les données séquentielles. Ces motifs seront définis à l'aide du formalisme général proposé dans la section 2.1. Définir un type de motif consiste à définir son environnement d'extraction, i.e., le triplet composé de la *base d'objets* $\mathcal{B}$, de l'*ensemble de motifs* $\mathcal{M}$ et de la *relation* $\mathcal{R}$ correspondants.

Nous considérons, par la suite, deux types de motifs dans les données séquentielles : les motifs séquentiels et les itemsets inter-transactionnels. Cependant, dans un premier temps, nous présentons la découverte d'itemsets fréquents qui est à l'origine de la problématique générale de l'extraction de motifs fréquents.

2.2.1 Itemsets fréquents

L'extraction de séquences fréquentes dans les données séquentielles est issue de la notion d'itemset fréquent initialement définie dans [AIS93].

Nous reprenons les notions présentées dans la section 2.1 et définissons l'environnement d'extraction associé. Pour ce faire, nous nous appuyons en premier lieu sur la notion d'itemset définie ci-dessous.

Définition 7 (Itemset) **:** Soit $\mathcal{I} = \{i_1, i_2, ..., i_m\}$ un ensemble fini d'**items**. Un **itemset** est un ensemble non vide d'items dans $\mathcal{I}$ noté $I = \{i_1 i_2 ... i_n\}$, i.e., pour $1 \leq j \leq n$, $i_j \in \mathcal{I}$.

Définition 8 (Base d'itemsets) **:** Soit $\mathcal{B} = \{I_1, I_2, ..., I_n\}$ une base d'objets, telle que $\forall i \in \{1, ..., n\}$, I_i est un itemset. $\mathcal{B}$ est appelée une **base d'itemsets**.

Exemple 7 : Considérons les données décrites dans le tableau 1.1. Nous traduisons ces données sous la forme de la base d'itemsets présentée dans le tableau 2.2, où pour chaque habitant l'itemset correspondant contient l'ensemble de ses activités. Notons que cette représentation entraîne la perte de certaines informations présentes dans la base de données initiales. En effet, il n'est, dans cette représentation, pas possible de savoir combien de fois une activité a été suivie par un même habitant (car les différentes occurrences d'une même activité apparaissent sous la forme d'un seul item dans l'itemset), ni dans quel ordre les activités ont été réalisées.

ID	Itemset
H_1	$\{a, b, d\}$
H_2	$\{a, b\}$
H_3	$\{a, b\}$
H_4	$\{a, b, c\}$
H_5	$\{a, b, c, d\}$
H_6	$\{a, b\}$
H_7	$\{a, b\}$
H_8	$\{a, b, c, d\}$
H_9	$\{a, b, d\}$
H_{10}	$\{b, c, d\}$
H_{11}	$\{a, b, d\}$
H_{12}	$\{a, b, c, d, e\}$
H_{13}	$\{b, d, e\}$
H_{14}	$\{a, b, e\}$

TABLE 2.2 – Base d'itemsets associée au tableau 1.1.

Nous définissons maintenant l'ensemble des motifs. Il devrait intuitivement être défini comme l'ensemble de tous les sous-ensembles non vides d'items dans $\mathcal{I}$. Dans ce cas, le nombre d'itemsets possibles est $2^{|\mathcal{I}|} - 1$. Comme nous l'annoncions dans la section précédente, ce nombre de possibilités rend impossible une énumération complète des motifs, y compris pour des tailles de $\mathcal{I}$ modérées. En effet, comme souligné dans [Raï08], pour une taille de $\mathcal{I}$ égale à 263, ce nombre serait supérieur au nombre estimé d'atomes dans l'univers, i.e., 10^{79}.

Définition 9 (Ensemble de motifs) **:** L'ensemble de motifs est l'ensemble de tous les itemsets construits à partir des items de $\mathcal{I}$. En d'autres termes, il s'agit de l'ensemble des parties de $\mathcal{I}$ privé de l'ensemble vide $\emptyset$.

Exemple 8 : Étant donné $\mathcal{B}$ issu du tableau 2.2, l'ensemble des motifs $\mathcal{M}$ contient $2^5 - 1 = 31$ itemsets, i.e., l'ensemble des parties de $\mathcal{I}$ privé de $\emptyset$. Une partie de $\mathcal{M}$ correspondant à notre exemple est décrite ci-dessous.

$$\mathcal{M} = \{\{a\}, \{a, b\}, \{a, b, c\}, \{a, b, c, d\}, \{a, b, c, d, e\}, \{b\}, \{b, c\}, ..., \{e\}\}.$$

Remarquons que les éléments de la base d'objets et de l'ensemble des motifs ont la même structure. Ce sont des itemsets. En outre, la relation de support entre les objets de $\mathcal{B}$ et les motifs de $\mathcal{M}$ est naturellement définie par le biais de l'inclusion ensembliste.

Définition 10 (Relation de support) **:** La **relation de support** $\mathcal{R}$ est une relation binaire entre $\mathcal{B}$ et $\mathcal{M}$ définie par $\mathcal{R} = \{(I, m) \in \mathcal{B} \times \mathcal{M} | m \subseteq I\}$.

Exemple 9 : D'après la relation de support ainsi définie, l'itemset $H_2 = \{a, b\}$ supporte les motifs $\{a\}$, $\{b\}$ et $\{a, b\}$.

Les éléments $\mathcal{B}$, $\mathcal{M}$ et $\mathcal{R}$ étant définis dans le cadre des itemsets, nous pouvons désormais formellement définir l'environnement d'extraction d'itemsets fréquents comme suit.

Définition 11 (Environnement d'extraction d'itemsets fréquents) **:** L'**environnement d'extraction d'itemsets fréquents** dans une base d'itemsets $\mathcal{B}$ est le triplet $(\mathcal{B}, \mathcal{M}, \mathcal{R})$.

	$\{a\}$	$\{a,b\}$	$\{a,b,c\}$	$\{a,b,c,d\}$	$\{a,b,c,d,e\}$	$\{b\}$	$\{b,c\}$	...	$\{e\}$
H_1	×	×				×			
H_2	×	×				×			
H_3	×	×				×			
H_4	×	×	×			×	×		
H_5	×	×	×	×		×	×		
H_6	×	×				×			
H_7	×	×				×			
H_8	×								
H_9	×	×	×	×		×	×		
H_{10}						×	×		
H_{11}	×	×				×			
H_{12}	×	×	×	×	×	×	×		×
H_{13}						×			×
H_{14}	×	×							×

TABLE 2.3 – Représentation de l'environnement d'extraction d'itemsets fréquents pour le tableau 1.1.

Exemple 10 : Comme nous l'avons vu dans la section 2.1, un environnement d'extraction peut être représenté sous la forme d'un tableau. Le tableau 2.3 montre l'environnement d'extraction d'itemsets fréquents pour la base d'itemsets $\mathcal{B}$.

Choisissons un seuil minimum de fréquence σ fixé à $0,5$. Si l'on se réfère au tableau 2.3 présentant notre environnement d'extraction, les itemsets fréquents sont les motifs supportés par au moins 7 itemsets de $\mathcal{B}$ (i.e., de fréquence supérieure ou égale à 7/14).

Considérons par exemple l'itemset $\{a,b\}$. Il est supporté par onze éléments de la base d'itemsets. Son support est donc $Supp_{\mathcal{B}}(\{a,b\}) = 11$. Sa fréquence est $Freq_{\mathcal{B}}(\{a,b\}) = \frac{Supp_{\mathcal{B}}(\{a,b\})}{|\mathcal{B}|} = \frac{11}{14} \geq 0,5$. Par conséquent, $\{a,b\}$ est un itemset fréquent dans $\mathcal{B}$.

En revanche, l'itemset $\{a,b,c\}$ n'est pas fréquent. En effet, $Freq_{\mathcal{B}}(\{a,b,c\}) = \frac{Supp_{\mathcal{B}}(\{a,b,c\})}{|\mathcal{B}|} = \frac{4}{14} < \sigma$.

2.2.1.1 Un détour par les règles d'association

De plus, [AIS93] étend les itemsets fréquents en définissant également la notion de règle d'association comme suit.

Définition 12 (Règle d'association) **:** Une **règle d'association** est une implication de la forme

$X \rightarrow Y$, où X est un itemset ($X \subset \mathcal{I}$) et Y est un item ($Y \in \mathcal{I}$) tel que $Y \notin X$).
Par la suite, [AS94] étend cette définition. Ainsi, Y est un itemset tel que $Y \cap X = \emptyset$.

Une règle d'association est associée à deux mesures d'intérêt : la fréquence et la confiance, définies comme suit.

Définition 13 (Fréquence et confiance) **:** La **fréquence** de la règle d'association $X \rightarrow Y$, noté $Freq_{\mathcal{B}}(X \rightarrow Y)$, est définie comme la fréquence de l'itemset $X \cup Y$, i.e., $Freq_{\mathcal{B}}(X \rightarrow Y) = Freq_{\mathcal{B}}(X \cup Y)$. La **confiance** de $X \rightarrow Y$, notée $conf_{\mathcal{B}}(X \rightarrow Y)$, établit la probabilité qu'une transaction de $\mathcal{B}$ supportant l'itemset X supporte également l'itemset Y, i.e., $conf_{\mathcal{B}}(X \rightarrow Y) = \frac{Supp_{\mathcal{B}}(X \cup Y)}{Supp_{\mathcal{B}}(X)}$.

Exemple 11 : Dans notre exemple, considérons la règle d'association $\{a\} \rightarrow \{b\}$. La fréquence de cette règle dans $\mathcal{B}$ est la fréquence de l'itemset $\{a, b\}$, i.e., $Freq_{\mathcal{B}}(\{a, b\}) = \frac{11}{14}$. Sa confiance est $conf_{\mathcal{B}}(\{a\} \rightarrow \{b\}) = \dfrac{Supp_{\mathcal{B}}(\{a, b\})}{Supp_{\mathcal{B}}(\{a\})} = \dfrac{11}{12} \approx 0,92$.

2.2.1.2 Extraction d'itemsets fréquents

L'extraction d'itemsets fréquents est une tâche difficile. En effet, nous avons déjà pu observer que l'espace de recherche est exponentiel selon le nombre d'items composant $\mathcal{I}$. L'extraction d'itemsets fréquents ne faisant pas l'objet de ce manuscrit, nous étudions uniquement ici *Apriori*, l'algorithme pionner proposé en 1994 de manière indépendante par [AS94] et [MTV94], avant de faire l'objet d'une publication conjointe dans [AMS+96].

L'idée centrale consiste à utiliser la propriété d'anti-monotonie (*Cf.* propriété 1) dans une approche de type *générer-élaguer* décrite dans l'algorithme 1 qui se décompose en quatre étapes :

1. Extraire l'ensemble de tous les itemsets fréquents de taille 1 (F_1 dans l'algorithme 1), i.e., ne contenant qu'un seul item. D'après la propriété d'anti-monotonie, tous les autres itemsets fréquents seront nécessairement des super-motifs de ceux-ci.
2. Utiliser les itemsets fréquents de taille $k = 1$ pour générer des candidats de taille $k+1 = 2$ (opération *GenererCandidats*).
3. Calculer la fréquence dans la base de chaque candidat (opération *CompterFrequence*) pour obtenir l'ensemble des itemsets fréquents de taille $k + 1 = 2$.
4. Réitérer les deux étapes précédentes pour $k = k + 1$, jusqu'à ce qu'aucun itemset fréquent de taille k ne soit extrait.

Les idées générales introduites par l'algorithme *Apriori* ont depuis été largement étendues. Étant hors de notre problématique, nous ne les détaillons pas ici, néanmoins le lecteur intéressé peut se reporter par exemple à [Goe02] pour de plus amples informations. Nous précisons toutefois quelques améliorations proposées ces dernières années. De nouvelles stratégies de recherche, ne fonctionnant plus par niveau comme *Apriori*, ont tout d'abord été proposées (par exemple *Eclat* [OZP+97] utilise un parcours en profondeur). Le constat d'un trop grand nombre de candidats générés a donné lieu à des approches telles que *FP-Growth* [HPY00] qui propose une structuration nouvelle de la base. De nombreux travaux se sont également intéressés à la

Algorithm 1 APRIORI

Entrées: une base d'itemsets $\mathcal{B}$, un seuil minimum de fréquence σ.
Sorties: l'ensemble $MFreq(\mathcal{B}, \sigma)$ des itemsets fréquents dans $\mathcal{B}$.

$F_1 \leftarrow \{i \in \mathcal{I} | Freq_{\mathcal{B}}(i) \geq \sigma\}$
$k \leftarrow 1$
tantque $F_k \neq \emptyset$ **faire**
  $C_{k+1} \leftarrow GenererCandidats(F_k)$
  $CompterFrequence(C_{k+1})$
  $F_{k+1} \leftarrow \{I \in C_{k+1} | Freq_{\mathcal{B}}(I) \geq \sigma\}$
  $k \leftarrow k + 1$
fin tantque
retourne $F = \bigcup_{j \in \{1,...,k\}} F_k$

réduction de l'espace de recherche et notamment à l'utilisation de représentations condensées dont l'objectif est d'obtenir un ensemble restreint de motifs conservant les données nécessaires pour en extraire tous les motifs (motifs fermés [PBTL99, ZH02], motifs libres [BBR00], motifs non-dérivables [CG02], etc.).

2.2.2 Motifs séquentiels

Dans la suite de ce travail, nous nous intéressons plus particulièrement à la découverte de motifs séquentiels. Contrairement aux itemsets fréquents, les motifs séquentiels introduits dans [AS95] prennent en compte un ordre (généralement temporel) lié aux itemsets afin de découvrir des corrélations fréquentes entre les items au cours du temps. Il s'agira, par exemple, d'extraire les ensembles fréquents d'activités qui sont effectuées dans un ordre fixe par des habitants.

De la même manière que l'extraction d'itemsets fréquents vise à trouver les sous-ensembles d'items fréquents dans une base d'itemsets, l'extraction de motifs séquentiels consiste à rechercher les sous-séquences fréquentes dans une base de séquences.

Comme nous l'avons fait pour les itemsets fréquents, nous présentons ci-dessous le problème de la découverte de motifs séquentiels par la définition de son environnement d'extraction.

Définition 14 (Base de séquences) **:** Soit $\mathcal{B} = \{S_1, S_2, ..., S_n\}$ une base d'objets, telle que $\forall i \in \{1, ..., n\}$ où S_i est une séquence. $\mathcal{B}$ est appelée une **base de séquences**.

Exemple 12 : Considérons les données décrites dans le tableau 1.1. Elles peuvent naturellement être traduites dans la base de séquences présente dans le tableau 2.4. Par exemple, la séquence S_1 indique qu'un habitant a suivi les activités a et d au cours de la même heure, puis a plus tard suivi l'activité d.

De la même manière que, dans le cadre de l'extraction d'itemsets fréquents, nous avons défini l'ensemble des motifs comme l'ensemble des itemsets contenus dans au moins un objet de la base, nous considérons ici l'ensemble des séquences contenues dans au moins un élément de la base de séquences. Les itemsets étant définis comme des ensembles d'items, leur inclusion est définie naturellement par l'inclusion ensembliste. En revanche, la notion de séquence requiert

ID	Séquence
S_1	$\langle(ad)(b)\rangle$
S_2	$\langle(ab)(b)\rangle$
S_3	$\langle(a)(a)(b)\rangle$
S_4	$\langle(c)(a)(bc)\rangle$
S_5	$\langle(d)(ab)(bcd)\rangle$
S_6	$\langle(b)(a)\rangle$
S_7	$\langle(a)(b)(a)\rangle$
S_8	$\langle(d)(a)(bc)\rangle$
S_9	$\langle(ab)(a)(bd)\rangle$
S_{10}	$\langle(bcd)\rangle$
S_{11}	$\langle(bd)(a)\rangle$
S_{12}	$\langle(e)(bcd)(a)\rangle$
S_{13}	$\langle(bde)\rangle$
S_{14}	$\langle(b)(a)(e)\rangle$

TABLE 2.4 – Une base de séquences.

une définition spécifique de l'inclusion.

Définition 15 (Sous-séquence) **:** Soient $s = \langle I_1 I_2 ... I_m\rangle$ et $s' = \langle I'_1 I'_2 ... I'_n\rangle$ deux séquences. s est une **sous-séquence** de s', noté $s \sqsubseteq s'$, si $\exists i_1, i_2, ... i_m$ avec $1 \leq i_1 < i_2 < ... < i_m \leq n$ tels que $I_1 \subseteq I'_{i_1}$, $I_2 \subseteq I'_{i_2}$, ..., $I_m \subseteq I'_{i_m}$.

Exemple 13 : Soient trois séquences $s_1 = (c)(a)(bc)$, $s_2 = \langle(c)(b)\rangle$ et $s_3 = \langle(b)(c)\rangle$. s_2 est une sous-séquence de s_1 ($s_2 \sqsubseteq s_1$). En revanche, s_3 n'est pas une sous-séquence de s_1.

Définition 16 (Ensemble de motifs) **:** Soit $\mathcal{B}$ une base de séquences. L'**ensemble de motifs** $\mathcal{M}$ correspondant est un ensemble de séquences tel que $\mathcal{M} = \bigcup_{S \in \mathcal{B}} \{m \sqsubseteq I\}$.

Exemple 14 : L'ensemble des motifs est par conséquent l'ensemble de toutes les séquences incluses dans au moins un élément de $\mathcal{B}$. Pour la base de séquences $\mathcal{B}$ donnée dans le tableau 5.3.2, il y a 108 motifs dans $\mathcal{M}$. Parmi ceux-ci se trouvent les séquences $\langle(a)\rangle$, $\langle(ab)\rangle$, $\langle(ad)\rangle$, $\langle(a)(b)\rangle$, $\langle(a)(a)\rangle$, etc.

Similairement à ce que nous avons pu observer dans le cadre des itemsets fréquents, les objets et les motifs ont une structure identique : ce sont des séquences. Par conséquent, la relation de support peut être explicitée par le biais de l'inclusion de séquences déjà définie.

Définition 17 (Relation de support) **:** La **relation de support** $\mathcal{R}$ est une relation binaire entre $\mathcal{B}$ et $\mathcal{M}$ définie par $\mathcal{R} = \{(S, m) \in \mathcal{B} \times \mathcal{M} | m \sqsubseteq S\}$.

Exemple 15 : D'après la relation de support ainsi définie, l'itemset $S_1 = \langle (ad)(b) \rangle$ supporte les motifs $\langle (a) \rangle$, $\langle (ad) \rangle$, $\langle (a)(b) \rangle$, $\langle (d) \rangle$, $\langle (d)(b) \rangle$ et $\langle (b) \rangle$.

Définition 18 (Environnement d'extraction de motifs séquentiels) **:** L'**environnement d'extraction de motifs séquentiels** dans une base de séquences $\mathcal{B}$ est le triplet $(\mathcal{B}, \mathcal{M}, \mathcal{R})$.

Exemple 16 : Nous représentons l'environnement d'extraction de motifs séquentiels pour la base de séquences $\mathcal{B}$ sous la forme du tableau 2.5. Au regard du grand nombre d'éléments dans $\mathcal{M}$, nous n'en sélectionnons qu'une partie pour illustrer notre propos.

	$\langle (a) \rangle$	$\langle (ab) \rangle$	$\langle (ad) \rangle$	$\langle (a)(b) \rangle$	$\langle (a)(a) \rangle$
S_1	×		×	×	
S_2	×	×		×	
S_3	×			×	×
S_4	×			×	
S_5	×	×		×	
S_6	×				
S_7	×			×	×
S_8	×			×	
S_9	×	×		×	×
S_{10}					
S_{11}	×				
S_{12}	×				
S_{13}					
S_{14}	×				

TABLE 2.5 – Représentation de l'environnement d'extraction de motifs séquentiels pour le tableau 1.1.

Définition 19 (Découverte de motifs séquentiels) **:** Étant donné un environnement d'extraction de motifs séquentiels $(\mathcal{B}, \mathcal{M}, \mathcal{R})$ et un seuil minimum de fréquence σ, le probléme de la découverte des motifs séquentiels consiste à extraire l'ensemble $MFreq(\mathcal{B}, \sigma)$.

Exemple 17 : Soit un seuil minimum de fréquence σ fixé à $0,5$. Dans le tableau 2.5, présentant notre environnement d'extraction, les motifs séquentiels sont les motifs supportés par au moins 7 séquences de $\mathcal{B}$ (i.e., de fréquence supérieure ou égale à $7/14$).

Considérons par exemple la séquence $\langle (a)(b) \rangle$. Elle est supportée par 8 éléments de la base de séquences. Son support est donc $Supp_{\mathcal{B}}(\langle (a)(b) \rangle) = 8$. Sa fréquence est $Freq_{\mathcal{B}}(\langle (a)(b) \rangle) = \frac{Supp_{\mathcal{B}}((b,d))}{|\mathcal{B}|} = \frac{8}{14} > 0,5$. Par conséquent, $\langle (a)(b) \rangle$ est un motif séquentiel dans $\mathcal{B}$. En revanche, la séquence $\langle (ab) \rangle$ n'est pas fréquente. En effet, $Freq_{\mathcal{B}}(\langle (ab) \rangle) = \frac{Supp_{\mathcal{B}}(\langle (ab) \rangle)}{|\mathcal{B}|} = \frac{3}{14} < \sigma$.

L'ensemble $MFreq(\mathcal{B}, \sigma)$ de tous les motifs séquentiels dans $\mathcal{B}$ (en considérant l'intégralité de $\mathcal{M}$) est $\{\langle (a) \rangle, \langle (a)(b) \rangle, \langle (b) \rangle, \langle (d) \rangle\}$.

Extraction de motifs séquentiels

Depuis l'introduction de la problématique d'extraction des motifs séquentiels dans [AS95] de nombreux algorithmes ont été proposés. Dans cette section, nous proposons un aperçu des principales approches et extensions. Une présentation complète est disponible dans [Raï08, Pla08]. Ce problème étant proche de celui de l'extraction des itemsets fréquents, on y retrouve les mêmes principes généraux.

Cependant, l'espace de recherche dans le cadre des séquences se révèle plus complexe que celui des itemsets. En particulier, étendre un motif séquentiel par l'ajout d'un item peut se faire selon deux manières différentes : soit en ajoutant un nouvel itemset de taille 1 à la séquence (par exemple, la séquence $\langle(ab)\rangle$ générera la séquence $\langle(ab)(c)\rangle$), soit en ajoutant un item au dernier itemset de la séquence (la séquence $\langle(ab)\rangle$ générera alors la nouvelle séquence $\langle(abc)\rangle$).

La première catégorie d'algorithmes apparue pour résoudre le problème de l'extraction de motifs séquentiels fait appel aux principes de l'algorithme *Apriori*, utilisé pour l'extraction d'itemsets fréquents. Ces approches reposent sur un parcours par niveau de l'espace de recherche s'appuyant sur le paradigme *générer-élaguer*.

Considérons en particulier l'algorithme *GSP* (*Generalized Sequential Patternns*) introduit dans [SA96], successeur des deux algorithmes pionniers *APrioriAll* et *APrioriSome* présentés dans [AS95].

GSP fait face aux contraintes liées aux données séquentielles par le biais d'une méthode de génération des candidats adaptée. Ainsi, deux motifs séquentiels s_1 et s_2 de taille k vont permettre de générer un motif séquentiel candidat de taille $k+1$ si et seulement si la séquence obtenue en supprimant le premier item de s_1 est identique à celle obtenue en supprimant de dernier item de s_2. Par exemple, si $s_1 = \langle\{a,b\}\{c\}\rangle$ et $\langle\{b\}\{c,d\}\rangle$ générera la séquence candidate $\langle\{a,b\}\{c,d\}\rangle$. La seconde particularité de *GSP* concerne la structure d'arbre de hachage utilisée pour organiser les candidats en fonction de leur préfixe. Les feuilles de cet arbre contiennent les candidats générés.

De nombreux autres algorithmes d'extraction de motifs séquentiels de type *Apriori* ont depuis été développés, tels que *PSP* [MCP98] ou *SPAM* [AFGY02].

La fin des années 1990 a vu l'essor de nouvelles approches visant à optimiser le processus d'extraction. En particulier, l'algorithme *SPADE* décrit dans [Zak01] offre deux nouveautés principales. Tout d'abord, la base fouillée est transformée dans un format vertical où chaque sous-séquence est associée à sa liste d'occurrences dans les données initiales. La deuxième proposition de *SPADE* est la définition d'une classe d'équivalence pour regrouper les motifs séquentiels. Ainsi, deux séquences de taille k sont dans la même classe d'équivalence si elles partagent le même préfixe de taille $k-1$.

L'approche par projections, déjà proposée dans le cadre de l'extraction d'itemsets fréquents [HPY00], a fait l'objet d'une adaptation dans le domaine des motifs séquentiels [PHMA+01].

De la même manière que pour les itemsets, de nombreux travaux se sont également focalisés sur la possibilité de représentation condensées. Ainsi des approches adaptées aux motifs séquentiels fermés ont été proposées [YHA03, WH04]. Cependant la particularité des données manipulées fait que d'autres types de représentation ne sont pas forcément possibles. Par exemple, dans [RCP08], les auteurs ont montré qu'il ne peut pas exister pour les motifs séquentiels de repré-

sentations condensées basées sur un calcul de support issues du cadre k-libre (non-dérivables, disjonctifs libres).

2.2.3 Itemsets inter-transactionnels

Ni les itemsets fréquents, ni les motifs séquentiels ne permettent l'extraction de motifs de la forme « *Fréquemment, l'habitant H_1 mange, puis allume l'ordinateur une heure après, puis va se coucher deux heures après* ». Un tel motif présente deux particularités qui le différencient des précédents motifs définis. La première concerne le fait que ce motif décrit un comportement propre à un habitant uniquement et non à un ensemble d'habitants. L'extraction de ce type de motifs nécessite donc de considérer une unique séquence de données (celle correspondant à l'habitant H_1 dans notre exemple) et d'extraire une suite d'activités (i.e., une sous-séquence) qui apparaît un nombre suffisant de fois à l'intérieur d'une même séquence. Ce principe tranche avec celui que nous avons développé jusqu'ici. L'intuition qui accompagne l'extraction d'un tel motif fréquent semble ne pas coïncider avec le formalisme général proposé dans la section 2.1. Pourtant, comme nous le montrons ci-dessous, le problème d'extraction de tels motifs peut facilement être réduit à un problème d'extraction de motifs classique dans une base d'objets.

La seconde particularité de ce type de motifs concerne la prise en compte du temps séparant les activités (i.e., l'heure dans notre cas d'étude). En effet, les itemsets fréquents ne tiennent pas du tout compte de la dimension temporelle. Les motifs séquentiels dans leur forme classique, quant à eux, considèrent uniquement l'ordre temporel sur les items. Le motif décrit plus haut, en revanche, considère à la fois l'ordre entre les items et l'écart de temps qui sépare ces items. Notons toutefois qu'une extension des motifs séquentiels, utilisée dans [SA96, MPT04], offre la possibilité d'intégrer des contraintes dites temporelles dans l'extraction de motifs séquentiels. Par exemple, une telle contrainte imposera d'extraire uniquement les motifs séquentiels dont les occurrences ne dépassent pas une certain intervalle de temps (i.e., le premier et le dernier itemset de la séquence sont séparés par un intervalle de temps inférieur à un paramètre fixé). Une autre contrainte temporelle consiste à imposer que l'intervalle de temps séparant deux itemsets consécutifs dans un même motif soit supérieur (ou inférieur) à un paramètre fixé. Cependant, de telles contraintes fournissent uniquement la possibilité de fixer des bornes sur les intervalles séparant les itemsets d'un même motif et non de considérer un écart de temps donné.

Ainsi, afin d'obtenir de tels motifs, [LHF98] et [TLHF99] introduisent les itemsets inter-transactionnels (appelés itemsets IT dans la suite). Les notions associées peuvent être définies comme suit.

Les séquences que nous avons considérées pour les motifs séquentiels ne conservent pas l'information de l'estampille associée à chaque itemset. Afin d'introduire les itemsets inter-transactionnels, nous définissons donc la notion de séquence étendue.

Définition 20 (Séquence étendue) **:** Une **séquence étendue** est une liste ordonnée d'itemsets notée $s = \langle I_1^{d_1} I_2^{d_2} ... I_n^{d_n} \rangle$, telle que chaque itemset I_i est associé à une estampille d_i.

Exemple 18 : Dans notre cas d'étude, les estampilles correspondent à l'heure durant laquelle les activités d'un itemset ont été enregistrées. Par la suite, nous considérerons la séquence étendue suivante comme exemple :

$$s = \langle (ab)^7 (bc)^9 (ad)^{11} (cd)^{13} (ab)^{14} (abc)^{16} \rangle.$$

Celle-ci peut se traduire par : « *À 7 heures ont été enregistrées les activités a et b, puis b et c à 9 heures, puis a et d à 11 heures, etc.* ».

La notion de séquence étendue nous permettant de représenter à la fois l'ordre défini sur les itemsets et les estampilles associées à ceux-ci, nous pouvons désormais présenter les itemsets inter-transactionnels.

Définition 21 (Itemset inter-transactionnel) **:** Un itemset inter-transactionnel est un ensemble de couples (i, k), où i est un item ($i \in \mathcal{I}$) et k est un entier positif ou nul.

Définition 22 (Itemset inter-transactionnel sur s) **:** Soient $s = \langle I_1^{d_1} I_2^{d_2} ... I_n^{d_n} \rangle$ une séquence étendue, d_k une estampille sur s et $maxSpan$ un seuil d'écart maximal. L'itemset inter-transactionnel sur s pour l'estampille d_k est l'ensemble des couples (i, k) tels que $i \in I_k$ et $d_k - d_1 \leq maxSpan$.

Exemple 19 : Soient la séquence étendue s précédemment donnée en exemple et le seuil maximum d'écart $maxSpan = 3$. L'itemset inter-transactionnel sur s pour l'estampille 11 est l'ensemble $\{(a, 0)(d, 0)(c, 2)(d, 2)(a, 3)(b, 3)\}$.

Notons qu'un itemset inter-transactionnel est un ensemble et peut être vu comme un simple itemset, défini sur un ensemble d'items enrichi $\mathcal{I}^+ = \{(i, k) \in \mathcal{I} \times \mathbb{N}\}$. Venons-en à la définition de l'environnement d'extraction des itemsets IT. Nous avons observé en introduction de cette section que les motifs recherchés ne semblent pas vérifier notre formalisme général puisque ces motifs doivent être recherchés dans un objet seulement et non dans une base d'objets. Néanmoins, le problème d'extraction des itemsets IT fréquents peut se traduire sous cette forme. Nous définissons donc la base d'objets correspondante.

Définition 23 (Base d'itemsets inter-transactionnels) **:** Soit $s = \langle I_1^{d_1} I_2^{d_2} ... I_n^{d_n} \rangle$ une séquence étendue. La **base d'itemsets inter-transactionnels** $\mathcal{B}$ pour la séquence étendue s est l'ensemble des itemsets inter-transactionnels sur s pour chaque estampille d_k (avec $k \in \{1, 2, ..., n\}$).

Exemple 20 : Considérons la séquence s et $maxSpan = 3$. La base d'itemsets IT sur s est présentée dans le tableau 2.6.

Afin de compléter l'environnement d'extraction, nous définissons également l'ensemble des motifs.

Définition 24 (Ensemble de motifs) **:** Pour la base $\mathcal{B}$ d'itemsets IT, l'ensemble de motifs $\mathcal{M}$ est l'ensemble de tous les itemsets inter-transactionnels inclus dans au moins un élément de $\mathcal{B}$, tels qu'ils contiennent au moins un couple de la forme $(i, 0)$, avec $i \in \mathcal{I}$.

Ainsi, un motif dans le cadre de l'extraction d'itemsets IT fréquents vérifie deux conditions : *(1)* il s'agit d'un itemset inter-transactionnel inclus dans au moins un élément de la base et

ID	Itemset IT
I_1	$\{(a,0)(b,0)(b,2)(c,2)\}$
I_2	$\{(b,0)(c,0)(a,2)(d,2)\}$
I_3	$\{(a,0)(d,0)(c,2)(d,2)(a,3)(b,3)\}$
I_4	$\{(c,0)(d,0)(a,1)(b,1)(a,3)(b,3)(c,3)\}$
I_5	$\{(a,0)(b,0)(a,2)(b,2)(c,2)\}$
I_6	$\{(a,0)(b,0)(c,0)\}$

TABLE 2.6 – Base d'itemsets IT correspondant à la séquence étendue s.

(2) cet itemset inter-transactionnel contient au moins un item ayant 0 pour estampille. Cette dernière condition vise à réduire la redondance entre les motifs. Par exemple, les itemsets IT $\{(a,0)(b,1)\}$ et $\{(a,1)(b,2)\}$ sont en réalité redondants, puisqu'ils expriment simplement le fait que fréquemment, b apparaît 1 heure après a. Par conséquent, seul le premier motif sera conservé par le formalisme des itemsets IT.

La relation de support, quant à elle, est identique à celle que l'on peut trouver dans le cas des itemsets fréquents puisqu'elle est liée à une inclusion ensembliste.

Définition 25 (Relation de support) **:** La **relation de support** $\mathcal{R}$ est une relation binaire entre $\mathcal{B}$ et $\mathcal{M}$ définie par $\mathcal{R} = \{(I,m) \in \mathcal{B} \times \mathcal{M} | m \subseteq I\}$.

Exemple 21 : D'après la relation de support ainsi définie, l'itemset inter-transactionnel I_1 de la base $\mathcal{B}$ supporte le motif $\{(a,0)(b,2)\}$.

Définition 26 (Extraction d'itemsets IT fréquents) **:** D'après l'environnement d'extraction d'itemsets IT formé par le triplet $(\mathcal{B}, \mathcal{M}, \mathcal{R})$ et pour un seuil minimum de fréquence σ, le problème de la découverte des itemsets IT fréquents dans la séquence étendue s consiste à extraire l'ensemble $MFreq(\mathcal{B}, \sigma)$.

Exemple 22 : D'après la séquence étendue s précédente, l'environnement d'extraction des itemsets IT est présenté dans le tableau 2.7. Comme précédemment, une partie seulement de l'ensemble des motifs est donnée.

	$\{(a,0)\}$	$\{(a,0)(b,0)\}$	$\{(a,0)(c,2)\}$	...	$\{(d,0)(b,3)\}$
I_1	×	×	×		
I_2					
I_3	×		×		×
I_4					×
I_5	×	×	×		
I_6	×	×			

TABLE 2.7 – Représentation de l'environnement d'extraction d'itemsets IT pour la séquence s.

Considérons une fréquence minimum σ fixée à 0,5. Alors l'itemset IT $I = \{(a,0)(c,2)\}$

est fréquent. En effet, $Freq_{\mathcal{B}}(I) = \frac{Supp_{db}(I)}{|\mathcal{B}|} = \frac{3}{6} \geq \sigma$. De même, les itemsets IT $\{(a,0)\}$ et $\{(a,0)(b,0)\}$ sont également fréquents.
En revanche, l'itemset IT $I' = \{(d,0)(b,3)\}$ n'est pas fréquent, puisque $Freq_{\mathcal{B}}(I) = \frac{Supp_{db}(I)}{|\mathcal{B}|} = \frac{1}{6} < \sigma$.

Extraction d'itemsets inter-transactionnels fréquents

Bien qu'ayant reçu moins d'attention que les itemsets fréquents ou les motifs séquentiels, de nombreuses approches ont été proposées dans le but d'extraire les itemsets IT fréquents dans une séquence étendue. Nous ne rentrerons pas ici dans le détails de chacune de ces approches. En effet, le problème d'extraction de tels motifs étant extrêmement proche de l'extraction d'itemsets fréquents, les algorithmes développés ont suivi les mêmes évolutions. Par conséquent, nous présentons simplement ici un historique des principales idées qui ont porté les différentes approches.

Ainsi, [LFH00] introduit deux algorithmes principalement basés sur l'algorithme *Apriori* : *E-Apriori* et *EH-Apriori*. Tandis que le premier est une adaptation directe de l'algorithme *Apriori*, le second intègre une technique de hachage particulière afin d'élaguer les candidats de taille 2.

En 2003, [TLHF03] présente l'algorithme *FITI* (*First-Intra-Then-Inter*), qui fonctionne en deux étapes. D'abord, les itemsets intra-transactionnels fréquents (i.e., les itemsets classiques) sont extraits à l'aide l'algorithme *Apriori*. Puis, les itemsets inter-transactionnels sont générés à l'aide des motifs obtenus dans la première phase.

Un nouvel algorithme a par la suite été proposé par [LW07] : *ITP-Miner*. Celui-ci a la particularité de s'appuyer sur un format vertical des données ainsi que sur une structure d'arbre particulière (*ITP-Tree*) pour extraire les motifs.

Dernièrement, [WC11] a proposé d'étendre le principe de la projection (*PrefixSpan*, *FP-Growth*) à l'extraction des itemsets inter-transactionnels. Cette adaptation, enrichie de stratégies d'élagage de l'espace de recherche, a donné jour à l'algorithme *PITP-Miner*.

2.3 Discussion

Dans ce chapitre, nous avons étudié la problématique de l'extraction de motifs fréquents d'un point de vue général. La section 2.1 a présenté un formalisme général pour l'extraction de motifs fréquents, inspiré de l'analyse de concepts formels. Dans la section 2.2, nous nous sommes focalisés sur les applications de l'extraction de motifs fréquents dans les données séquentielles et avons brièvement présenté les motifs qui nous intéresserons particulièrement dans ce travail : les motifs séquentiels et les itemsets inter-transactionnels. Notons cependant que la littérature liée à l'extraction de séquences fréquentes offre d'autres types de motifs. Par exemple, [MTIV97] introduit le problème de la découverte d'épisodes fréquents dans des données séquentielles. Un épisode est ainsi défini comme une collection partiellement ordonnée d'événements qui apparaissent fréquemment proches les uns des autres (au sens d'une fenêtre glissante de taille fixée par l'utilisateur) dans une séquence de données. Ou encore, la découverte de motifs périodiques initialement présentée dans [ORS98] vise à extraire les motifs récurrents, qui se répètent au cours du temps dans une séquence de données.

La raison du choix des motifs séquentiels et des itemsets inter-transactionnels est liée à leurs caractéristiques complémentaires. Nous résumons ci-après les différences qui les opposent. La

première provient du type de base sur laquelle s'effectue l'extraction. Les motifs séquentiels abordent l'extraction de sous-séquences apparaissant dans un nombre suffisamment grand de séquences, tandis que les itemsets IT permettent de rechercher les sous-séquences qui apparaissent fréquemment au sein d'une même séquence.

La seconde différence concerne la gestion de l'ordre lié aux items dans les sous-séquences extraites. Les motifs séquentiels se basent en effet uniquement sur l'ordre d'occurrences des items sans considérer l'estampille associée à ceux-ci. En revanche, les itemsets IT considèrent l'ordre d'occurrence des items, mais également l'écart séparant ceux-ci.

Dans le reste de ce mémoire, nous constaterons que la complémentarité observée entre ces types de motifs permet d'offrir une réponse adaptée à de nombreux cas d'application.

Finalement, ce chapitre nous a permis de préciser différentes définitions qui s'avéreront utiles pour la suite. Il a également permis de proposer un rapide tour d'horizon des principales approches d'extraction de motifs fréquents. Toutefois, en introduction de ce manuscrit, nous avons précisé que les informations contextuelles pouvaient avoir une influence sur ce qui se produit dans les données. Dans le chapitre suivant, nous étudions justement cette influence dans l'extraction de motifs fréquents.

Chapitre 3

Extraction de motifs fréquents contextuels

Sommaire

Introduction

Comme nous l'avons vu précédemment, la découverte de motifs fréquents dans les données séquentielles présente un éventail important d'applications, telles que l'étude des comportements des utilisateurs, de données issues de capteurs, de puces à ADN, etc. Pourtant, bien souvent, les données disponibles sont associées à des informations additionnelles, visant à décrire le contexte dans lequel les données ont été collectées. Le degré de connaissance offert par les motifs fréquents est limité lorsque ces informations contextuelles ne sont pas considérées.

Par exemple, dans les données liées à notre cas d'étude, un motif séquentiel peut être *"fréquemment, les habitants regardent la télé puis vont se coucher"*. Un tel motif apporte une information générale sur le comportement des habitants sans pour autant l'associer à un contexte particulier (par exemple, l'âge des personnes suivant ce comportement ou la saison).

L'importance du contexte se vérifie pourtant dans de nombreuses applications : le panier d'achats typique d'un étudiant n'est pas le même que celui d'un retraité, l'usage d'Internet n'est pas le même dans un cadre professionnel ou personnel, une pression sanguine jugée normale diffère selon l'âge ou le sexe du patient, etc.

Notre cas d'étude ne fait pas exception. Par exemple, la consommation énergétique dans un logement dépend de la saison, du jour de la semaine, du type d'habitants, etc. La prise en compte de ces informations contextuelles permet ainsi à l'expert de répondre à des questions telles que : *"Quels sont les comportements les plus représentatifs parmi les habitants âgés ?"*, *"Existe-t-il des comportements représentatifs des jeunes habitants en été ?"* ou encore *"Quels sont les comportements représentatifs qui ne dépendent pas du contexte ?"*.

Ainsi, l'extraction de motifs fréquents doit tenir compte de l'impact du contexte dans les données et fournir à l'utilisateur une vue contextuelle des connaissances extraites. Dans ce chapitre, nous définissons les motifs fréquents contextuels et proposons les algorithmes nécessaires à leur extraction. En effet, intuitivement, l'extraction de tels motifs est difficile car elle nécessite de prendre en compte les différents niveaux de généralisation/spécialisation des contextes. Par exemple, le contexte correspondant aux *jeunes habitants* est plus général que celui des *jeunes habitants en été*. Les possibilités de contextes étant nombreuses, extraire les motifs représentatifs dans chacun d'eux est particulièrement coûteux et il est donc indispensable de développer des algorithmes efficaces.

La suite de ce chapitre est organisée de la manière suivante. La section 3.1 explique pourquoi les motifs fréquents traditionnels ne sont pas adaptés pour manipuler les données contextuelles et présente un panorama des travaux similaires. Nous définissons dans la section 3.2 les notions liées aux motifs fréquents contextuels. La section 3.3 présente des propriétés importantes pour l'extraction. L'algorithme proposé est décrit dans la section 3.4. L'évaluation de notre approche est exposée dans la section 3.5. Enfin, nous concluons par une discussion dans la section 3.6.

3.1 Contexte et motifs fréquents

Dans cette section, nous montrons, dans un premier temps, pourquoi les motifs fréquents « *traditionnels* » ne conviennent pas lorsque les données étudiées sont associées à des informations

contextuelles. Puis, nous étudions les différents travaux qui peuvent être rapprochés de notre problématique.

3.1.1 Pourquoi intégrer les informations contextuelles dans le processus d'extraction ?

id	Age	Saison	Sequence
s_1	jeune	été	$\langle(ad)(b)\rangle$
s_2	jeune	été	$\langle(ab)(b)\rangle$
s_3	jeune	été	$\langle(a)(a)(b)\rangle$
s_4	jeune	été	$\langle(c)(a)(bc)\rangle$
s_5	jeune	été	$\langle(d)(ab)(bcd)\rangle$
s_6	jeune	hiver	$\langle(b)(a)\rangle$
s_7	jeune	hiver	$\langle(a)(b)(a)\rangle$
s_8	jeune	hiver	$\langle(d)(a)(bc)\rangle$
s_9	âgé	été	$\langle(ab)(a)(bd)\rangle$
s_{10}	âgé	été	$\langle(bcd)\rangle$
s_{11}	âgé	été	$\langle(bd)(a)\rangle$
s_{12}	âgé	hiver	$\langle(e)(bcd)(a)\rangle$
s_{13}	âgé	hiver	$\langle(bde)\rangle$
s_{14}	âgé	hiver	$\langle(b)(a)(e)\rangle$

TABLE 3.1 – Une base contextuelle de séquences.

Exemple 23 : Le tableau 3.1 rappelle la base de séquences $\mathcal{B}$ décrivant les activités effectuées par différents habitants présentée en introduction de ce mémoire. Dans la première colonne est fourni l'identifiant de chaque séquence. Ici, a, b, c, d, e sont des activités. Les colonnes *Age* et *Saison* sont des informations additionnelles relatives aux séquences. Elles ne sont pas considérées dans l'extraction de motifs séquentiels traditionnels. La taille de $\mathcal{B}$ est $|\mathcal{B}| = 14$. La première séquence décrit la séquence d'activités d'un habitant d'identifiant s_1 : il a suivi les activités a et d, puis l'activité b.

Dans la suite, nous fixons le seuil minimum de fréquence σ à $0,5$. Considérons la séquence $s = \langle(a)(b)\rangle$. Sa fréquence dans $\mathcal{B}$ est $Freq_{\mathcal{B}}(s) = 8/14$. Ainsi, $Freq_{\mathcal{B}}(s) \geq \sigma$ et s est un motif séquentiel dans $\mathcal{B}$.

En considérant l'exemple précédent, les informations contextuelles disponibles sont l'âge des habitants (jeune ou âgé) et la saison (été ou hiver) [1]. Un contexte dans ce cas pourra être *habitant jeune en été* ou encore *habitant âgé pour n'importe quelle saison.* Afin de comprendre les inconvénients posés par la découverte de motifs fréquents traditionnels dans de telles données, considérons les deux exemples suivants.

1. Nous limitons le nombre de valeurs par dimension à 2 afin de simplifier l'exemple.

Cas 1. La séquence $s = \langle(a)(b)\rangle$ est fréquente dans $\mathcal{B}$. Pourtant, ce motif est seulement représentatif des jeunes habitants : 7 *jeunes* habitants sur 8 supportent cette séquence, contre seulement 1 habitant *âgé* sur 6. Nous dirons par la suite que ce motif n'est pas « *général dans l'ensemble de la base* », i.e., il n'est fréquent que dans une sous-partie de la base et non dans tous les contextes. En revanche, $\langle(a)(b)\rangle$ est « *générale chez les jeunes habitants* » car elle est fréquente à la fois chez les jeunes en été (5 habitants sur 5) et chez les jeunes en hiver (2 habitants sur 3).

Cas 2. La séquence $s' = \langle(bd)\rangle$ n'est pas fréquente dans $\mathcal{B}$ (6 habitants sur 14 la supportent). Pourtant, elle est fréquente pour les habitants *âgés* : 5 sur 6 la supportent.

L'extraction de motifs fréquents traditionnels peut ainsi mener à considérer certains comportements dépendant d'un contexte comme représentatifs de l'ensemble des données (*Cf.* Cas 1) alors qu'ils ne sont, en réalité, représentatifs que d'une sous-partie de la base seulement. À l'inverse, ils peuvent également mener à ne pas les considérer comme représentatifs parce que le contexte associé n'est lui-même pas fréquent (*Cf.* Cas 2). Ainsi, dès lors que des informations contextuelles sont disponibles, leur prise en compte apporte une réelle valeur ajoutée pour les connaissances extraites.

3.1.2 Motifs et contextes dans les travaux existants

Nous avons montré que les motifs fréquents « *traditionnels* » ne tirent pas parti des informations contextuelles. Néanmoins, d'autres travaux liés à la découverte de motifs peuvent être rapprochés de notre problématique. Nous en discutons ci-dessous les limites.

Extraction de motifs multidimensionnels D'un point de vue général, l'extraction de motifs multidimensionnels, introduite dans [KHC97], vise à intégrer des données décrites sur plusieurs dimensions. Dans ce cadre, deux approches se rapprochent de la problématique que nous abordons dans ce chapitre.

La plus proche est sans doute celle décrite dans [GLWX01]. Les auteurs s'intéressent aux « *circonstances* » dans lesquelles un itemset est fréquent et décrivent ces circonstances par le biais de dimensions équivalentes à celles que nous utilisons dans notre cas d'étude. Cependant, les auteurs considèrent uniquement le support absolu (et non la fréquence). Ils remarquent ainsi que, lorsqu'un motif est fréquent dans un contexte d'après la mesure de support absolu, alors il est forcément fréquent dans les contextes plus généraux. Par exemple, si un motif m est fréquent chez les jeunes en été, alors il sera fréquent chez les jeunes (pour n'importe quelle saison) ou encore dans l'ensemble de la base (i.e., pour n'importe quel âge et n'importe quelle saison). Cette propriété est exploitée pour extraire les contextes les plus spécifiques où un motif est fréquent. Cette approche possède deux inconvénients qui la rendent inapplicable pour notre problématique. D'abord, nous affirmons que la mesure de support absolu n'est pas adaptée pour extraire les motifs dans les données contextuelles. En effet, celle-ci est dépendante de la répartition des données dans les différents contextes. Par exemple, si les données contiennent un grand nombre d'habitants jeunes et moins d'habitants âgés (comme c'est le cas dans notre base exemple) alors les contextes impliquant des jeunes habitants seront favorisés. Par conséquent, nous considérons dans la suite la fréquence et non le support comme mesure principale. La principale propriété exploitée dans [GLWX01] n'est donc plus applicable. Un deuxième inconvénient est lié à l'itemset

auquel on souhaite associer des circonstances qui est supposé déjà connu. Ainsi, le problème posé est « *Étant donné un itemset, quelles sont les circonstances associées ?* », tandis que nous nous intéressons à la question « *Étant donné une base d'objets associée à des informations contextuelles, quels sont tous les motifs généraux dans un contexte ?* ». Cette deuxième question est bien plus difficile puisqu'elle nécessite d'extraire l'ensemble de tous les motifs.

Dans ce mémoire, nous nous focalisons principalement sur les données séquentielles. Dans ce type de données, quelques travaux ont abordé la problématique de l'extraction de motifs séquentiels multidimensionnels. Les premiers travaux sont ceux de [PHP+01]. L'approche proposée considère des données en tous points identiques à celle que nous décrivons dans le tableau 3.1, i.e., chaque élément d'une base de séquences est associé à des valeurs sur des dimensions contextuelles. Les motifs séquentiels multidimensionnels définis dans ces travaux sont composés d'une séquence et d'un contexte donné, tels que la combinaison des deux est fréquente dans l'ensemble de la base. Dans notre cas d'étude, pour un seuil minimum de fréquence $\sigma = 0.5$, la séquence $\langle(a)(b)\rangle$ associée au contexte correspondant aux *jeunes* forme un motif séquentiel multidimensionnel, avec une fréquence de 7/14 (car dans l'ensemble de la base, il existe 7 jeunes qui supportent $\langle(a)(b)\rangle$). On remarque alors aisément qu'il n'est pas possible d'extraire des motifs séquentiels multidimensionnels associés au contexte des habitants âgés sur cette base, puisque le contexte est lui-même infréquent dans la base (il existe seulement 6 habitants âgés dans la base). Bien sûr, baisser le seuil de fréquence minimum est une solution, mais dans ce cas, le problème demeurera pour les contextes plus spécifiques (par exemple, le contexte correspondant aux habitants âgés en été, qui ne compte que trois éléments) et le nombre de motifs associés aux contextes les plus fréquents sera d'autant plus important.

Notons que la définition des motifs séquentiels multidimensionnels a été largement enrichie depuis, notamment dans [Pla08], de manière à intégrer la multidimensionnalité dans les items (au travers d'items multidimensionnels) et la possibilité de considérer des hiérarchies sur les dimensions prises en compte. Par exemple, on pourra considérer que la valeur *jeune* sur la dimension *Age* est une généralisation des valeurs plus spécifiques *adolescent* et *jeune adulte*. Malgré cette amélioration, les inconvénients déjà soulignés pour les travaux de [PHP+01] restent vrais : pour être extrait, un motif multidimensionnel doit être fréquent sur l'ensemble de la base.

Le même inconvénient s'applique encore sur [SZ05] qui, dans des données similaires à celles décrites dans [Pla08], enrichit la notion de support en introduisant des fonctions de similarité pour remplacer le comptage binaire (présent/absent) du support habituellement utilisé pour l'extraction de motifs fréquents.

Par conséquent, aucune des approches développées dans le cadre de l'extraction de motifs fréquents dans les données multidimensionnelles ne répond entièrement à nos attentes, qui peuvent être résumées ainsi :

- Nous souhaitons rechercher les motifs qui sont représentatifs (au sens de la fréquence) d'un contexte, même si ce contexte est lui-même très peu fréquent dans l'ensemble de la base. Les travaux existants qui soit considèrent uniquement le support absolu comme mesure [GLWX01], soit recherchent les motifs fréquents dans l'ensemble de la base [PHP+01, SZ05, Pla08], ne répondent pas à ce premier objectif.
- De plus, à notre connaissance, aucune approche ne tient compte de ce que nous appelons la « *généralité* » d'un motif (*Cf.* Cas 1 dans la section précédente) dans un contexte. Pourtant,

cette notion nous paraît fondamentale dans l'étude contextualisée des motifs fréquents. Comme nous l'avons souligné dans la section précédente, elle permet en particulier de corriger le déséquilibre de la répartition des données dans les différents contextes.

Extraction de motifs d'intérêt Une autre problématique est liée à l'extraction de motifs dans les données contextuelles : la découverte de motifs d'intérêt. De manière générale, il s'agit d'extraire les motifs qui respectent une certaine contrainte d'intérêt, où l'intérêt est mesuré via des propriétés statistiques d'une mesure, dite mesure d'intérêt. De nombreuses mesures d'intérêt ont été proposées et étudiées dans la littérature [TKS02, GH06, JHZ10]. Un exemple consiste à extraire les motifs émergents, i.e., les motifs qui sont significativement plus fréquents dans un sous-ensemble de données par rapport au reste de la base. Appliquée aux données contextuelles, on peut imaginer que la découverte de motifs émergents dans un contexte peut répondre à notre problématique en révélant les motifs qui y sont discriminants. Cependant, nous nous intéressons uniquement dans ce chapitre à la notion de fréquence d'un motif dans les différents contextes. Nous verrons pourtant dans le chapitre 4 que les mesures d'intérêt (i.e., autres que la fréquence) ne sont pas incompatibles avec les données contextuelles.

3.2 Motifs fréquents contextuels

Nous proposons dans cette section une description formelle de la notion de contexte et définissons les notions nécessaires pour appréhender les motifs fréquents contextuels. Les nouveaux concepts exposés dans cette section étendent le formalisme général des motifs fréquents présenté dans le chapitre 2. Par conséquent, les notions sont présentées ici de manière généralisée et sont exploitables pour tous les types de motifs fréquents satisfaisant ce formalisme (en particulier les itemsets, les motifs séquentiels et les itemsets inter-transactionnels).

3.2.1 Données contextuelles

Dans un premier temps, nous définissons la notion de base contextuelle d'objets, dans laquelle chaque objet est associée à diverses informations contextuelles.

Définition 27 (Base contextuelle d'objets) **:** Une **base contextuelle d'objets** $\mathcal{CB}$ est définie comme une relation $\mathcal{R}(\mathcal{B}, D_1, ...D_n)$, où $\mathcal{B}$ est un ensemble d'objets et $dom(D_i)$, pour $1 \leq i \leq n$, est l'ensemble de toutes les valeurs possibles de D_i. D_1, D_2, ... D_n sont appelées les **dimensions contextuelles** de $\mathcal{CB}$. Un **tuple** $u \in \mathcal{CB}$ est noté $\prec o, d_1, ..., d_n \succ$.

Exemple 24 : Le tableau 3.1 présente une base contextuelle de séquences. Les dimensions contextuelles sont données par les colonnes *Age* et *Saison*. Nous remarquons que le domaine de la dimension *Age* contient les valeurs *jeune* et *âgé*, i.e., $dom(Age) = \{jeune, âgé\}$. De même, $dom(Saison) = \{été, hiver\}$.

Un tuple de cette base est, par exemple, $\prec \langle (ad)(b) \rangle, jeune, été \succ$, signifiant que la séquence d'activités $\langle (ad)(b) \rangle$ a été enregistrée pour un habitant jeune en été.

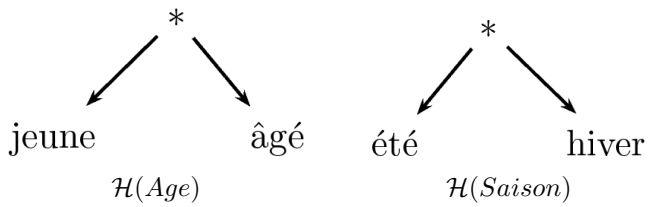


FIGURE 3.1 – Hiérarchies sur les dimensions *Age* et *Saison*

.

Les valeurs sur chaque dimension contextuelle peuvent être organisées sous la forme d'une hiérarchie. Pour $1 \leq i \leq n$, $dom(D_i)$ peut être étendu à $dom'(D_i)$, où $dom(D_i) \subseteq dom'(D_i)$. De plus, $dom'(D_i)$ est associé à un ordre partiel $\subseteq_{D_i}$ tel que $dom(D_i)$ est l'ensemble des éléments minimaux de $dom'(D_i)$ selon $\subseteq_{D_i}$.

Définition 28 (Hiérarchie sur la dimension D_i) **:** L'ensemble partiellement ordonné $(dom'(D_i), \subseteq_{D_i})$ est la **hiérarchie sur la dimension** D_i, notée $\mathcal{H}_{D_i}$.

Exemple 25 : Considérons les hiérarchies $\mathcal{H}_{Age}$ et $\mathcal{H}_{Saison}$ présentées dans la figure 3.1.
Dans cet exemple, $dom'(Age) = dom(Age) \cup \{*\}$. L'ordre partiel $\subseteq_{Age}$ est défini tel que *jeune* $\subseteq_{Age} *$ and *âgé* $\subseteq_{Age} *$. Ainsi, les valeurs *jeune* et *âgé* sont toutes deux des spécialisations de la valeur $*$ sur la dimension *Age*.
De même, $dom'(Saison) = dom(Saison) \cup \{*\}$. L'ordre partiel $\subseteq_{Saison}$ est défini tel que *été* $\subseteq_{Saison} *$ and *hiver* $\subseteq_{Saison} *$. Ainsi, les valeurs *été* et *hiver* sont toutes deux des spécialisations de la valeur $*$ sur la dimension *Saison*.

D'après la définition des dimensions contextuelles, nous pouvons définir la notion de contexte comme suit, ainsi que l'ordre de généralisation/spécialisation sur l'ensemble des contextes.

Définition 29 (Contexte) **:** Un *contexte* c dans $\mathcal{CB}$ est noté $[d_1, ...d_n]$, où $d_i \in dom'(D_i)$. Si pour $1 \leq i \leq n$, $d_i \in dom(D_i)$, alors c est un *contexte minimal*.

Définition 30 (Ordre sur les contextes) **:** Soient c_1 et c_2 deux contextes dans $\mathcal{CB}$, tels que $c_1 = [d_1^1, ..., d_n^1]$ et $c_2 = [d_1^2, ...d_n^2]$. Alors $c_1 \leq c_2$ si et seulement si $\forall i$ avec $1 \leq i \leq n$, $d_i^1 \subseteq_{D_i} d_i^2$. S'il existe un entier i avec $1 \leq i \leq n$ tel que $d_i^1 \subset_{D_i} d_i^2$, alors $c_1 < c_2$. Dans ce cas, c_1 est dit ***plus spécifique*** que c_2 et c_2 est ***plus général*** que c_1. De plus, si $c_1 \not\geq c_2$ et $c_1 \not\leq c_2$, alors c_1 et c_2 sont **incomparables**.

Exemple 26 : La base contextuelle de séquences présentée dans le tableau 3.1 se compose de quatre contextes minimaux : $[j, h]$, $[j, e]$, $[a, h]$ et $[a, e]$, où j, a, h et e représentent respectivement *jeune*, *âgé*, *hiver* et *été*. Le contexte $[*, *]$ est plus général que $[j, *]$. $[j, *]$ et $[*, h]$ sont en revanche incomparables.

Définition 31 (Hiérarchie de contextes) **:** Soit $\mathcal{C}$ l'ensemble de tous les contextes de $\mathcal{CB}$. L'ensemble partiellement ordonné $(\mathcal{C}, \leq)$ de tous les contextes joint à l'ordre partiel $\leq$ constitue la *hiérarchie de contextes*, notée $\mathcal{H}$. Etant donné deux contextes c_1 et c_2 tels que $c_1 < c_2$, c_1 est un *descendant* de c_2 et c_2 est un *ancêtre* de c_1.

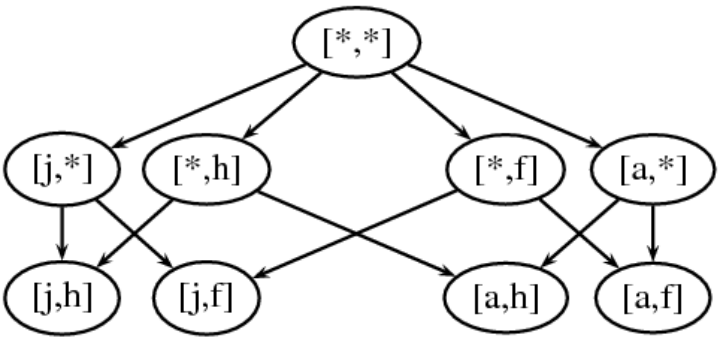


FIGURE 3.2 – La hiérarchie de contextes $\mathcal{H}$.

Exemple 27 : La figure 3.2 montre une représentation visuelle de $\mathcal{H}$ pour la base contextuelle de séquences $\mathcal{CB}$ associée aux hiérarchies précédemment définies sur les dimensions *Age* et *Saison*.

Nous pouvons désormais considérer les tuples de $\mathcal{CB}$ conformément aux contextes définis plus tôt et manipuler chaque contexte avec la base de séquences qui lui est associée.

Définition 32 (Contexte d'un tuple) **:** Soit $u = \prec id, s, d_1, ...d_n \succ$ un tuple dans $\mathcal{CB}$. Le contexte c tel que $c = [d_1, ...d_n]$ est appelé le *contexte de* u. Notons que ce contexte est minimal puisque $\forall i$ tel que $1 \leq i \leq n$, $d_i \in dom(D_i)$.

Soit u un tuple dans $\mathcal{CB}$ et c le contexte de u. Un contexte c' *est associé à* u (et u est associé à c') si et seulement si $c' \geq c$.

Définition 33 (Base d'objets d'un contexte) **:** Soient $c = [d_1, ...d_n]$ un contexte (minimal ou non) dans $\mathcal{CB}$ et $\mathcal{U}$ l'ensemble des tuples contenus par c. La *base d'objets de* c, notée $\mathcal{B}(c)$, est l'ensemble des tuples $\prec id, s \succ$ tels que $\exists u \in \mathcal{U}$ avec $u = \prec id, s, d_1, ...d_n \succ$. Nous définissons la *taille d'un contexte* c, notée $|c|$, comme la taille de sa base de séquences, i.e., $|c| = |\mathcal{B}(c)|$.

Exemple 28 : Dans le tableau 3.1, $\mathcal{B}([a, *]) = \{s_9, s_{10}, s_{11}, s_{12}, s_{13}, s_{14}\}$ et $|[a, *]| = 6$.

3.2.2 Motifs contextuels

Nous avons montré dans la section précédente comment une base contextuelle de séquences peut être décomposée en s'appuyant sur les contextes. Nous pouvons à présent définir la notion de motif fréquent associée.
Soient un contexte c et un motif m.

Définition 34 (Motif c-fréquent) **:** m est un **motif fréquent dans** c ou c-fréquent, si et seulement si m est fréquent dans $\mathcal{B}(c)$, i.e., si $Freq_{\mathcal{B}(c)}(s) \geq \sigma$. Par la suite, nous noterons $Freq_{\mathcal{B}(c)}(m)$ par $Freq_c(m)$.

Comme nous l'avons montré dans la section précédente, nous nous intéressons aux motifs représentatifs d'un contexte, i.e., dont la fréquence se propage parmi les descendants de ce contexte. Afin de manipuler un contexte et ses descendants, nous définissons d'abord l'ensemble des composants d'un contexte.

Définition 35 (Ensemble des composants) **:** L'*ensemble des composants* de c, noté $\hat{c}$, est défini comme l'ensemble composé de c et de tous ses descendants, i.e.,

$$\hat{c} = \{c' \in \mathcal{H} | c' \leq c\}.$$

Exemple 29 : L'ensemble des composants du contexte $[j, *]$ est constitué de $[j, *]$, $[j, e]$ et $[j, h]$, i.e., $\widehat{[j, *]} = \{[j, *], [j, e], [j, h]\}$.

Nous pouvons, à présent, définir les motifs dont la fréquence se propage dans tous les composants d'un contexte, appelés motifs généraux.

Définition 36 (Motif c-général) **:** Le motif m est **général dans** c ou c-général, si et seulement si $\forall c' \in \hat{c}$, m est c'-fréquent.

Séquences	$[j, h]$	$[j, e]$	$[a, h]$	$[a, e]$
$\langle(a)\rangle$	**5/5**	**3/3**	**2/3**	**2/3**
$\langle(b)\rangle$	**5/5**	**3/3**	**3/3**	**3/3**
$\langle(d)\rangle$	2/5	1/3	**3/3**	**2/3**
$\langle(e)\rangle$	0/5	0/3	0/3	**3/3**
$\langle(a)(b)\rangle$	**5/5**	**2/3**	1/3	0/3
$\langle(b)(a)\rangle$	0/5	**2/3**	**2/3**	**2/3**
$\langle(bd)\rangle$	1/5	0/3	**3/3**	**2/3**

TABLE 3.2 – Motifs séquentiels dans les contextes minimaux de $\mathcal{CB}$.

Exemple 30 : Le tableau 3.2 présente, d'après la base contextuelle de séquences associée au tableau 3.1, les séquences fréquentes dans au moins un contexte minimal, ainsi que leur fréquence (de la forme $Supp_c(s)/|\mathcal{B}(c)|$). Lorsque la fréquence est affichée en gras, la séquence est fréquente dans le contexte minimal associé.

Considérons le contexte $[a, *]$ (correspondant aux habitants âgés). D'après la définition 36, une séquence s est générale dans le contexte $[a, *]$ si et seulement si s est fréquente dans tous les composants de $[a, *]$, i.e., dans le contexte lui-même $[a, *]$, ainsi que dans ses descendants $[a, h]$ et $[a, e]$.

Toutes les séquences $\langle(a)\rangle$, $\langle(b)\rangle$, $\langle(d)\rangle$, $\langle(b)(a)\rangle$ et $\langle(bd)\rangle$ vérifient ces conditions. Elles sont $[a,*]$-générales. En revanche, la séquence $\langle(e)\rangle$ est fréquente dans $[a,*]$ (elle est supportée par 3 habitants âgés sur 6) mais pas dans son descendant $[a,h]$. Par conséquent, $\langle(e)\rangle$ n'est pas générale dans $[a,*]$.

La propriété de c-généralité assure qu'un motif fréquent dans un contexte est également fréquent dans tous les descendants de ce contexte. De tels motifs ne sont pas sensibles au problème de la répartition des données dans les contextes mis en lumière dans la section 3.1.1. Par exemple, la séquence $\langle(a)(b)\rangle$ qui est fréquente dans le contexte $[*,*]$ (i.e., dans l'ensemble de la base de séquences) n'est pas générale dans ce même contexte. Elle n'est en effet pas fréquente dans le contexte correspondant aux habitants âgés (i.e., $[a,*]$). Notons par ailleurs que, par définition, l'ensemble des motifs généraux dans un contexte est inclus dans l'ensemble des motifs fréquents de ce contexte. Les motifs généraux dans un contexte c peuvent donc être considérés comme le sous-ensemble des motifs fréquents de c qui possèdent une propriété particulière, la c-généralité, vis-à-vis des descendants de ce contexte.

Définition 37 (Motif fréquent contextuel) **:** Un **motif fréquent contextuel** est un couple $\alpha = (c, m)$, tel que m est c-général. α est alors **généré par** m.

L'ensemble de tous les motifs fréquents contextuels de $\mathcal{CB}$ pour un seuil de fréquence minimum σ est noté $MFCont(\mathcal{CB}, \sigma)$.

Exemple 31 : L'exemple précédent montre que la séquence $s = \langle(b)(a)\rangle$ est générale dans le contexte $[a,*]$. Par conséquent, le couple $([a,*], s)$ est un motif contextuel.

3.2.3 Stratégie de sélection des motifs fréquents contextuels

L'ensemble des motifs fréquents contextuels dans $\mathcal{CB}$ comporte des redondances. En effet, pour un même motif m, il peut exister plusieurs motifs fréquents contextuels générés par m, de la forme (c_1, m), (c_2, m), etc. En particulier, notons qu'un motif général dans un contexte c est, par définition, également général dans tous les descendants de c. Par exemple, la séquence $s = \langle(b)(a)\rangle$ est générale dans $[a,*]$, mais également dans $[a,e]$ et $[a,h]$ et génère donc les trois motifs fréquents contextuels correspondant : $([a,*], s)$, $([a,e], s)$ et $([a,h], s)$. Or, les deux derniers peuvent être déduits du premier.

Plus formellement, notons $\{c_1, c_2, ..., c_n\}$ l'ensemble de tous les descendants de c. Il est possible, par simple application de la c-généralité, de déduire les motifs fréquents contextuels (c_1, m), (c_2, m), ..., (c_n, m) à partir du seul motif fréquent contextuel (c, m). Nous proposons par conséquent d'extraire uniquement les motifs fréquents contextuels maximaux par le contexte que nous définissons de la manière suivante.

Définition 38 (Maximal par le contexte) **:** Un motif contextuel (c, m) est dit *maximal par le contexte* si il n'existe pas de contexte c' tel que c' est un ancêtre de c et (c', m) est un motif contextuel, i.e.,

$$\nexists c' \in \mathcal{H} | (c' > c) \text{ et } m \text{ est } c'\text{-général.}$$

Exemple 32 : Considérons la séquence $s = \langle (b)(a) \rangle$. Les supports de s dans les contextes minimaux de $\mathcal{CB}$ sont présentés ci-dessous.

	$[j,h]$	$[j,e]$	$[a,h]$	$[a,e]$
$\langle (b)(a) \rangle$	$0/5$	$\mathbf{2/3}$	$\mathbf{2/3}$	$\mathbf{2/3}$

Le couple $([a,*],s)$ est un motif fréquent contextuel (voir exemple précédent). De plus, s n'est pas $[*,*]$-générale (car il existe des descendants de $[*,*]$ dans lesquels s n'est pas fréquente). Par conséquent, $([a,*],s)$ est un **motif fréquent contextuel maximal par le contexte**.

L'ensemble des motifs fréquents contextuels maximaux par le contexte constitue une représentation condensée de l'ensemble total des motifs fréquents contextuels. En effet, tout motif fréquent contextuel non maximal peut être déduit d'un motif maximal.

3.3 Extraction de motifs fréquents contextuels

Dans cette section, nous nous penchons sur la question suivante : *Comment, à partir d'une base contextuelle de séquences, extraire l'ensemble des motifs fréquents contextuels ?*

Une approche naïve consiste à extraire les motifs fréquents indépendamment dans chaque élément de la hiérarchie des contextes, puis pour chaque contexte à éliminer les motifs non-généraux. Cette approche soulève cependant deux difficultés :

- **Les contextes à fouiller sont nombreux.** En effet, le nombre d'éléments d'une hiérarchie de contextes est $\prod_{i=1}^{n} |dom'(D_i)|$, où $D_1, ...D_n$ sont les dimensions contextuelles. En comparaison, le nombre de contextes minimaux est $\prod_{i=1}^{n} |dom(D_i)|$. Par exemple, la hiérarchie de contextes utilisée dans ce chapitre, bien que très simple, contient 9 contextes pour seulement 4 contextes minimaux.

- **Éliminer les motifs séquentiels n'ayant pas les propriétés requises est coûteux**. En effet, vérifier qu'un motif est général dans un contexte c donné nécessite de vérifier sa fréquence dans tous les autres contextes de la hiérarchie.

Afin de surmonter ces difficultés, nous étudions les propriétés de la hiérarchie de contextes et montrons que les motifs fréquents contextuels peuvent être générés en considérant uniquement les motifs fréquents des contextes minimaux. Dans ce but, nous définissons tout d'abord la décomposition d'un contexte comme suit.

Définition 39 (Décomposition minimales d'un contexte) **:** Soit un contexte c dans $\mathcal{CB}$. La **décomposition minimale** de c dans $\mathcal{CB}$, notée $decomp(c)$, est l'ensemble non-vide $\{c_1, c_2, ..., c_n\}$ des composants minimaux de c.

La décomposition minimale, appelée décomposition par la suite, est donc constituée de l'ensemble des éléments minimaux de $\hat{c}$.

Exemple 33 : La décomposition minimale de $[j, *]$ est $\{[j, h], [j, e]\}$.

D'après la définition de $\mathcal{B}(c)$, nous pouvons immédiatement dégager plusieurs propriétés intéressantes sur la décomposition de c.

Propriété 2 : La décomposition d'un contexte c dans $\mathcal{CB}$ vérifie les propriétés suivantes :

1. $\bigcap_{i=1}^{n} \mathcal{B}(c_i) = \emptyset$;
2. $\bigcup_{i=1}^{n} \mathcal{B}(c_i) = \mathcal{B}(c)$;
3. $|c| = |\mathcal{B}(c)| = \sum_{i=1}^{n} |c_i|$;
4. $Supp_c(m) = \sum_{i=1}^{n} Supp_{c_i}(m)$.

Les propriétés de la décomposition d'un contexte impliquent le lemme suivant.

Lemme 1 : Soit un contexte c, tel que $decomp(c) = \{c_1, c_2, ..., c_n\}$. Si $\forall i \in \{1, ..., n\}$, m est fréquent dans c_i (respectivement n'est pas fréquent), alors m est fréquent dans c (respectivement n'est pas fréquent dans c). De plus, m est fréquent (respectivement n'est pas fréquent) dans les descendants de c.

Démonstration : *Pour tout c_i tel que $i \in \{1, ..n\}$, $Freq_{c_i}(m) \geq \sigma \times |c_i|$. Cela signifie que $\sum_{i=1}^{k} Freq_{c_i}(m) \geq \sum_{i=1}^{n} \sigma \times |c_i|$. Cependant, $\sum_{i=1}^{n} \sigma \times |c_i| = \sigma \times \sum_{i=1}^{n} |c_i| = \sigma \times |c|$. Comme $\sum_{i=1}^{k} Freq_{c_i}(m) = Freq_c(m)$, $Freq_c(m) \geq \sigma \times |c|$.*

Soit un contexte c' tel que $c > c'$. Alors $decomp(c') \subseteq decomp(c)$, i.e., m est un motif fréquent dans chaque élément de $decomp(c')$. Par application du résultat précédent, m est un motif fréquent dans c'.

Un raisonnement similaire est appliqué si m n'est fréquent dans aucun des éléments de $decomp(c)$. □

Une conséquence immédiate du lemme 1 est la redéfinition de la notion de c-généralité, en ne tenant compte que de la décomposition des contextes de la hiérarchie.

Dans la suite de cette section, nous notons $\mathcal{F}_m$ l'ensemble des contextes minimaux dans lesquels m est fréquent. En exploitant le lemme 1, nous constatons que m est c-général si et seulement si $decomp(c) \subseteq \mathcal{F}_m$. L'ensemble des contextes vérifiant ces conditions est appelé la *couverture* de $\mathcal{F}_m$ et noté $cov(\mathcal{F}_m)$. Nous proposons dans la section 3.4 un algorithme visant à retrouver la couverture de $\mathcal{F}$ à partir de la hiérarchie de contextes, où $\mathcal{F}$ est un ensemble non vide de contextes minimaux.

Exemple 34 : Soit $\mathcal{F} = \{[j,h],[j,f],[a,f]\}$, alors $cov(\mathcal{F}) = \{[j,*],[*,f]\}$ et $([j,*],s)$ et $([*,f],s)$ sont les motifs séquentiels contextuels générés par s.

Théorème 1 : Soit $\mathcal{M}^f$ l'ensemble des motifs fréquents dans au moins un contexte minimal de $\mathcal{CB}$. L'ensemble des motifs fréquents contextuels $MFCont(\mathcal{CB},\sigma)$ est l'ensemble de tous les couples (c,m) où $m \in \mathcal{M}^f$ et (c,m) est généré par m.

Démonstration : *Ce résultat est une conséquence immédiate de la définition d'un motif c-général. En effet, si m n'est fréquent dans aucun contexte minimal, i.e., $\mathcal{F} = \emptyset$, alors il n'est fréquent dans aucun élément de la hiérarchie de contextes (voir lemme 1) et il n'existe aucun contexte c tel que m est c-général. Ainsi, tout motif séquentiel contextuel est généré par un motif qui est fréquent dans au moins un contexte minimal.* □

Le théorème 1 est essentiel dans le problème de l'extraction de motifs fréquents contextuels. En effet, il assure que tous les motifs contextuels peuvent être déduits des motifs fréquents des contextes minimaux. Dans la section 3.4, nous nous appuyons sur les propriétés des motifs séquentiels contextuels pour proposer un algorithme d'extraction efficace.

3.4 Algorithmes

Dans cette section, nous exploitons les propriétés mises en lumière pour proposer un algorithme qui, à partir d'une base contextuelle d'objets, extrait tous les motifs fréquents contextuels maximaux par le contexte. Nous abordons cette section en suivant les trois étapes suivantes :

Extraction des motifs fréquents Nous avons vu dans la section précédente que générer tous les motifs fréquents contextuels de la base nécessite dans un premier temps d'extraire tous les motifs qui sont fréquents dans au moins un contexte minimal. Nous verrons donc dans la sous-section 3.4.1 comment les algorithmes classiques d'extraction de motifs fréquents peuvent être adaptés pour accomplir efficacement cette étape.

Génération des motifs fréquents contextuels Pour chaque motif extrait dans la première étape, le cœur de notre approche génère les motifs fréquents contextuels correspondant en exploitant les propriétés dévoilées dans la section 3.3.

Optimisation du processus global Les deux premières étapes sont à l'origine indépendantes. Cependant, nous verrons qu'il est possible de mieux assembler ces étapes afin d'optimiser l'ensemble du processus.

3.4.1 Extraction des motifs fréquents

Selon les propriétés mises en lumière dans la section 3.3, un motif fréquent contextuel (c,m) peut être généré à partir d'un motif m et de l'ensemble non-vide de contextes minimaux où il est fréquent. Dans la suite, nous notons $\mathcal{F}_m$ cet ensemble. Par conséquent, la première étape de notre algorithme vise à extraire chacun de ces motifs ainsi que l'ensemble de contextes minimaux dans lesquels ces motifs sont fréquents.

Une première approche pourrait consister à extraire les motifs fréquents dans chacun des contextes minimaux $c_1, c_2, ..., c_n$ de la hiérarchie. Pour chaque motif extrait, une étape de post-traitement peut ensuite être réalisée pour rechercher les contextes minimaux où le motif a été extrait. Une telle approche a l'avantage d'être générique puisque tout algorithme d'extraction de motifs fréquents peut être appliqué sans modification préalable.

Néanmoins, nous montrons ci-dessous qu'il est possible d'adapter les algorithmes existants afin d'obtenir les motifs souhaités plus efficacement en appliquant l'idée suivante : plutôt qu'extraire les motifs fréquents séparément dans chacun des contextes minimaux, nous pouvons extraire les motifs sur l'ensemble de la base et indexer leur fréquence en fonction des différents contextes minimaux. Bien que cette approche requiert une modification de l'algorithme d'extraction employé, elle possède l'avantage suivant : chaque motif est associé à l'ensemble des contextes minimaux où il est fréquent sans post-traitement.

Dans la suite de cette section, nous appliquons cette deuxième approche sur l'algorithme *PrefixSpan* [PHMA+01] dédié à l'extraction de motifs séquentiels. Dans l'exemple suivant, nous décrivons le principe de *PrefixSpan* sur notre base de séquences sans considérer les informations contextuelles, afin de mieux souligner les adaptations proposées par la suite.

Exemple 35 : Un premier parcours de la base de séquences extrait tous les items fréquents de la base, i.e., les motifs séquentiels de la forme $\langle(i)\rangle$, où i est un item. Dans l'exemple, on obtient les motifs : $\langle(a)\rangle$, $\langle(b)\rangle$, $\langle(d)\rangle$. On ne trouve pas les motifs $\langle(c)\rangle$ et $\langle(e)\rangle$ qui ne sont pas fréquents.

Par conséquent, l'ensemble des motifs séquentiels dans $\mathcal{B}$ peut être partitionné en sous-ensembles, chacun d'eux étant l'ensemble des motifs séquentiels ayant $\langle(i)\rangle$ pour préfixe. *PrefixSpan* repose sur le fait que ces sous-ensembles peuvent être extraits des **bases projetées** de chaque préfixe, i.e., pour chaque $\langle(i)\rangle$. Une base projetée contient, pour chaque séquence de $\mathcal{B}$, sa sous-séquence contenant tous les items fréquents suivant la première occurrence du préfixe donné. Une telle sous-séquence est appelée **postfixe**. Si le premier item x du postfixe est présent dans le même itemset que le dernier item du préfixe, le postfixe est noté $\langle(\_x...)...\rangle$.

Considérons le motif séquentiel $\langle(a)\rangle$ extrait dès la première passe. La base projetée de $\langle(a)\rangle$ contient 9 postfixes présentés dans le tableau 3.3. On extrait ensuite de cette base projetée tous les items i tels que $\langle(ai)\rangle$ ou $\langle(a)(i)\rangle$ soit fréquent. Le motif séquentiel $\langle(a)(b)\rangle$ est extrait. Le processus peut ainsi continuer en retournant $\langle(a)(b)\rangle$, puis en l'utilisant comme un nouveau préfixe.

Dans l'adaptation que nous faisons de *PrefixSpan*, la première idée est la suivante : bien que l'ensemble de la base soit fouillée, la fréquence de chaque motif est calculée pour chaque contexte minimal. Pour ce faire, la base contextuelle de séquences est transformée de manière à ce que chaque séquence soit associée à un contexte minimal. Le résultat de cette transformation pour notre base d'exemple est présenté dans le tableau 3.4.

Exemple 36 : L'extraction des motifs séquentiels requis se produit de manière similaire à l'algorithme original *PrefixSpan*. Dans un premier temps, la base est parcourue et la fréquence de chaque item est calculée pour chaque contexte minimal. Le résultat de ce parcours est présenté dans le tableau 3.5 où la fréquence est affichée en gras lorsque la séquence correspon-

Postfixes
$\langle(\_d)(b)\rangle$
$\langle(\_b)(b)\rangle$
$\langle(a)(b)\rangle$
$\langle(bc)\rangle$
$\langle(\_b)(bcd)\rangle$
$\langle(b)(a)\rangle$
$\langle(bc)\rangle$
$\langle(\_b)(a)(bd)\rangle$
$\langle(e)\rangle$

TABLE 3.3 – La base projetée pour le préfixe $\langle(a)\rangle$.

dante est fréquente. On s'intéresse aux motifs fréquents dans au moins un contexte minimal : $\langle(a\rangle), \langle(b\rangle), \langle(d\rangle), \langle(e\rangle)$.

Chacun de ces motifs séquentiels m est alors retourné avec l'ensemble $\mathcal{F}_m$ des contextes minimaux où il est fréquent. Par exemple, le motif $\langle(a)\rangle$ est retourné sous la forme du couple $(\langle(a)\rangle, \mathcal{F}_{\langle(a)\rangle})$, où $\mathcal{F}_{\langle(a)\rangle} = \{[j,e],[j,h],[a,e],[a,h]\}$.

La suite du processus suit les principes de *PrefixSpan*. Chaque motif est utilisé comme préfixe et de nouveaux motifs sont trouvés dans les nouvelles bases projetées construites. Cependant, la construction des bases projetées est modifiée afin d'optimiser le processus général.

Considérons par exemple le préfixe $\langle(d)\rangle$. La base projetée correspondante suivant la définition de *PrefixSpan* est présentée dans le tableau 3.5(a). Elle contient tous les postfixes de la base $\mathcal{CB}$. Il est cependant inutile de conserver les postfixes associés aux contextes où $\langle(d)\rangle$ n'est pas fréquent. En effet, aucun super-motif fréquent ne sera trouvé dans ces contextes (d'après la propriété d'anti-monotonie). Par conséquent, nous proposons de ne construire la base projetée que sur les contextes minimaux où un motif m est fréquent (i.e., $\mathcal{F}_m$).

Le tableau 3.5(b) décrit la base projetée ainsi construite pour le préfixe $\langle(d)\rangle$.

Les modifications apportées à l'algorithme *PrefixSpan* présentée ci-dessus permettent d'obtenir l'ensemble des couples $(m, \mathcal{F}_m)$, où m est un motif séquentiel et $\mathcal{F}_m$ est l'ensemble non-vide des contextes minimaux où m est fréquent. Deux adaptations ont été proposées. La première concerne le calcul de la fréquence de chaque motif dans la base d'origine. Alors que l'algorithme original la calcule sur l'ensemble de la base, nous la calculons de manière séparée sur chacun des contextes minimaux divisant la base. Cette modification a l'avantage d'être facilement applicable sur nombre des algorithmes d'extraction de motifs fréquents autres que `PrefixSpan` utilisé ici, qu'ils soient basés sur le paradigme *générer-élaguer* ou non.

La deuxième modification est en revanche une optimisation spécifique à *PrefixSpan* puisqu'elle se base sur la réduction des bases projetées aux seuls contextes minimaux de $\mathcal{F}_m$.

La modification de l'algorithme d'extraction des motifs séquentiels ne permet cependant pas d'obtenir les motifs fréquents contextuels souhaités. Nous nous intéressons donc, dans la sous-section suivante, à la génération de tels motifs à partir des motifs obtenus dans la première

id	Contexte minimal	Séquence
s_1	$[j, e]$	$\langle(ad)(b)\rangle$
s_2	$[j, e]$	$\langle(ab)(b)\rangle$
s_3	$[j, e]$	$\langle(a)(a)(b)\rangle$
s_4	$[j, e]$	$\langle(c)(a)(bc)\rangle$
s_5	$[j, e]$	$\langle(d)(ab)(bcd)\rangle$
s_6	$[j, h]$	$\langle(b)(a)\rangle$
s_7	$[j, h]$	$\langle(a)(b)(a)\rangle$
s_8	$[j, h]$	$\langle(d)(a)(bc)\rangle$
s_9	$[a, e]$	$\langle(ab)(a)(bd)\rangle$
s_{10}	$[a, e]$	$\langle(bcd)\rangle$
s_{11}	$[a, e]$	$\langle(bd)(a)\rangle$
s_{12}	$[a, h]$	$\langle(e)(bcd)(a)\rangle$
s_{13}	$[a, h]$	$\langle(bde)\rangle$
s_{14}	$[a, h]$	$\langle(b)(a)(e)\rangle$

TABLE 3.4 – La base contextuelle de séquences transformée pour l'extraction.

Séquence	$[j, e]$	$[j, h]$	$[a, e]$	$[a, h]$
$\langle(a)\rangle$	**5/5**	**3/3**	**2/3**	**2/3**
$\langle(b)\rangle$	**5/5**	**3/3**	**3/3**	**3/3**
$\langle(c)\rangle$	2/5	1/3	1/3	1/3
$\langle(d)\rangle$	2/5	1/3	**3/3**	**2/3**
$\langle(e)\rangle$	0/3	0/3	0/3	**3/3**

TABLE 3.5 – Recherche des items fréquents dans un contexte minimal au moins.

étape.

3.4.2 Génération des motifs fréquents contextuels

La sous-section précédente montre les principes généraux qui permettent d'obtenir chaque motif m associé à l'ensemble $\mathcal{F}_m$ de contextes minimaux où il est fréquent. En nous appuyant sur cet ensemble, nous déduisons les motifs fréquents contextuels maximaux par le contexte générés par m. Ceci est réalisé par $Couverture(\mathcal{F}, \mathcal{H})$, décrit dans l'algorithme 2. Cet algorithme repose sur un parcours ascendant de la hiérarchie de contextes (i.e., des feuilles vers la racine), dans le but de collecter les contextes maximaux dont la décomposition est un sous-ensemble de $\mathcal{F}$ (*Cf.* section 3.2).

3.4.3 Algorithme général

Les deux précédentes sous-sections répondent aux deux interrogations suivantes :

(a)

Contexte	Postfixes
$[j, e]$	$\langle(b)\rangle$
$[j, e]$	$\langle(ab)(bcd)\rangle$
$[j, h]$	$\langle(a)(bc)\rangle$
$[a, e]$	$\langle(a)\rangle$
$[a, h]$	$\langle(a)\rangle$
$[a, h]$	$\langle(\_e)\rangle$

(b)

Contexte	Postfixes
$[a, e]$	$\langle(a)\rangle$
$[a, h]$	$\langle(a)\rangle$
$[a, h]$	$\langle(\_e)\rangle$

TABLE 3.6 – Bases projetées de $\langle(d)\rangle$ selon *PrefixSpan* (a) ou restreinte à $\mathcal{F}_{\langle(d)\rangle}$ (b)

Algorithm 2 Couverture($\mathcal{F}$, $\mathcal{H}$)

ENTRÉES: Un ensemble de contextes minimaux $\mathcal{F}$, une hiérarchie de contextes $\mathcal{H}$.
Soit $C = \emptyset$
Soit $\mathcal{L}$ l'ensemble des feuilles de $\mathcal{H}$
pour tout $l \in \mathcal{L}$ **faire**
$C = C \cup auxiliaireCouverture(l, \mathcal{F}, \mathcal{H})$
fin pour
retourne C la couverture de $\mathcal{F}$ dans $\mathcal{H}$

Routine $auxiliaireCouverture(c, \mathcal{F}, \mathcal{H})$
ENTRÉES: Un contexte c, un ensemble de contextes minimaux $\mathcal{F}$, une hiérarchie de contextes $\mathcal{H}$.
Soit $C = \emptyset$
si $decomp(c) \subseteq \mathcal{F}$ **alors**
pour tout p parent de c dans $\mathcal{H}$ **faire**
$C = C \cup auxiliaireCouverture(p, \mathcal{F}, \mathcal{H})$
fin pour
si $C = \emptyset$ **alors**
$C = \{c\}$
finsi
finsi
retourne C

1. *Comment extraire les motifs fréquents dans la base contextuelle d'objets ?*
2. *Comment, à partir des motifs fréquents extraits, générer les motifs fréquents contextuels ?*

Associées, les deux réponses fournies permettent aisément d'obtenir une méthode générale d'extraction des motifs contextuels maximaux par le contexte décrit dans l'algorithme 3. Celui-ci est basé sur les étapes suivantes :

1. Dans un premier temps, l'ensemble des couples $(m, \mathcal{F}_m)$ est extrait dans la base $\mathcal{CB}$. Lorsque les motifs recherchés sont des motifs séquentiels, cette étape peut être effectuée au travers des modifications apportées à l'algorithme *PrefixSpan* dans la sous-section 3.4.1. Pour d'autres types de motifs, les mêmes principes peuvent cependant s'appliquer.
2. Puis, pour chacun des motifs extraits m, l'algorithme 2 fournit l'ensemble $\mathcal{C}$ de tous les contextes maximaux où m est général. Il ne reste alors plus qu'à générer les couples (c, m) où c est un tel contexte.

Algorithm 3 *CFPM sans optimisation*

Entrées: une base contextuelle de séquences $\mathcal{CB}$, un seuil minimum de fréquence σ, une hiérarchie de contextes $\mathcal{H}$.

`/* Extraction des motifs fréquents requis */`
Extraire tous les couples $(m, \mathcal{F}_m)$ tels que $\mathcal{F}_m \neq \emptyset$

`/* Indexation de chaque motif` m `en fonction de` $\mathcal{F}_m$ `*/`
pour tout $(m, \mathcal{F}_m)$ **faire**
  $\mathcal{C} \leftarrow Couverture(\mathcal{F}_m, \mathcal{H})$
  pour tout $c \in \mathcal{C}$ **faire**
    générer le motif fréquent contextuel (c, m)
  fin pour
fin pour

Cet algorithme possède néanmoins un inconvénient majeur : la méthode $Couverture(\mathcal{F}_m, \mathcal{H})$ est exécutée pour chaque motif extrait. Pourtant, nous pouvons facilement remarquer que deux motifs m et m' fréquents dans les mêmes contextes minimaux (i.e., $\mathcal{F}_m = \mathcal{F}_{m'}$) sont généraux dans les mêmes contextes et généreront les mêmes motifs fréquents contextuels. Afin de répondre à ce problème, nous proposons l'algorithme 4. Dans celui-ci, un motif extrait m est d'abord stocké dans une table de hachage T dont les clés sont les différents ensembles $\mathcal{F}_m$ rencontrés au cours du processus. Aussi, à la fin du processus, $T[\mathcal{F}]$ renvoie l'ensemble de tous les motifs fréquents m fréquents dans les éléments de $\mathcal{F}$ (i.e., tels que $\mathcal{F}_m = \mathcal{F}$). Lorsque tous les motifs ont été extraits et stockés dans T, les motifs contextuels sont générés de manière groupée : pour chaque clé $\mathcal{F}$ de la table de hachage T, l'opération $Couverture(\mathcal{F}_m, \mathcal{H})$ est exécutée. Puis tous les motifs fréquents contextuels sont générés à partir du résultat obtenu. Cette approche a un avantage majeur : le nombre d'exécutions de l'opération $Couverture(\mathcal{F}_m, \mathcal{H})$ ne dépend plus du nombre de motifs extraits (potentiellement très grand) mais seulement du nombre de combinaisons de $\mathcal{F}$ rencontrées.

3.5 Expérimentations

Les approches proposées dans ce manuscrit ont été évaluées sur trois jeux de données différents. Dans cette section, nous décrivons ces jeux de données ainsi que leurs spécificités puis nous présentons les résultats obtenus dans le cadre de l'extraction de motifs fréquents contextuels. Toutes les expérimentations présentées au sein de ce mémoire ont été réalisées sur un système équipé de 16GB de mémoire centrale et d'un processeur cadencé à 3GHz. Les méthodes sont implémentées en C++.

Algorithm 4 *CFPM*

Entrées: une base contextuelle de séquences $\mathcal{CB}$, un seuil minimum de fréquence σ, une hiérarchie de contextes $\mathcal{H}$.

```
/* Extraction des motifs fréquents requis */
```
Extraire tous les couples $(m, \mathcal{F}_m)$ tels que $\mathcal{F}_m \neq \emptyset$

```
/* Indexation de chaque motif m en fonction de F_m */
```
pour tout $(m, \mathcal{F}_m)$ **faire**
  insérer m dans $T[\mathcal{F}_m]$
fin pour

```
/* Génération groupée des motifs fréquents contextuels */
```
pour tout $\mathcal{F} \in clés(T)$ **faire**
  $\mathcal{C} \leftarrow Couverture(\mathcal{F}, \mathcal{H})$
  pour tout $c \in \mathcal{C}$ **faire**
    pour tout $m \in T[\mathcal{F}]$ **faire**
      générer le motif fréquent contextuel (c, m)
    fin pour
  fin pour
fin pour

3.5.1 Description des données

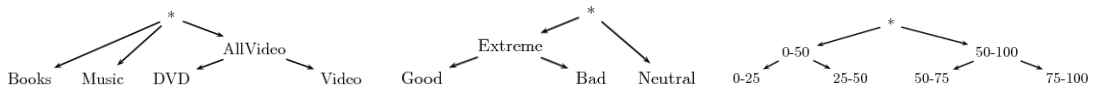
Commentaires Amazon

Notre premier jeu de données est constitué d'environ 100000 commentaires d'utilisateurs sur des produits du site `amazon.com`. Notre objectif est d'étudier le vocabulaire utilisé en fonction du type de commentaire. Ce jeu de données est une partie de celui utilisé dans [JL08]. Les commentaires, en anglais, ont été lemmatisés [2] et grammaticalement filtrés à l'aide de l'outil *tree tagger* [Sch94] afin de supprimer les termes jugés inintéressants. Nous avons conservé les verbes (mis à part les verbes modaux et le verbe "être"), les noms, les adjectifs et les adverbes. La base de séquences a été construite suivant les principes suivants : chaque commentaire est une séquence, chaque phrase est un itemset (i.e., l'ordre des mots dans une phrase n'est pas considéré) et chaque mot lemmatisé est un item.

Nous recherchons des motifs séquentiels de la forme $\langle(eat\ mushroom)(hospital)\rangle$, signifiant que fréquemment, une phrase contient les mots *eat* et *mushroom* et une des phrases suivantes contient *hospital*.

Chaque commentaire est associé aux dimensions contextuelles suivantes :
- le type de *Produit* (*Book*, *DVD*, *Music* ou *Video*).
- la *Note* (à l'origine, une valeur numérique r entre 0 et 5). Pour ces expérimentations, r a

2. i.e., les différentes formes d'un mot ont été regroupées sous la forme d'un item unique. Par exemple, les différentes formes du verbe *être* (est, sont, était, été, etc.) sont toutes retournées en « *être* ».



$\mathcal{H}(Produit)$ $\mathcal{H}(Note)$ $\mathcal{H}(Retour)$

FIGURE 3.3 – Hiérarchies sur les dimensions contextuelles pour les données `Amazon`.

été traduit en valeurs qualitatives : *Bad* (si $0 \leq r < 2$), *Neutral* (si $2 \leq r \leq 3$) et *Good* (si $3 < r \leq 5$).
- la *Retour* des utilisateurs : pourcentage de retours positifs sur un commentaire[3], i.e., *0-25%*, *25-50%*, *50-75%* or *75-100%*.

Nous définissons les hiérarchies sur les dimensions contextuelles comme décrites dans la figure 3.3. Le nombre de contextes est $|dom'(Produit)| \times |dom'(Note)| \times |dom'(Retour)| = 6 \times 5 \times 7 = 210$. Le nombre de contextes minimaux est $|dom(Produit)| \times |dom(Note)| \times |dom(Retour)| = 4 \times 3 \times 4 = 48$.

Notons que le domaine des dimensions contextuelles est enrichi avec de nouvelles valeurs. Par exemple, la hiérarchie $\mathcal{H}(note)$ contient une valeur $Extreme$, qui permettra par la suite d'obtenir les motifs spécifiques aux opinions extrêmes, qu'elles soient positives ou négatives.

Puces à ADN

Le deuxième jeu de données exploité dans nos travaux est constitué de puces à ADN obtenues afin d'étudier les différents cancers du sein. Notre objectif est d'analyser les variations existant entre différents grades (sévérités) de cancers. Les puces à ADN sont des outils puissants permettant de dresser un véritable portrait génétique d'un échantillon biologique (ici des échantillons de tumeurs) en comparant l'expression de milliers de gènes dans différents tissus, cellules ou conditions.

Les données que nous avons exploitées proviennent de plusieurs enregistrements issus du NCBI[4] (*National Center for Biotechnology Information*). Un tri des enregistrements a été nécessaire afin de sélectionner un nombre suffisant de puces adéquates avec notre approche (i.e., contenant toutes les informations contextuelles souhaitées). Le jeu de données utilisé pour nos expérimentations est constitué de 649 puces à ADN.

Les puces à ADN fournissent l'expression de milliers de gènes pour chacun des patients. Or, la plupart de ces gènes ne sont pas connus pour avoir une implication dans le cancer du sein. Nous nous sommes donc appuyés sur les 128 gènes identifiés par [SWL+06] pour leur implication dans cette maladie.

Dans cette application, l'originalité vient du fait que nous exploitons les puces à ADN sous la forme de séquences de gènes ordonnés selon leur niveau d'expression. Il n'y a donc pas la notion usuelle de temps associée aux séquences. Le tableau 3.4 montre comment les puces à ADN sont transformées en séquences. Le tableau 3.4(a) présente trois puces $P1$, $P2$ et $P3$ (les lignes du tableau) et les gènes $G_1, G_2, ..., G_5$ (les colonnes du tableau). Une puce à ADN contient

3. Sur `amazon.com`, chaque utilisateur peut déclarer s'il a trouvé un commentaire utile ou non.
4. `http://www.ncbi.nlm.nih.gov`

Puce	G_1	G_2	G_3	G_4	G_5
$P1$	7.3	6.6	6.6	9.5	8.1
$P2$	5.6	7.4	5.6	5.3	7.9
$P3$	5.7	5.2	8.7	6.8	6.2

(a) Ensemble des puces à ADN.

Puce	Séquence
$P1$	$\langle(G_2G_3)(G_1)(G_5)(G_4)\rangle$
$P2$	$\langle(G_4)(G_1G_3)(G_2)(G_5)\rangle$
$P3$	$\langle(G_2)(G_1)(G_5)(G_4)(G_3)\rangle$

(b) Base de séquences correspondante.

FIGURE 3.4 – Extraction de motifs contextuels avec ou sans restriction sur la hiérarchie de contextes.

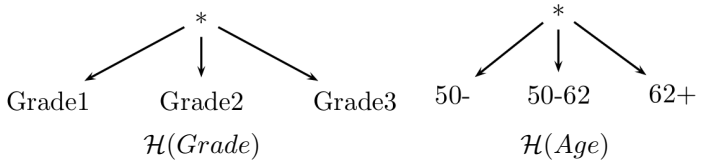


FIGURE 3.5 – Hiérarchies sur les dimensions contextuelles pour les puces à ADN.

l'expression de chaque gène (les cellules du tableau). Le tableau 3.4(b) présente les séquences correspondant à chaque puce. Les gènes sont des items, ordonnés dans la séquence en fonction de leur expression croissante. Ainsi, la puce $P1$ est traduite sous la forme d'une séquence signifiant : « *Les gènes G_2 et G_3 ont une expression identique, inférieure à celle de G_1, elle-même inférieure à celle de G_5, elle-même inférieure à celle de G_4* ».

L'extraction de motifs séquentiels dans une telle base de séquences, définie et discutée en détails dans [Sal10], permet d'étudier les puces à ADN sous un œil nouveau en considérant l'ordre des expressions de gènes pour une tumeur au lieu d'analyser directement les différentes valeurs d'expression. Un motif séquentiel extrait dans ces données est de la forme $\langle(G_1)(G_5)\rangle$, signifiant que fréquemment, le gène G_1 a une expression inférieure au gène G_5.

Chaque puce à ADN est associée aux dimensions contextuelles suivantes :

- **Grade** : chaque tumeur peut être associée à un grade de malignité. Ces grades dépendent de l'évaluation de trois critères que sont la différenciation, le degré d'anisocaryose, ainsi que le nombre de mitoses. Valeurs : *grade 1, 2 ou 3.*
- **Age** : l'âge du patient lors du diagnostic. Valeurs : *50- (moins de 50 ans), 50-62 (entre 50 et 62 ans), 62+ (plus de 62 ans).*

Les hiérarchies définies sur les dimensions contextuelles sont présentées dans la figure 3.5. Le nombre de contextes est $|dom'(Grade)| \times |dom'(Age)| = 4 \times 4 = 16$. Le nombre de contextes minimaux est $|dom(Grade)| \times |dom(Age)| = 3 \times 3 = 9$.

Consommation énergétique

Le troisième jeu de données que nous utiliserons dans ce manuscrit est issu du site `data.gov.uk`. Il s'agit d'une collection des données de consommation énergétique du siège du département anglais de la santé *Richmond House*.

La consommation d'électricité et la consommation de carburant ont été collectées toutes les 30 minutes depuis le 1er septembre 2008 jusqu'au 9 juin 2011, totalisant ainsi environ 50000

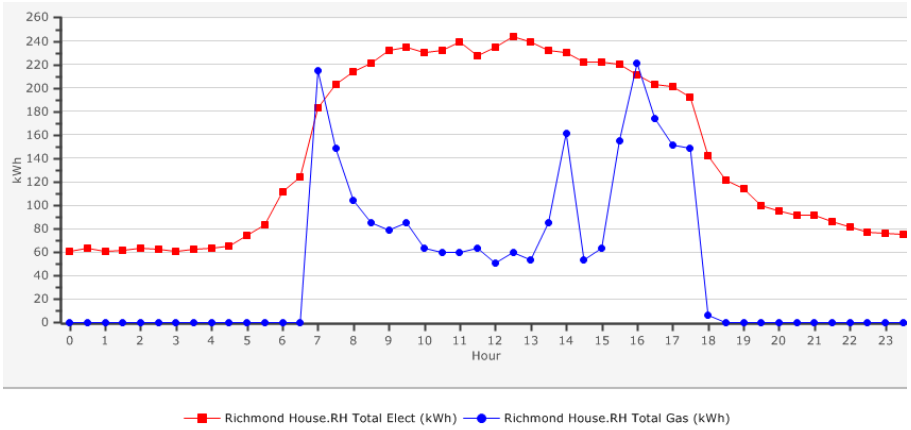


FIGURE 3.6 – Consommation d'électricité et de gaz (en kWh) le mardi 3 mai 2011.

relevés. À titre d'exemple, la figure 3.6 décrit la consommation énergétique du mardi 3 mai 2011.

Afin d'analyser ces données par le biais de l'extraction de motifs, les valeurs numériques ont été discrétisées et les données transformées en une séquence étendue, telle que chaque itemset correspond à un relevé et les estampilles temporelles associées correspondent au numéro de l'itemset dans la séquence. Un extrait de cette séquence étendue, correspondant au 3 mai 2011 entre 6h et 7h30, est proposé ci-dessous :

$$\langle ...(E_{105\text{–}117}G_0)^k(E_{117\text{–}138}G_0)^{k+1}(E_{138\text{–}187}G_{208+})^{k+2}(E_{187\text{–}234}G_{124\text{–}208})^{k+3}...\rangle.$$

La discrétisation des valeurs numériques a été réalisée sur le principe d'intervalles de fréquence constante[5]. La consommation électrique a 8 valeurs, tandis que la consommation de gaz a 7 valeurs.

Chaque itemset inter-transactionnel construit sur cette séquence étendue est associé à différentes informations contextuelles :

- Le **jour** de la semaine. Valeurs : *lundi, mardi, ..., samedi, dimanche*. De plus, le domaine enrichi de cette dimension est associé à de nouvelles valeurs : *jour de semaine, jour de week-end,* $*$.
- Le **mois**. Valeurs : *janvier, février, ..., décembre*. Le domaine enrichi de cette dimension est composé de valeurs plus générales : *hiver, printemps, été, automne,* $*$.
- L'**heure**. Valeurs : *0h-6h, 6h-10h, 10h-14h, 14h-20h, 20h-0h*. Les valeurs plus générales du domaine enrichi sont : *20h-6h, 6h-14h,* $*$.

Le nombre de contextes de la hiérarchie est $dom'(Jour) \times dom'(Mois) \times dom'(Heure) = 10 \times 17 \times 8 = 1360$. Le nombre de contextes minimaux est $dom(Jour) \times dom(Mois) \times dom(Heure) = 7 \times 12 \times 5 = 420$. Cette hiérarchie est donc sensiblement plus grande que les deux précédentes.

Notons que l'extraction d'itemsets inter-transactionnels dans la suite est réalisée par le biais de la modification de `PrefixSpan` que nous avons décrite dans la section 3.4. En effet, l'extraction

5. Chaque intervalle correspond à un même nombre d'exemples dans les données initiales. Ainsi, les valeurs rares sont associées à des intervalles plus larges que les valeurs très fréquentes.

d'itemsets inter-transactionnels peut être vue comme un cas particulier de l'extraction de motifs séquentiels où les séquences ne contiennent qu'un seul itemset. De plus, une légère modification a été implémentée de manière à n'extraire que les motifs qui contiennent au moins un item étendu de la forme $(i, 0)$, comme le requiert la définition des itemsets inter-transactionnels.

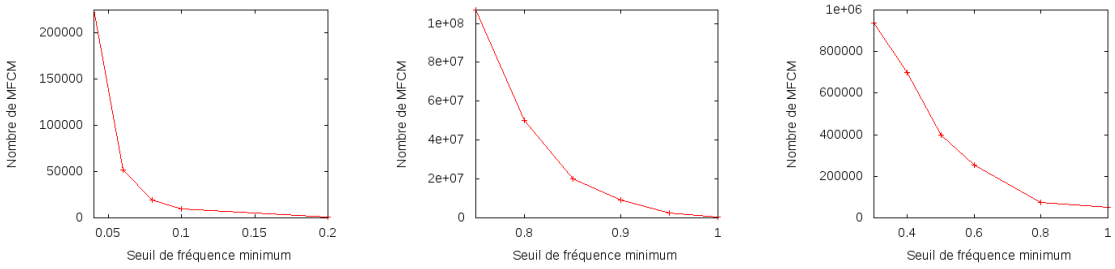
3.5.2 Résultats expérimentaux

Les expérimentations que nous présentons ont deux objectifs principaux :
- Tout d'abord, nous évaluons la méthode et les algorithmes proposés pour l'extraction de motifs fréquents contextuels. Nous étudions donc l'efficacité de notre approche et des différentes optimisations que nous avons proposées dans l'algorithme `CFPM`.
- Dans un second temps, nous nous intéressons aux motifs extraits et à leur pertinence par rapport aux motifs fréquents traditionnels.

3.5.2.1 Évaluation des algorithmes

Nous nous intéressons dans un premier temps aux algorithmes développés dans la section 3.4. Les figures 3.7 et 3.8 présentent le comportement général (nombre de motifs extraits et temps d'exécution global) de l'algorithme `CFPM` pour l'extraction des motifs fréquents contextuels maximaux (MFCM) dans chacun des jeux de données.

Plus précisément, l'algorithme `CFPM` est constitué de deux parties : l'extraction des motifs fréquents dans les contextes minimaux de la hiérarchie par le biais d'une adaptation de l'algorithme `PrefixSpan` et la génération des motifs fréquents contextuels à partir des motifs extraits.

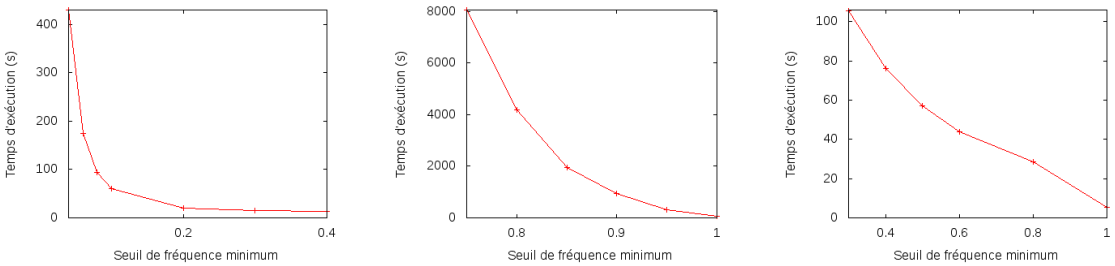


(a) *Amazon* (b) *Puces à ADN* (c) *Consommation énergétique*

FIGURE 3.7 – Nombre de motifs fréquents contextuels maximaux (MFCM) extraits en fonction du seuil de fréquence minimum σ pour chaque jeu de données.

Extraction des motifs fréquents dans les contextes minimaux D'abord, nous avons proposé un algorithme pour extraire, dans une base contextuelle de séquences $\mathcal{B}$, tous les couples $(m, \mathcal{F}_m)$ où m est un motif fréquent et $\mathcal{F}_m$ est l'ensemble non-vide des contextes minimaux où m est fréquent.

Nous avons notamment exploité le paradigme des bases projetées de `PrefixSpan` pour optimiser le processus, en supprimant des bases projetées toutes les séquences inutiles, i.e., appartenant



(a) *Amazon* (b) *Puces à ADN* (c) *Consommation énergétique*

FIGURE 3.8 – Temps d'exécution (en s) pour l'extraction des motifs fréquents contextuels maximaux (MFCM) en fonction du seuil de fréquence minimum σ pour chaque jeu de données.

à un contexte minimal où un préfixe n'est pas fréquent. Le tableau 3.7 montre pour chacun des jeux de données la réduction des temps d'extraction des MFCM due cette optimisation. En effet, elle permet de supprimer de nombreuses séquences des bases projetées et réduit donc le temps global de traitement de celles-ci.

Données	σ	sans	avec
amazon	0.06	740.69	174.08
puces à ADN	0.9	3357.42	937.09
consommation énergétique	0.3	246.64	105.52

TABLE 3.7 – Temps d'extraction (en s) des MFCM sans et avec l'optimisation des bases projetées en fonction du jeu de données et du seuil de fréquence minimum (σ).

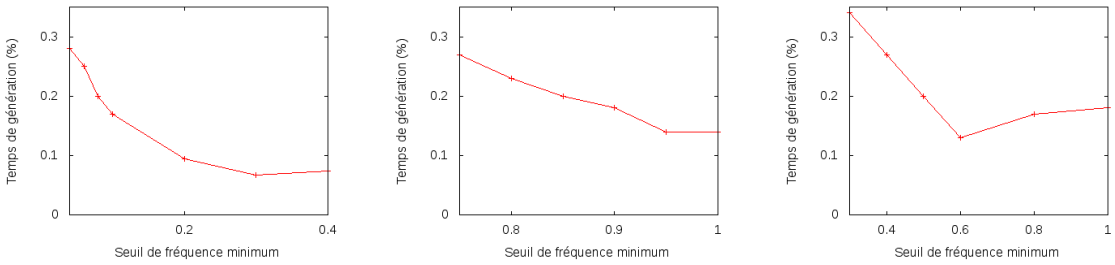
Génération des motifs fréquents contextuels Notre approche d'extraction des motifs fréquents contextuels repose sur la génération de tels motifs d'après les motifs fréquents extraits dans les contextes minimaux. Une première approche proposée au travers de l'algorithme 3 génère les motifs contextuels en calculant la couverture de $\mathcal{F}_m$ (méthode $Couverture(\mathcal{F}_m, \mathcal{H})$) séquentiellement pour chaque motif fréquent m extrait. Afin d'optimiser le nombre d'appels à la méthode $Couverture(\mathcal{F}_m, \mathcal{H})$, nous avons proposé dans l'algorithme 4 de regrouper la génération des motifs fréquents contextuels. Le tableau 3.8 présente les résultats obtenus avec chacun de ces algorithmes et montre que cette optimisation réduit considérablement le temps dédié à la génération des motifs fréquents contextuels.

Étudions à présent le temps dédié à la génération des motifs fréquents contextuels en comparaison avec le temps global de l'algorithme `CFPM` (i.e., le temps d'extraction des motifs fréquents dans les contextes minimaux ajouté au temps de génération des motifs fréquents contextuels). La figure 3.9 présente la proportion du temps de génération des motifs fréquents contextuels dans le temps total d'exécution en fonction du seuil de fréquence minimum. Nous notons que le temps dédié à la génération des motifs fréquents contextuels est négligeable par rapport au temps total pris par le processus. Une très large proportion du temps nécessaire à l'exécution de

Données	σ	sans	avec
amazon	0.06	4.77	0.44
puces à ADN	0.9	143.8	1.72
consommation énergétique	0.3	0.62	0.01

TABLE 3.8 – Temps de génération (en s) des MFCM sans et avec l'optimisation des appels à la méthode *Couverture* en fonction du jeu de données et du seuil de fréquence minimum (σ).

l'algorithme `CFPM` est donc liée à l'extraction des motifs fréquents dans les contextes minimaux. Par exemple, pour un seuil de fréquence minimum $\sigma = 0.04$, cette partie de l'algorithme prend 0.28% du temps total.



(a) *Amazon* (b) *Puces à ADN* (c) *Consommation énergétique*

FIGURE 3.9 – Étude de la proportion (en %) du temps dédié à la génération des MFCM dans le processus global `CFPM` en fonction du seuil de fréquence minimum.

Comparaison avec l'approche naïve Finalement, nous comparons l'algorithme `CFPM` avec l'approche naïve décrite dans la section 3.3. Cette dernière nécessite au préalable d'extraire les motifs fréquents dans chacun des contextes de la hiérarchie, au lieu de les extraire uniquement dans les contextes minimaux comme le propose `CFPM`. Le tableau 3.9 compare le temps d'exécution nécessaire à `CFPM` avec le temps nécessaire à l'extraction des motifs fréquents dans chaque contexte de la hiérarchie (nous n'avons pas considéré le temps supplémentaire de génération des motifs fréquents contextuels pour l'approche naïve). Sans surprise, `CFPM` s'exécute bien plus rapidement que l'approche naïve. En effet, nous avons montré que le temps d'exécution de `CFPM` est en grande partie dû à l'extraction des motifs fréquents dans les contextes minimaux. Naturellement, lorsque tous les contextes de $\mathcal{H}$ sont considérés, le temps nécessaire explose. Pour le jeu de données *Amazon* par exemple, il s'agit de fouiller 210 contextes dans l'approche naïve contre 48 pour l'algorithme `CFPM`.

Ainsi, le gain obtenu par l'algorithme `CFPM` en comparaison avec l'approche naïve provient en grande partie de la définition même d'un motif général dans un contexte et des propriétés qu'elle induit. En effet, nous avons montré par le biais du lemme 1 que cette définition pouvait être réécrite sans tenir compte des nombreux contextes non-minimaux de la hiérarchie.

À ce sujet, soulignons un aspect intéressant de `CFPM`. Nous avons vu tout au long de ce

chapitre que le domaine des dimensions contextuelles pouvait être enrichi de valeurs supplémentaires, telles que la valeur *Extreme* sur la dimension *Note* du jeu de données *Amazon*. Cette valeur appartient à $dom'(Note)$ (i.e., le domaine enrichi) mais pas à $dom(Note)$ (i.e., le domaine simple). Par conséquent, l'ajout de telles valeurs ne modifie pas le nombre de contextes minimaux dans la hiérarchie. Or, le temps d'exécution de `CFPM` étant très largement dû à l'extraction des motifs fréquents dans les contextes minimaux, nous en concluons que l'enrichissement du domaine des dimensions contextuelles a un effet négligeable sur le temps nécessaire à l'extraction des motifs fréquents contextuels.

Données	σ	Extraction naïve	CFPM
amazon	0.1	756.4	61.21
puces à ADN	0.9	2329.61	937.09
consommation énergétique	1	9340.48	5.5

TABLE 3.9 – Temps d'extraction (en s) des motifs fréquents dans les contextes minimaux pour l'approche naïve et `CFPM` en fonction du jeu de données et du seuil de fréquence minimum (σ).

L'ensemble des expérimentations que nous avons menées dans cette sous-section montrent l'efficacité de l'algorithme `CFPM` et des optimisations qu'il intègre. Cependant, elles ne garantissent pas l'intérêt des motifs extraits par rapport aux motifs fréquents traditionnels. Nous présentons par conséquent dans la suite l'analyse de ces résultats.

3.5.2.2 Évaluation de la pertinence des motifs fréquents contextuels

Intéressons nous à présent au nombre de motifs extraits. Le tableau 3.10 présente pour le jeu de données *Amazon* le nombre de motifs extraits dans les différents contextes. Nous ne pouvons pas décrire les résultats pour chacun des 210 contextes de la hiérarchie. Par conséquent, nous proposons une sélection de contextes plus ou moins généraux :

- $[*, *, *]$ est le contexte le plus général de la hiérarchie. Il correspond à tous les commentaires de la base contextuelle.
- $[Books, *, *]$ est le contexte correspondant à tous les commentaires décrivant un livre.
- $[Books, bad, *]$ est un contexte plus spécifique que $[Books, *, *]$. Il correspond aux commentaires décrivant une mauvaise opinion d'un livre.
- $[Books, bad, 75-100]$ est un contexte minimal dans la hiérarchie. Il correspond aux commentaires décrivant une mauvaise opinion d'un livre, qui ont été considérés comme utiles par plus de 75% des votants.

Nous observons que la proportion de motifs fréquents contextuels maximaux (MFCM) parmi les motifs séquentiels fréquents (MF) est basse dans les contextes non-minimaux (*Cf.* pourcentages entre parenthèses dans la colonne MFC). En effet, par définition, un motif général dans un contexte est nécessairement fréquent dans ce contexte. Par conséquent, le nombre de motifs généraux dans un contexte est inférieur ou égal au nombre de motifs fréquents. Considérons plus attentivement l'exemple du contexte $[Books, *, *]$. Seulement 19% des motifs fréquents sur l'en-

semble des commentaires associés à un livre génèrent un motif fréquent contextuel maximal. En d'autres termes, 80% des motifs fréquents extraits ne sont en réalité fréquents que pour certains types de commentaires sur les livres. Nous soutenons donc que 81% des motifs fréquents extraits ne sont pas représentatifs du contexte $[Books, *, *]$ mais d'une partie de ses sous-contextes seulement.

Contexte	MF	MFCM
$[*, *, *]$	1577	155 (9.8%)
$[Books, *, *]$	2467	469 (19%)
$[Books, bad, *]$	4523	1060 (23%)
$[Books, bad, 75 - 100]$	5830	2157 (37%)

TABLE 3.10 – Nombre de motifs fréquents (MF) et de motifs fréquents contextuels maximaux par le contexte (MFCM) pour $\sigma = 0.04$ dans le jeu de données *Amazon*.

Comme le montrent les tableaux 3.11 et 3.12, ces observations se vérifient également dans les jeux de données *Puces à ADN* et *Consommation énergétique*.

Ces résultats sont particulièrement intéressants car ils montrent que l'extraction de motifs fréquents contextuels présente deux avantages majeurs par rapport à l'extraction de motifs fréquents dans les contextes. D'une part, les contraintes liées aux motifs fréquents contextuels permettent de diminuer considérablement le nombre de motifs associés à un contexte. D'autre part, cette réduction n'est pas synonyme de perte d'information mais au contraire d'un nouvel apport de connaissances pour l'utilisateur. En effet, les motifs fréquents contextuels, au travers de la contrainte de c-généralité, garantissent une représentativité qu'une large proportion des motifs fréquents ne possède pas. Ainsi, ils fournissent des motifs qui sont, d'un point de vue contextuel, plus intéressants.

Contexte	MF	MFCM
$[*, *]$	455842	116042 (25.5%)
$[Grade1, *]$	1229012	517068 (42%)
$[Grade1, 50 - 62]$	2605662	1918226 (73.6%)

TABLE 3.11 – Nombre de motifs séquentiels fréquents (MF) et de motifs fréquents contextuels maximaux par le contexte (MFCM) pour $\sigma = 0.9$ dans le jeu de données *Puces à ADN*.

3.6 Discussion

Les motifs fréquents définis dans la littérature ne pouvant pas réellement intégrer et exploiter des informations contextuelles associées aux données fouillées, nous avons dans ce chapitre proposé une solution au problème suivant : « *Dans un cadre où les données à fouiller sont associées à des informations contextuelles, quels motifs fréquents extraire et comment ?* ». Afin de

Contexte	MF	MFCM
$[*, *, *]$	7012	266 (3.8%)
$[lundi, *, *]$	12484	799 (6.4%)
$[lundi, 14h-20h, *]$	66023	7130 (10.8%)
$[lundi, 14h-20h, juin]$	174597	21475 (12.3%)

TABLE 3.12 – Nombre de motifs séquentiels fréquents (MF) et de motifs fréquents contextuels maximaux par le contexte (MFCM) pour $\sigma = 0.3$ dans le jeu de données *Consommation énergétique*.

répondre à la question posée, nous avons dans la section 3.2 formalisé un cadre pour les données contextuelles, ainsi que la notion de motif fréquent contextuel, telle qu'un motif fréquent est associé à un contexte dans lequel il est général. Nous avons proposé dans les sections 3.3 et 3.4 une méthode ainsi que des algorithmes pour extraire ces motifs. Nous avons finalement démontré dans la section 3.5 l'efficacité de nos propositions sur trois jeux de données.

Dans la section 3.1, nous avons vu que nos travaux semblaient posséder des points communs avec les motifs séquentiels multidimensionnels [PHP+01, Pla08]. Nous revenons, dans cette discussion, sur ce point. De tels motifs correspondent à des séquences multidimensionnelles fréquentes dans l'ensemble de la base. Ces séquences sont associées à un contexte à condition que cette association s'avère fréquente. La sémantique portée par ces motifs multidimensionnels est donc différente de celle fournie par les motifs fréquents contextuels. En effet, alors que les motifs multidimensionnels utilisent les informations additionnelles pour caractériser l'ensemble de la base, nous les utilisons pour caractériser les contextes eux-mêmes.

Revenons à présent sur la généralité de l'approche par rapport à d'autres types de motifs. Le grand nombre de contextes à considérer pour extraire l'ensemble des motifs fréquents contextuels rend le problème difficile. Néanmoins, nous avons démontré certaines propriétés théoriques intéressantes qui ont permis la proposition d'un algorithme efficace d'extraction de motifs fréquents contextuels. Au cours de ce chapitre nous avons, par souci de simplification, illustré les résultats en utilisant les motifs séquentiels. Toutefois, comme nous l'avons montré dans les expérimentations, l'approche est également adaptée aux itemsets inter-transactionnels (*Cf.* jeu de données *Consommation énergétique*). Dans le chapitre précédent, nous avions proposé un cadre général d'étude qui nous a permis d'abstraire l'extraction de différents types de motifs : itemsets fréquents, motifs séquentiels, etc. En analysant spécifiquement la notion de motifs fréquent présentée dans ce chapitre, nous constatons que nos extensions sont indépendantes du type de motifs considérés. Ainsi, toutes les propriétés définies pour les motifs fréquents contextuels sont utilisables pour n'importe quel type de motifs pour lesquels le problème d'extraction peut être rattaché au formalisme proposé dans le chapitre précédent.

Ce chapitre a mis en exergue le fait que la notion de fréquence était tout à fait adaptée à l'extraction de motifs contextuels. Cependant, est-ce que la fréquence est toujours la mesure la plus appropriée pour les différentes tâches d'analyse ? Nous apportons une réponse à cette question dans le chapitre suivant.

Chapitre 4

Extraction de motifs contextuels d'intérêt

Sommaire

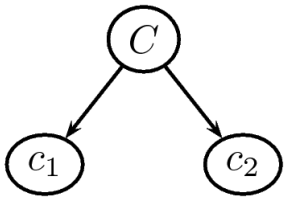


FIGURE 4.1 – Exemple de hiérarchie de contextes.

Introduction

Nous avons montré dans le chapitre 3 qu'il était possible de tenir compte des informations contextuelles associées aux données pour extraire des motifs fréquents représentatifs d'un contexte. La notion de motif fréquent contextuel ainsi définie repose exclusivement sur la fréquence d'un motif dans un contexte et ses descendants (motif c-général). Toutefois, la fréquence n'est pas toujours la mesure la plus adaptée pour découvrir les motifs d'intérêt caractéristiques ou discriminants d'un contexte par rapport au reste de la base. Par exemple, un décideur pourrait souhaiter une réponse à la question suivante : « *Quelles sont les activités qui différencient les jeunes habitants des autres habitants ?* »

Les motifs fréquents contextuels proposés dans le chapitre 3 ne fournissent pas une réponse adaptée. Pour nous en convaincre, examinons l'exemple suivant. La figure 4.1 montre une hiérarchie de contextes simple, composée de deux contextes minimaux c_1 et c_2 et d'un contexte C ancêtre des deux précédents. Considérons maintenant les deux motifs m et m', tels que leur fréquence dans les deux contextes minimaux est fournie dans le tableau 4.1. Ainsi, m a une fréquence de $0,5$ dans c_1 et de $0,45$ dans c_2.

Motif	c_1	c_2
m	$0,5$	$0,45$
m'	1	$0,5$

TABLE 4.1 – Fréquence des motifs dans les contextes minimaux.

Pour un seuil minimum de fréquence fixé à $0,5$, le motif m est uniquement représentatif du contexte c_1 (car il est fréquent dans c_1 uniquement), tandis que m' est représentatif de C (car il est fréquent à la fois dans c_1 et c_2).

Le motif m est-il pour autant plus caractéristique de c_1 que ne l'est m' ? Pour répondre, nous devons nous intéresser aux variations de fréquence de m' et m dans tous les contextes de la hiérarchie. Ce faisant, nous remarquons que m' est deux fois plus fréquent dans c_1 que dans c_2 ($Freq_{c_1}(m') = 2 \times Freq_{c_2}(m')$). En revanche, la fréquence de m' varie très peu entre c_1 et c_2 ($Freq_{c_1}(m) \approx 1,1 \times Freq_{c_2}(m)$). Selon ce critère (appelé taux d'émergence [1]), le motif m' peut être considéré comme plus caractéristique de c_1 que m.

1. Nous revenons plus en détail dans la suite sur ce critère ainsi que sur d'autres types de critères qui ont fait l'objet de nombreux travaux.

Les motifs fréquents contextuels, en s'appuyant uniquement sur la notion de fréquence d'un motif dans un contexte, ne peuvent révéler ce type d'information. De nombreux travaux ont proposé de prendre en compte des mesures plus complexes afin d'évaluer la qualité d'un motif pour caractériser une classe de données parmi d'autres. Ces mesures d'intérêt, bien que basées sur la mesure de fréquence, mettent en avant les motifs présentant des propriétés statistiques intéressantes et propres à chaque mesure.

Dans ce chapitre, nous proposons, dans la section 4.1, de généraliser les concepts liés aux motifs contextuels afin d'intégrer les mesures d'intérêt. Puis, dans la section 4.2, nous étudions le problème de l'extraction de tels motifs contextuels. Notamment, nous mettons en valeur les propriétés théoriques de certaines mesures d'intérêt dans la section 4.3 et les exploitons pour proposer un algorithme efficace d'extraction dans la section 4.4. Des expérimentations valident ces résultats dans la section 4.5 qui sont discutés dans la section 4.6.

4.1 Motifs et mesures d'intérêt

L'extraction de motifs pour caractériser les différences entre deux bases d'objets est un problème clé de la fouille de données. On le retrouve sous différents noms et différentes formes dans la littérature, tel que l'extraction de motifs discriminants [CYHH07, YCHY08], de motifs d'intérêt [JHZ10], de motifs de contraste [BP99], de motifs émergents [DL99], de sous-groupes [Wro97, GRW08] ou encore de motifs corrélés [MS00, NGDR09] [2].

Ces différentes apparitions dans la littérature, bien qu'utilisant des termes et des définitions différentes, abordent généralement un problème similaire de recherche de motifs dont la présence est corrélée avec une classe donnée. Ainsi, [NLW09] offre une synthèse des problèmes de recherche de sous-groupes, de motifs de contraste et de motifs émergents et soutient que leurs différences résident principalement dans la terminologie employée.

Dans notre cas d'étude, la recherche de tels motifs pourra par exemple mettre en valeurs les motifs caractérisant les activités des jeunes habitants, par opposition à celles des autres habitants.

Les motifs recherchés sont généralement évalués par leur significativité, mesurée suivant différents critères statistiques. La littérature fournit ainsi de nombreuses mesures basées sur la fréquence ou le support, que nous appellerons mesures d'intérêt. Le tableau 4.2 en présente quelques unes parmi les plus utilisées. Une liste plus complète pourra être trouvée dans [TKS02] ou [GH06]. Étant donné un motif m et deux bases d'objets $\mathcal{B}_1$ et $\mathcal{B}_2$, ces mesures ont la particularité de ne dépendre que de $Supp_{\mathcal{B}_1}(m)$, $Supp_{\mathcal{B}_2}(m)$, $|\mathcal{B}_1|$ et $|\mathcal{B}_2|$, i.e., une combinaison des supports de m et de la taille des bases d'objets considérées.

Dans ce chapitre, nous nous intéressons particulièrement à la notion de motif discriminant qui permet de révéler les motifs dont la fréquence contraste d'un ensemble de données à un autre. La popularité de l'extraction de motifs discriminants provient en grande partie des applications

2. Ces termes sont des traductions des différents termes anglais *discriminative pattern mining, interesting pattern mining, contrast pattern mining, emerging pattern mining, subgroup discovery, correlated pattern mining.*

Mesure	Expression
Fréquence	$\frac{Supp_{\mathcal{B}}(m)}{\lvert\mathcal{B}\rvert} = Freq_{\mathcal{B}}(m)$
Fréquence conditionnelle	$\frac{Supp_{\mathcal{B}}(m)}{\lvert\mathcal{B}\cup\mathcal{B}'\rvert}$
Confiance	$\frac{Supp_{\mathcal{B}}(m)}{Supp_{\mathcal{B}\cup\mathcal{B}'}(m)}$
Gain d'Information (GI)	$log\left(\frac{Supp_{\mathcal{B}}(m)\times\lvert\mathcal{B}\cup\mathcal{B}'\rvert}{\lvert\mathcal{B}\rvert\times Supp_{\mathcal{B}\cup\mathcal{B}'}(m)}\right) = log\left(\frac{Freq_{\mathcal{B}}(m)}{Freq_{\mathcal{B}\cup\mathcal{B}'}(m)}\right)$
Émergence (Em)	$\frac{Supp_{\mathcal{B}}(m)\times\lvert\mathcal{B}'\rvert}{Supp_{\mathcal{B}'}(m)\times\lvert\mathcal{B}\rvert} = \frac{Freq_{\mathcal{B}}(m)}{Freq_{\mathcal{B}'}(m)}$
Lift	$\frac{Supp_{\mathcal{B}}(m)\times\lvert\mathcal{B}\cup\mathcal{B}'\rvert}{Supp_{\mathcal{B}\cup\mathcal{B}'}(m)\times\lvert\mathcal{B}\rvert} = \frac{Freq_{\mathcal{B}}(m)}{Freq_{\mathcal{B}\cup\mathcal{B}'}(m)}$
Specificity	$\frac{\lvert\mathcal{B}'\rvert - Supp_{\mathcal{B}'}(m)}{\lvert\mathcal{B}\cup\mathcal{B}'\rvert}$

TABLE 4.2 – Exemples de mesures d'intérêt.

qu'ils recouvrent. Tout d'abord, leur intérêt réside dans leur aspect descriptif. Ils fournissent en effet à l'utilisateur une vue compréhensible des différences qui peuvent exister entre deux classes de données. Ces motifs peuvent également être utilisés pour répondre, dans une certaine mesure, au problème de la quantité excessive de motifs fréquents généralement extraite. En effet, se concentrer uniquement sur les motifs qui contrastent d'une base d'objets à l'autre permet de ranger les motifs extraits par ordre d'intérêt, pour ne conserver que les meilleurs (au sens d'un seuil fixé). Ces motifs ont également été largement utilisés dans des tâches de classification [CYHH07, JHZ10]. Notamment, [CYHH07] étudie l'apport théorique des motifs fréquents et discriminants dans le problème de classification.

Bien que la recherche de motifs discriminants ait avant tout abordé le domaine des itemsets [CYHY08, NGDR09], ce problème a également fait l'objet de diverses adaptations pour des motifs de structures différentes, telles que les séquences [HHS+00, DZ09], les graphes [YCHY08, CLZ+09] ou encore les arbres [ZB, HTS+08].

Dans ce chapitre, nous intégrons les mesures d'intérêt dans la découverte de motifs contextuels. Les bases d'objets considérées sont alors associées aux différents contextes possibles. Dans ce cas, la première difficulté rencontrée provient de l'organisation des données non pas en classes distinctes, mais en contextes, qui peuvent être plus ou moins généraux ou spécifiques. En effet, nous devrons prendre en compte le fait que le contexte $[j, e]$, correspondant aux jeunes habitants pendant l'été, est un sous-contexte de $[j, *]$ correspondant aux jeunes habitants. Nous obtenons alors un nombre très grand de contextes à traiter pendant l'extraction. En outre, notre objectif est toujours de conserver le principe de représentativité d'un motif contextuel. Un motif, pour être qualifié de contextuel, doit propager ses propriétés d'intérêt dans ses sous-contextes. Par

exemple, un motif jugé intéressant dans le contexte $[j, *]$ doit conserver les mêmes propriétés dans ses descendants et être ainsi également discriminant dans $[j, e]$ et $[j, h]$. Enfin, nous ne nous intéressons pas dans ces travaux à une catégorie spécifique de motifs telle que les motifs séquentiels ou les itemsets inter-transactionnels et souhaitons définir des propriétés applicables à des motifs au sens général.

4.2 Motifs contextuels et mesures d'intérêt

La définition d'un motif contextuel fréquent présentée dans le chapitre précédent est dédiée uniquement à la mesure de fréquence. Elle ne considère pas les autres contextes de la hiérarchie (car la mesure de fréquence ne dépend pas du reste de la hiérarchie) et n'est pas directement applicable pour d'autres mesures d'intérêt. Dans cette section nous généralisons cette définition pour considérer les mesures d'intérêt présentées dans la section 4.1.

Afin de faciliter la lecture de cette section, le tableau 4.3 rappelle les principales notations nécessaires introduites dans le chapitre 3. De la même manière, il est rappelé dans le tableau 4.4 et la figure 4.2 la base de séquences ainsi que la hiérarchie de contextes associées à notre cas d'étude.

Notation	Description
$\mathcal{CB}$	la base contextuelle d'objets
$\mathcal{H}$	la hiérarchie de contextes sur $\mathcal{CB}$
$c = [d_1, ..., d_n]$	c est un contexte, tel que $\forall i \in \{1, ..., n\}$, d_i est la valeur de c pour la dimension contextuelle D_i
$\mathcal{B}(c)$	la base d'objets de c, i.e., tous les objets de $\mathcal{CB}$ qui sont contenus dans c
$c > c'$	c' est un descendant de c dans $\mathcal{H}$ et c est un ancêtre de c'
$\hat{c}$	l'ensemble des composants de c, i.e., l'ensemble muni de c et de ses descendants
$decomp(c)$	la décomposition de c, i.e., l'ensemble des composants minimaux de c

TABLE 4.3 – Rappel des notations.

Nous avons observé plus tôt que les mesures d'intérêt sont définies pour mettre en valeur les motifs caractéristiques d'une base d'objets, relativement à une autre base. Dans le cas qui nous concerne, ces deux bases correspondent *(1)* à un contexte de la hiérarchie et *(2)* au reste de la hiérarchie. Afin d'étudier le reste d'un contexte dans la hiérarchie, nous définissons la notion de complément d'un contexte. Dans la suite, nous considèrons M une mesure d'intérêt, σ un seuil minimum sur M et m un motif.

id	Age	Saison	Sequence
s_1	jeune	été	$\langle(ad)(b)\rangle$
s_2	jeune	été	$\langle(ab)(b)\rangle$
s_3	jeune	été	$\langle(a)(a)(b)\rangle$
s_4	jeune	été	$\langle(c)(a)(bc)\rangle$
s_5	jeune	été	$\langle(d)(ab)(bcd)\rangle$
s_6	jeune	hiver	$\langle(b)(a)\rangle$
s_7	jeune	hiver	$\langle(a)(b)(a)\rangle$
s_8	jeune	hiver	$\langle(d)(a)(bc)\rangle$
s_9	âgé	été	$\langle(ab)(a)(bd)\rangle$
s_{10}	âgé	été	$\langle(bcd)\rangle$
s_{11}	âgé	été	$\langle(bd)(a)\rangle$
s_{12}	âgé	hiver	$\langle(e)(bcd)(a)\rangle$
s_{13}	âgé	hiver	$\langle(bde)\rangle$
s_{14}	âgé	hiver	$\langle(b)(a)(e)\rangle$

TABLE 4.4 – Une base contextuelle de séquences.

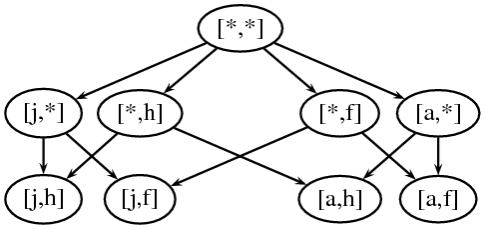


FIGURE 4.2 – La hiérarchie de contextes $\mathcal{H}$.

Définition 40 (Complément de c) **:** Soit c et c' deux contextes. Le contexte c' est dit *disjoint de* c si $\mathcal{B}(c) \cap \mathcal{B}(c') = \emptyset$. L'ensemble de tous les contextes disjoints de c, noté $\overline{c}$, est appelé le *complément de* c, i.e.,

$$\overline{c} = \{c' \in \mathcal{H} | \mathcal{B}(c) \cap \mathcal{B}(c') = \emptyset\}.$$

Exemple 37 : Considérons le contexte $[a, *]$. Comme le montre la figure 4.3, le complément de $[a, *]$ est $\overline{[a, *]} = \{[j, *], [j, e], [j, h]\}$. Le contexte $[*, h]$ n'est en revanche pas disjoint de $[a, *]$ car ils partagent un ensemble non vide d'objets $\{s_{12}, s_{13}, s_{14}\}$. En effet, $[a, *]$ et $[*, h]$ ont un descendant commun : le contexte $[a, h]$.

Le complément de c est donc composé de tous les contextes qui n'ont aucun objet en commun avec c. Nous définissons également la base du complément de c, afin de manipuler les objets contenus dans ces contextes.

Définition 41 (Base du complément de c) **:** Soit c un contexte. La *base du complément de* c,

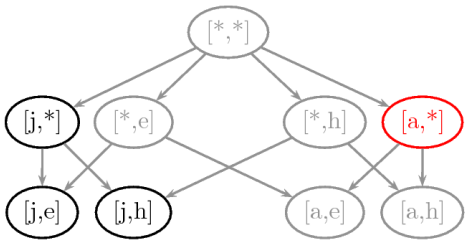


FIGURE 4.3 – Complément du contexte $[a, *]$ dans la hiérarchie de contextes.

notée $\mathcal{B}(\overline{c})$, est l'ensemble de tous les objets contenus dans les contextes disjoints de c, i.e.,

$$\mathcal{B}(\overline{c}) = \bigcup_{c' \in \overline{c}} \mathcal{B}(c').$$

Exemple 38 : La base du complément de $[a, *]$ est l'ensemble d'objets $\mathcal{B}(\overline{[a,*]}) = \{s_9, s_{10}, s_{11}, s_{12}, s_{13}, s_{14}\}$. La base du complément de $[j, h]$ est $\mathcal{B}(\overline{[j,h]}) = \{s_6, s_7, s_8\}$.

De même, nous définissons la décomposition du complément de c afin de manipuler les contextes minimaux de $\overline{c}$.

Définition 42 (Décomposition du complément de c) **:** Soit c un contexte. La *décomposition du complément de* c, notée $decomp(\overline{c})$, est l'ensemble de tous les contextes minimaux disjoints de c, i.e.,

$$decomp(\overline{c}) = \{c' \in \overline{c} |\ c' \text{ est minimal }\}.$$

Les définitions précédentes manipulent un contexte ainsi que la partie de la hiérarchie disjointe de ce contexte. Dès lors, nous sommes en mesure d'évaluer l'intérêt d'un motif m dans le contexte c, d'après une mesure d'intérêt M et un seuil minimum σ sur cette mesure. Un motif intéressant selon ces paramètres est appelé un motif valide et est défini comme suit.

Définition 43 (Motif c-valide) **:** Soit deux bases d'objets $\mathcal{B}$ et $\mathcal{B}'$. Le motif m est valide dans $\mathcal{B}$ relativement à $\mathcal{B}'$ si :

1. $M(m, \mathcal{B}, \mathcal{B}') \geq \sigma$
2. $M(m, \mathcal{B}, \mathcal{B}') > M(m, \mathcal{B}', \mathcal{B})$

De plus, s'il existe un contexte c tel que $\mathcal{B} = \mathcal{B}(c)$ et $\mathcal{B}' = \mathcal{B}(\overline{c})$, alors m est simplement dit valide dans c ou *c-valide*.

Remarquons que la c-validité d'un motif nécessite de satisfaire deux conditions. La première est naturellement liée à la valeur de la mesure qui doit être supérieure à un seuil minimum donné. La deuxième assure que le motif recherché est effectivement un motif caractéristique de c et non de son complément. En effet, si le seuil fixé par l'utilisateur est excessivement bas,

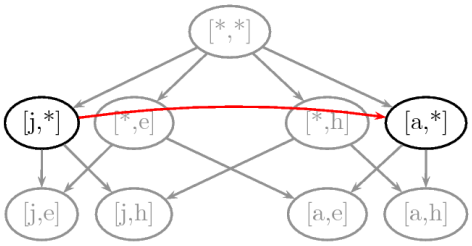


FIGURE 4.4 – Vérification de la validité d'un motif dans $[a, *]$.

l'absence de cette condition mènerait à considérer comme c-valides des motifs qui sont en réalité caractéristiques de $\overline{c}$.

Dans la suite de cette section, nous instancions cette définition sur la mesure d'émergence Em et nous fixons le seuil minimum d'émergence σ à 2. Ainsi, nous nous intéressons aux motifs dont la fréquence est au moins deux fois plus élevée dans un contexte donné, par rapport au reste de la base contextuelle.

Exemple 39 : La séquence $s = \langle (a)(b) \rangle$ est $[j, *]$-valide. En effet, les deux conditions nécessaires sont vérifiées :

1. $Em(s, \mathcal{B}([j,*]), \mathcal{B}(\overline{[j,*]})) = \frac{Freq_{[j,*]}(s)}{Freq_{\mathcal{B}(\overline{[j,*]})}(s)} = \frac{7}{8} \times \frac{6}{1} = 5,25 \geq \sigma$.
2. $Em(s, \mathcal{B}([j,*]), \mathcal{B}(\overline{[j,*]})) > Em(s, \mathcal{B}(\overline{[j,*]}), \mathcal{B}([j,*])) = 0,19$.

Comme le montre la figure 4.4, dans ce cas le calcul du taux d'émergence s'effectue uniquement entre les deux contextes $[j, *]$ et $[a, *]$. En effet, la base du complément de $[j, *]$ dans cet exemple coïncide avec la base du contexte $[a, *]$.
En revanche, s n'est pas $[j, h]$-valide. En effet, la première condition n'est pas vérifiée :

$$Em(s, \mathcal{B}([j,h]), \mathcal{B}(\overline{[j,h]})) = \frac{Freq_{[j,h]}(s)}{Freq_{\mathcal{B}(\overline{[j,h]})}(s)} = \frac{2}{3} \times \frac{11}{6} = 1,22 \leq \sigma.$$

Comme le montre l'exemple précédent, vérifier la c-validité d'un motif donné nécessite le calcul de deux valeurs : $M(m, \mathcal{B}(c), \mathcal{B}(\overline{c}))$ et $M(m, \mathcal{B}(\overline{c}), \mathcal{B}(c))$.

La notion de motif c-valide généralise celle de motif c-fréquent. En effet, un motif c-fréquent est un motif c-valide, lorsque la mesure M choisie est la fréquence. Nous nous posons à présent la question suivante : « *Que signifie la c-généralité pour les mesures d'intérêt ?* ». Dans le cadre de l'extraction de motifs fréquents contextuels, nous avons mis en avant la nécessité de considérer les descendants du contexte donné : la c-fréquence, afin d'être jugée représentative d'un contexte, doit se propager dans tous les descendants de ce contexte. Nous généralisons ce principe à l'ensemble des mesures d'intérêt.

Définition 44 (Motif c-général) **:** Le motif m est général dans c ou *c-général* pour une mesure d'intérêt $\mathcal{M}$, si $\forall (c, c') \in \hat{c} \times \overline{c}$, m est valide dans $\mathcal{B}(c)$ relativement à $\mathcal{B}(c')$.

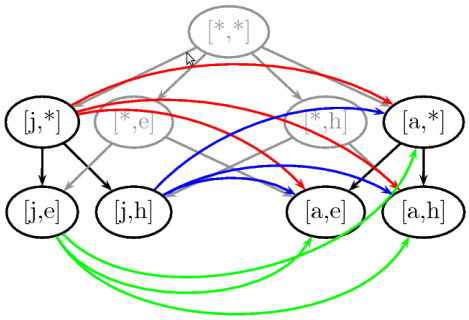


FIGURE 4.5 – Généralisation de la $[j,*]$-validité d'un motif.

Exemple 40 : La séquence $s = \langle (a)(b) \rangle$ est $[j,*]$-générale. En effet, quelque soit le contexte choisi impliquant des jeunes (i.e., $[j,*]$, $[j,e]$ ou $[j,h]$) et le contexte choisi impliquant des non-jeunes (i.e., $[a,*]$, $[a,e]$ ou $[a,h]$), alors s est émergent chez les jeunes par rapport aux moins jeunes. Par exemple[3],

- $Em(s, \mathcal{B}([j,e]), \mathcal{B}([a,*])) = 6 \geq \sigma$.
- $Em(s, \mathcal{B}([j,h]), \mathcal{B}([a,e])) = 2 \geq \sigma$.

La figure 4.5 illustre les calculs nécessaires pour vérifier la $[j,*]$-généralité de s chaque composant c de $[j,*]$ ($c \in \hat{[j,*]}$) doit être comparé à chaque contexte c' disjoint de $[j,*]$ ($c' \in \overline{[j,*]}$).

En nous appuyant sur la c-généralité, nous définissons désormais la notion de motif contextuel dans le cadre des mesures d'intérêt.

Définition 45 (Motif contextuel) **:** Un *motif contextuel* α est un couple (c,m), où c est un contexte et m un motif c-général. Dans ce cas, le motif contextuel α est dit généré par m.

Exemple 41 : Le couple $([j,*], \langle (a)(b) \rangle)$ est un motif contextuel, car la séquence $\langle (a)(b) \rangle$ est $[j,*]$-générale (voir exemple précédent).

De même que les motifs fréquents sont associés à leur valeur de fréquence, un motif contextuel est naturellement associé à une valeur quantifiant son intérêt selon la mesure choisie. Dans le cadre des motifs contextuels, cette valeur peut être définie de plusieurs manières différentes.

Définition 46 (Valeur d'intérêt de α) **:** Soit un motif contextuel $\alpha = (c,m)$. La *valeur d'intérêt* de α, notée $M_{min}(\alpha)$, est la valeur minimale de M mesurée entre un élément de $\hat{c}$ et un élément de $\overline{c}$, i.e.,

$$M(\alpha) = \min_{(c,c') \in \hat{c} \times \overline{c}} M(m, \mathcal{B}(c), \mathcal{B}(c')).$$

Exemple 42 : Considérons le motif contextuel $\alpha = ([j,*], \langle (a)(b) \rangle)$ (*Cf.* exemples précédents). Sa valeur d'intérêt est :

3. Nous ne montrons pas les calculs correspondant à chaque combinaison de contextes possibles. Il y en a $3 \times 3 = 9$.

$$Em(\alpha) = Em(s, \mathcal{B}([j,e]), \mathcal{B}([a,h])) = \tfrac{2}{3} \times \tfrac{3}{1} = 2$$

La valeur d'intérêt d'un motif contextuel (c, m) est ainsi désignée par la valeur minimale de M trouvée entre un contexte de $\hat{c}$ et un contexte de $\overline{c}$. Cette valeur joue en effet un rôle clé : par définition, le couple (c, m) est un motif contextuel si et seulement cette valeur est supérieure ou égale à σ. Nous verrons par la suite que cette remarque permet de simplifier le problème de l'extraction des motifs contextuels dans la hiérarchie de contextes.

Un motif contextuel peut également être associé à sa fréquence.

Définition 47 (Fréquence de α) **:** Soit un motif contextuel $\alpha = (c, m)$. La *fréquence de* α, notée notée $Freq(\alpha)$, est définie comme la fréquence de m dans c : $Freq(\alpha) = Freq_c(m)$.

Exemple 43 : Considérons le motif contextuel $\alpha = ([j, *], \langle (a)(b) \rangle)$ (*Cf.* exemples précédents). Sa fréquence est :

$$Freq(\alpha) = Freq_{[j,*]}(\langle (a)(b) \rangle) = \frac{7}{8}$$

4.2.1 Stratégies de sélection des motifs contextuels

Comme nous l'avons observé pour les motifs fréquents contextuels dans la section 3.2, l'ensemble de tous les motifs contextuels peut introduire une forme de redondance, car un même motif m peut générer plusieurs motifs contextuels de la forme (c, m), (c', m), etc. Ceci résulte du fait que m peut être général dans différents contextes.

Pour résoudre ce problème, nous proposons dans cette sous-section des stratégies visant à limiter le nombre de motifs contextuels à extraire. Cette sélection repose sur le principe suivant : pour chaque motif m, nous souhaitons conserver un seul motif contextuel généré par m.

Sélection par le contexte maximal. Une première stratégie de sélection repose sur les contextes associés aux motifs contextuels, en se concentrant sur le contexte maximal où la généralité d'un motif s'applique. Nous définissons dans ce but la maximalité par le contexte.

Définition 48 (Maximal par le contexte) **:** Un motif contextuel (c, m) est dit *maximal par le contexte* s'il n'existe pas de contexte c' tel que c' est un ancêtre de c et (c', m) est un motif contextuel, i.e.,

$$\nexists c' \in \mathcal{H} | (c' > c) \text{ et } m \text{ est } c'\text{-général.}$$

Cette stratégie de sélection s'apparente à celle exploitée dans le cadre des motifs contextuels fréquents. Cependant, contrairement à ce dernier cas, il ne s'agit pas ici d'une représentation condensée de l'ensemble total des motifs contextuels. En effet, nous avions noté dans le chapitre 3 que la généralité d'un motif fréquent dans c impliquait également la généralité de ce même motif dans tous les descendants de c. Dans le cas général des motifs contextuels, cette remarque ne tient plus. Par conséquent, il n'est pas possible, d'après un motif contextuel (c, m), de déduire les motifs contextuels de la forme (c', m), où c' est un descendant de c.

Sélection par la valeur maximale. Comme nous l'avons observé, un motif contextuel est associé à sa valeur d'intérêt. Celle-ci ayant pour objectif de mesurer la qualité d'un motif contextuel, un décideur peut souhaiter conserver uniquement les motifs contextuels les plus intéressants, i.e., dont la valeur d'intérêt est la plus élevée.

Définition 49 (Maximal par la valeur d'intérêt) **:** Un motif contextuel (c, m) est dit *maximal par la valeur d'intérêt* si il n'existe pas de contexte c' tel que la valeur d'intérêt de (c', m) est supérieure à celle de (c, m), i.e.,

$$\nexists c' \in \mathcal{H} | M((c', m)) > M((c, m)).$$

La définition des motifs contextuels généralisés aux mesures d'intérêt repose sur le principe clé déjà mis en avant dans le chapitre précédent : la propriété satisfaite par un motif doit pouvoir se propager dans tous les sous-contextes impliqués. Cependant, les propriétés et l'algorithme proposés pour extraire les motifs contextuels fréquents dans le chapitre 3 ne sont plus applicables. Dans la section suivante, nous mettons en valeur les propriétés théoriques des motifs contextuels généralisés pour proposer un algorithme d'extraction efficace.

4.3 Extraction de motifs contextuels

Dans cette section, nous étudions l'extraction de motifs contextuels. Pour ce faire, nous réduisons le problème de l'extraction de motifs contextuels au sous-problème suivant : *Comment, à partir d'un motif m, extraire les motifs contextuels générés par m ?*

Dans la suite, nous considérons un contexte C. Un motif m général dans un contexte C doit vérifier deux conditions pour tout $c \in \hat{C}$ et tout $c' \in \overline{C}$ (*Cf.* définition 44) :

- $M(m, c, c') \geq \sigma$;
- $M(m, c, c') > M(m, c', c)$.

Vérifier si ces conditions sont vraies nécessite de calculer deux valeurs : $M(m, c, c')$ et $M(m, c', c)$. Ainsi, une approche naïve pour contrôler la c-généralité de m nécessiterait un grand nombre de calculs : $2 \times |\hat{C}| \times |\overline{C}|$. D'autre part, notre objectif étant de trouver tous les contextes où m est général, ces calculs doivent être répétés pour chaque contexte de la hiérarchie $\mathcal{H}$. Le nombre total de calculs à effectuer pour un seul motif m sera donc : $2 \times |\hat{C}| \times |\overline{C}| \times |\mathcal{H}|$.

Afin d'extraire l'ensemble des motifs contextuels, nous nous appuyons sur l'idée suivante. La C-généralité d'un motif m est définie selon la valeur de $M(m, \mathcal{B}(c), \mathcal{B}(c)')$, pour chaque contexte $c \in \hat{C}$ et chaque contexte $c' \in \overline{C}$. Pourtant, elle dépend en réalité d'un couple de contextes uniquement : les deux contextes qui minimisent $M(m, \mathcal{B}(c), \mathcal{B}(c)')$. Dans la suite, nous focalisons sur ces deux contextes. En effet, si nous pouvons les identifier sans faire les tests requis par l'approche naïve, il sera inutile de s'intéresser aux autres contextes de la hiérarchie.

Pour cela, nous isolons quelques propriétés partagées par certaines mesures d'intérêt parmi les plus utilisées. L'exploitation de ces propriétés simplifient grandement le problème d'extraction des motifs contextuels. Nous montrons notamment qu'il est suffisant de considérer les contextes minimaux de la hiérarchie pour trouver tous les motifs générés par un motif m.

4.3.1 Propriétés pour l'extraction

Nous isolons ci-dessous trois propriétés P1, P2 et P3 vérifiées par certaines mesures d'intérêt. Nous montrons par la suite que lorsqu'une mesure vérifie ces trois propriétés le problème de l'extraction des motifs contextuels peut être simplifié.

P1. $M(m, \mathcal{B}(c), \mathcal{B}(c'))$ croît si $Freq_c(m)$ augmente et $Freq_{c'}(m)$ reste inchangée.

P2. $M(m, \mathcal{B}(c), \mathcal{B}(c'))$ décroît si $Freq_{c'}(m)$ augmente et $Freq_c(m)$ reste inchangée.

P3. Si $Freq_c(m) > Freq_{c'}(m)$, alors $M(m, \mathcal{B}(c), \mathcal{B}(c')) > M(m, \mathcal{B}(c'), \mathcal{B}(c))$.

Mesure	P1	P2	P3
Fréquence	✓	✓	✓
Fréquence conditionnelle	✓	✓	×
Confiance	✓	✓	×
Gain d'Information (GI)	✓	✓	✓
Émergence (Em)	✓	✓	✓
Lift	✓	✓	✓
Specificity	×	✓	×

TABLE 4.5 – Propriétés des mesures d'intérêt.

Le tableau 4.5 présente les propriétés vérifiées par chacune des mesures déjà détaillées dans le tableau 4.2.

4.3.2 Comment exploiter ces propriétés ?

Nous allons, au cours de cette sous-section, progressivement exploiter les propriétés de certaines mesures d'intérêt, afin de simplifier les conditions nécessaires pour tester la généralité d'un motif m dans un contexte. Nous considérons M une mesure d'intérêt vérifiant les propriétés P1, P2 et P3 et $\alpha = (C, m)$ un motif contextuel selon M.

Lemme 2 : Soit $(c, c') \in \hat{C} \times \overline{C}$.
$M(\alpha) = M(m, \mathcal{B}(c), \mathcal{B}(c'))$ si $Freq_c(m) = \min_{x \in \hat{C}}(Freq_x(m))$ et $Freq_{c'}(m) = \max_{x \in \overline{C}}(Freq_x(m))$.

Démonstration : *Ce lemme résulte des propriétés P1 et P2. En effet, pour n'importe quel c' fixé, $M(m, \mathcal{B}(c), \mathcal{B}(c'))$ est minimal quand $Freq_c(m)$ est minimal (propriété P1). De même, pour n'importe quel c fixé, $M(m, \mathcal{B}(c), \mathcal{B}(c'))$ est minimal quand $Freq_{c'}(m)$ est maximal (propriété P2).* □

Ce résultat dévoile une première information clé sur les contextes qui limitent la valeur d'intérêt d'un motif contextuel. Il s'agit des contextes c et c' tels que d'une part c minimise la fréquence de m dans $\hat{C}$ et d'autre part c' maximise la fréquence de m dans $\overline{C}$. Il suffit donc de s'appuyer sur la fréquence de m dans chacun des contextes de la hiérarchie pour les isoler. En outre, nous montrons que les contextes recherchés peuvent être trouvés parmi les contextes minimaux de la hiérarchie.

Lemme 3 : Soient m un motif et f_{min} et f_{max} respectivement les fréquences minimale et maximale de m dans un élément de $decomp(C)$:
- $f_{min} = \min_{x \in decomp(C)} Freq_x(m)$;
- $f_{max} = \max_{x \in decomp(C)} Freq_x(m)$.

Alors $Freq_C(m)$ est bornée par f_{min} et f_{max} : $f_{min} \leq Freq_C(m) \leq f_{max}$.

Démonstration : *La preuve du lemme 3 peut être trouvée dans l'annexe A.* □

Selon le lemme 3, la fréquence d'un motif m dans C est bornée. La borne inférieure est fournie par la fréquence minimale de m dans les éléments de la décomposition de C et la borne supérieure par sa fréquence maximale. Par extension, ce lemme garantit que pour tout motif m, sa fréquence la plus élevée comme sa fréquence la plus basse sont mesurées dans un contexte minimal.

Or, nous recherchons précisément les contextes qui d'une part minimisent et d'autre part maximisent la fréquence de m. Nous savons désormais que ces contextes peuvent être trouvés parmi les contextes minimaux. L'association des lemmes 2 et 3 nous fournit donc le résultat suivant.

Lemme 4 : Soit $(c, c') \in \hat{C} \times \overline{C}$. $M(\alpha) = M(m, \mathcal{B}(c), \mathcal{B}(c'))$ si $Freq_c(m) = \min_{x \in decomp(C)}(Freq_x(m))$ et $Freq_{c'}(m) = \max_{x \in decomp(\overline{C})}(Freq_x(m))$.

Démonstration : *Le lemme 4 résulte directement de l'application des lemmes 2 et 3 : les deux contextes c et c' qui minimisent $M(m, \mathcal{B}(c), \mathcal{B}(c'))$ sont tels que la fréquence de m est minimisée dans c et maximisée dans c' (Cf. lemme 2) et de plus ces contextes sont minimaux (Cf. lemme 3).* □

Le lemme 4 nous munit d'un moyen simple de retrouver la valeur d'intérêt minimale mesurée entre un contexte de $\hat{C}$ relativement à un contexte de $\overline{C}$ est supérieure ou égale au seuil fixé σ. Cependant, cela ne suffit pas à affirmer que le motif m est C-général, puisqu'il nous faut également vérifier que $M(m, \mathcal{B}(c), \mathcal{B}(c'))$ est supérieur à $M(m, \mathcal{B}(c'), \mathcal{B}(c))$ (*Cf.* la définition 44 d'un motif C-général). Nous remarquons que cette condition est facilement vérifiable pour les mesures qui satisfont la propriété P3 : elle est vraie si $Freq_c(m) > Freq_{c'}(m)$.

Par conséquent, pour les mesures d'intérêt vérifiant les propriétés P1, P2 et P3, la C-généralité d'un motif peut facilement être déterminée au travers du théorème suivant.

Théorème 2 : Soient une mesure d'intérêt M vérifiant les propriétés P1, P2 et P3 et deux contextes c et c' tels que $c = \underset{x \in decomp(C)}{\text{argmin}} (Freq_x(m))$ et $c' = \underset{y \in decomp(\overline{C})}{\text{argmax}} (Freq_y(m))$.
Un motif m est C-général si et seulement si :

1. $M(m, \mathcal{B}(c), \mathcal{B}(c')) \geq \sigma$;
2. $Freq_c(m) > Freq_{c'}(m)$.

Démonstration : *Le théorème 2 peut facilement être prouvé par le biais des résultats précédents. La définition d'un motif C-général (définition 44) implique que m est valide dans tout élément de $\hat{C}$ relativement à tout élément de $\overline{C}$. Or, pour deux contextes $c \in \hat{C}$ et $c' \in \overline{C}$, m est valide dans c relativement à c' si :*

1. $M(m, \mathcal{B}(c), \mathcal{B}(c')) \geq \sigma$;
2. $M(m, \mathcal{B}(c), \mathcal{B}(c')) > M(m, \mathcal{B}(c'), \mathcal{B}(c))$.

D'après le lemme 4, la première condition est vraie si la première condition du théorème 2 est remplie. De plus, la propriété P3 garantit que la deuxième condition est vraie si la deuxième condition du théorème 2 est remplie. Par conséquent, les conditions requises dans le théorème 2 sont suffisantes pour vérifier qu'un motif est C-général. □

Exemple 44 : Appliquons le théorème 2 sur le motif $m = \langle (a)(b) \rangle$ et le contexte $c = [j, *]$. D'après ce théorème, nous devons d'abord trouver le contexte de la décomposition de c dans lequel m a sa fréquence la plus basse. La décomposition de c est formée des deux contextes $[j, e]$ et $[j, h]$. Comme $Freq_{[j,e]}(m) = 1$ et $Freq_{[j,h]}(m) = \frac{2}{3}$, le contexte recherché est $[j, h]$.
De même, nous recherchons le contexte minimal disjoint de c qui offre la fréquence de m la plus élevée. Le contexte c possède deux contextes minimaux disjoints : $[a, e]$ et $[a, h]$. Comme $Freq_{[a,e]}(m) = \frac{1}{3}$ et $Freq_{[a,h]}(m) = 0$, le contexte recherché est $[a, e]$.
Ainsi, d'après le théorème 2, deux conditions doivent être remplies pour que s soit $[j, *]$-générale :

1. $Freq_{[j,h]}(s)$ doit être supérieure à $Freq_{[a,e]}(s)$. Cette condition est vérifiée.
2. $M(m, \mathcal{B}([j, h]), \mathcal{B}([a, e]))$ doit être supérieure ou égale à $\sigma = 2$. C'est également le cas car $M(m, \mathcal{B}([j, h]), \mathcal{B}([a, e])) = 2$.

Par conséquent, s est $[j, *]$-générale et le couple $([j, *], s)$ est un motif contextuel.
Comme le montre la figure 4.6, l'application du théorème 2 permet de vérifier la $[j, *]$-généralité de $\langle (a)(b) \rangle$ en considérant deux contextes seulement $[j, h]$ et $[a, e]$ et en ne faisant qu'un seul calcul d'émergence entre ces deux contextes.

Le théorème 2 est essentiel. Il montre que vérifier la généralité d'un motif m dans un contexte C donné repose uniquement sur deux contextes facilement identifiables parmi les contextes minimaux :

- le contexte c de $decomp(C)$ dans lequel m a une fréquence minimum,
- le contexte c' de $decomp(\overline{C})$ dans lequel m a une fréquence maximum.

Une fois ces contextes identifiés, nous obtenons deux conditions simples à tester :

- la valeur $M(m, \mathcal{B}(c), \mathcal{B}(c'))$ doit être supérieure à σ ;

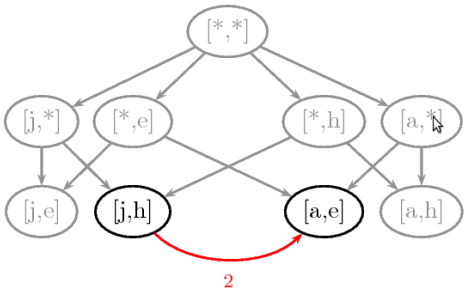


FIGURE 4.6 – Application du théorème 2 pour le motif m dans le contexte $[j, *]$.

– la fréquence de m dans c doit être supérieure à sa fréquence dans c' (i.e., $Freq_c(m) > Freq_{c'}(m)$).

Cependant, nous nous intéressons à la question suivante : « *Comment retrouver, pour un motif m donné, tous les contextes dans lesquels m est général ?* ». Pour y répondre, une exploitation naïve du théorème 2 nécessiterait de tester la généralité de m dans chaque contexte de la hiérarchie. Bien que le théorème 2 facilite ce test, le problème du nombre de contextes à traiter reste important.

Nous étudions donc la possibilité de mieux cerner les contextes dans lesquels m est général. Pour ce faire, nous étudions les contextes minimaux de la hiérarchie, ainsi que les fréquences de m dans ceux-ci, par le biais d'une liste de fréquences.

Définition 50 (Liste des fréquences de m) **:** La *liste des fréquences de m*, notée $\mathcal{L}_m = (c_1, c_2, ..., c_n)$, est la liste de tous les contextes minimaux de $\mathcal{H}$ ordonnés par ordre de fréquence de m décroissante, i.e., pour tout entier $1 \leq i < n$:

$$Freq_{c_i}(m) \geq Freq_{c_{i+1}}(m).$$

Exemple 45 : Soit la séquence $s = \langle (a)(b) \rangle$. Sa liste de fréquences $\mathcal{L}_s$ est $([j,e],[j,h],[a,e],[a,h])$. En effet, $Freq_{[j,e]}(m) < Freq_{[j,h]}(m) < Freq_{[a,e]}(m) < Freq_{[a,h]}(m)$.

La manipulation de la liste des fréquences permet facilement d'identifier les contextes nécessaires pour exploiter le théorème 2.

Théorème 3 : Soit $\mathcal{L}_m = (c_1, ..., c_n)$ la liste des fréquences de m et C un contexte. Le motif m est C-général si et seulement si il existe un entier i, avec $1 \leq i < n$, tel que $decomp(C) = \{c_1, ..., c_i\}$ et $M(m, \mathcal{B}(c_i), \mathcal{B}(c_{i+1})) \geq \sigma$.
Dans ce cas, $\alpha = (C, m)$ est un motif contextuel et de plus $M(\alpha) = M(m, \mathcal{B}(c_i), \mathcal{B}(c_{i+1}))$.

Démonstration : *Tout d'abord, notons que si $decomp(C) = \{c_1, c_2, ..., c_i\}$ alors nous obtenons, par construction de la liste de fréquences, $c_i = \underset{x \in decomp(C)}{\text{argmin}} Freq_x(m)$ et $c_{i+1} = \underset{x \in \overline{C}}{\text{argmax}} Freq_x(m)$.*
Donc, par application du théorème 2, si $M(m, \mathcal{B}(c_i), \mathcal{B}(c_{i+1})) \geq \sigma$ alors m est C-général. □

Exemple 46 : Soit la séquence $s = \langle (a)(b) \rangle$. Sa liste de fréquences est :

$$\mathcal{L}_s = ([j,e],[j,h],[a,e],[a,h])$$

Notons que cette liste peut être décomposée en deux parties telles que la partie gauche correspond à la décomposition d'un contexte $C = [j,*]$ (et donc la partie droite à la décomposition de son complément) :

$$(\overbrace{[j,e],[j,h]}^{decomp([j,*])},[a,e],[a,h])$$

Par conséquent, d'après le théorème 3, s est $[j,*]$-générale si et seulement si le dernier élément de la partie gauche $([j,h])$ et le premier élément de la partie droite $([a,e])$ vérifient :

$$M(m, \mathcal{B}([j,h]), \mathcal{B}([a,e])) \geq \sigma$$

Comme $M(m, \mathcal{B}([j,h]), \mathcal{B}([a,e])) = 2$, la séquence s est $[j,*]$-générale et le couple $\alpha = ([j,*], s)$ forme un motif contextuel de valeur d'intérêt $M(\alpha) = 2$.

Le théorème 3, en exploitant les différentes propriétés des motifs contextuels, offre un moyen d'obtenir tous les motifs contextuels générés par un motif m. Nous tirons parti de ce résultat dans la section suivante pour proposer un algorithme d'extraction de motifs contextuels approprié.

4.4 Algorithmes

Dans cette section, nous utilisons le théorème 3 pour introduire un algorithme d'extraction des motifs contextuels pour une mesure d'intérêt donnée. L'algorithme d'extraction de motifs contextuels repose sur deux étapes :

1. D'abord, il énumère les motifs m et construit leur liste de fréquences $\mathcal{L}_m$, nécessaire pour exploiter le théorème 3.
2. Puis, pour chaque motif m énuméré, il découvre les motifs contextuels générés par m. Cette deuxième partie peut être associée à une sélection des motifs contextuels, en fonction de différents critères (*Cf.* section 4.3).

4.4.1 Construction de $\mathcal{L}_m$

La première étape de l'algorithme consiste à énumérer un motif m et sa liste de fréquences, afin de générer les motifs contextuels correspondants. Une première question s'impose : « *Comment énumérer l'ensemble des motifs possibles ?* ». Comme nous l'avons observé dans le chapitre 2, le nombre de ces motifs est trop grand pour en faire une énumération exhaustive. Nous choisissons par conséquent de restreindre cette énumération aux motifs fréquents dans au moins un contexte minimal de la base. Ce choix, bien que contraignant, comporte des avantages :

– D'abord, il permet de bénéficier de la propriété d'anti-monotonie de la fréquence, déjà décrite dans le chapitre 2 et d'exploiter ainsi les algorithmes d'extraction de motifs fréquents existants. En effet, la plupart des mesures d'intérêt et en particulier le taux d'émergence et le gain d'information, ne possèdent pas une telle propriété.

– Nous conservons ainsi uniquement les motifs qui couvrent une proportion minimum choisie des données. En outre, une étude théorique effectuée dans [CYHH07] montre que sur certaines mesures telles que le gain d'information, l'intérêt d'un motif est limité lorsque celui-ci a une fréquence basse.

Lors de la recherche de motifs fréquents contextuels, nous avons déjà eu besoin d'extraire les motifs contextuels fréquents dans au moins un contexte minimal. Nous pouvons par conséquent utiliser l'approche d'extraction décrite dans le chapitre 3 (section 3.4.1). L'unique différence apportée ici est dans le tri effectué en post-traitement afin de construire la liste de fréquences $\mathcal{L}_m$.

4.4.2 Génération des motifs contextuels

La deuxième étape de l'algorithme d'extraction de motifs contextuels réside dans la génération des motifs contextuels, pour chaque motif m énuméré dans la première étape.

Cette étape, basée sur le théorème 3, est effectuée par l'algorithme 5 afin de trouver, pour un motif donné m, les contextes où m est général. D'après la liste de fréquences du motif m, il repose sur deux étapes principales :

1. Parcourir la liste pour trouver les contextes minimaux tels que les éléments à leur gauche (i.e., dans lesquels m a une fréquence égale ou supérieure) forment la décomposition d'un contexte.
2. Pour chaque contexte minimal trouvé, tester la validité de m relativement à l'élément suivant de la liste.

En effet, le théorème 3 montre que ces deux conditions sont à la fois nécessaires et suffisantes pour générer un motif contextuel, validant ainsi la correction et la complétude de l'algorithme associé.

Revenons à présent à la première étape. Celle-ci nécessite de répondre à la question suivante : « *Pour un élément donné de la liste $\mathcal{L}_m$, existe-t-il un contexte dont la décomposition correspond à la partie gauche de la liste ?* » Pour répondre à cette question, l'algorithme s'appuie sur l'information déjà obtenue pour l'élément précédent. Par exemple, considérons la liste de fréquences $\mathcal{L}_m = ([j,e],[j,h],[a,e],[a,h])$. Lors de la première itération, la partie gauche de la liste est uniquement constituée du contexte $[j,e]$. La réponse est alors immédiate : le contexte correspondant est $[j,e]$ lui-même. Lors de la deuxième itération, la partie gauche est constituée des deux contextes $[j,e]$ et $[j,h]$. Nous remarquons que s'il existe un contexte répondant à nos attentes, sa décomposition contient $[j,e]$. Il s'agit par conséquent d'un ancêtre de $[j,e]$. L'algorithme utilise cette remarque pour construire à chaque itération un ensemble de contextes candidats correspondant aux parents du contexte précédemment visité.

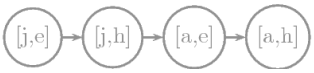
Algorithm 5 Génération de motifs généraux

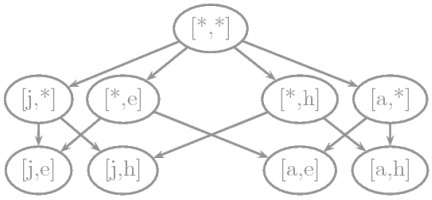
Entrées: une mesure de fréquences M, une hiérarchie de contextes $\mathcal{H}$, un seuil minimum σ.
Sorties: l'ensemble de tous les motifs contextuels générés par m.

$\mathcal{C} \leftarrow \{c_1\}$
pour tout $i \in \{1, ..., n\}$ **faire**
  $\mathcal{C}' \leftarrow \emptyset$
  pour tout $c \in \mathcal{C}$ **faire**
    si $c_i \in decomp(c)$ **alors**
      si $i = |decomp(c)|$ **alors**
        si $M(m, \mathcal{B}(c_i), \mathcal{B}(c_{i+1})) \geq \sigma$ **alors**
          output (c, m)
        finsi
        $\mathcal{P} \leftarrow$ { parents de c dans $\mathcal{H}$ }
        $\mathcal{C}' \leftarrow \mathcal{C}' \cup \mathcal{P}$
      sinon
        $\mathcal{C}' \leftarrow \mathcal{C}' \cup \{c\}$
      finsi
    finsi
  fin pour
  $\mathcal{C} \leftarrow \mathcal{C}'$
fin pour

Afin d'illustrer le fonctionnement de l'algorithme, nous en présentons ci-dessous un exemple d'exécution.

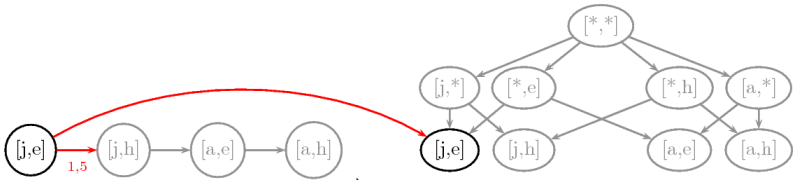
Exemple 47 : Considérons la séquence $s = \langle (a)(b) \rangle$. Dans un premier temps, nous construisons $\mathcal{L}_s$ la liste des contextes minimaux triée en fonction de la fréquence de s et obtenons $\mathcal{L}_s = ([j, e], [j, h], [a, e], [a, h])$.



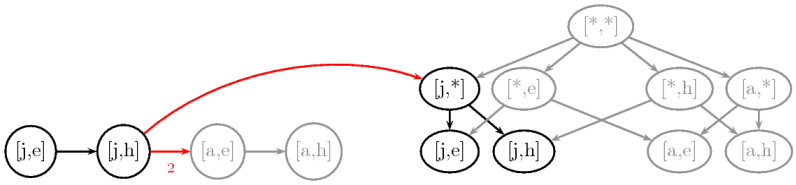


Nous parcourons ensuite cette liste de gauche à droite, afin de trouver les contextes candidats, i.e., dont tous les éléments de la décomposition sont à gauche de $\mathcal{L}_s$:

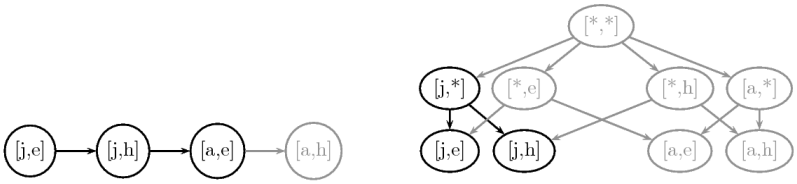
1. Le premier contexte minimal parcouru est $[j, e]$. Celui-ci est un candidat, puisque tous les éléments de sa décomposition sont en effet à gauche de la liste (car $decomp([j, e]) = \{[j, e]\}$). Afin de vérifier si s est $[j, e]$-générale, il suffit désormais de calculer l'émergence de s dans $[j, e]$, relativement au contexte minimal suivant, i.e., $[j, h]$. Nous trouvons $M(s, \mathcal{B}([j, e]), \mathcal{B}([j, h])) = 1, 5 < \sigma$. Par conséquent, s n'est pas $[j, e]$-générale.



2. Nous passons à l'élément suivant de la liste : $[j,h]$. Afin de vérifier s'il existe un contexte candidat à ce stade, i.e., un contexte dont la décomposition est composée de $[j,e]$ et $[j,h]$, il est uniquement nécessaire de considérer les parents de $[j,e]$ dans la hiérarchie de contextes. Ceux-ci sont $[j,*]$ et $[*,e]$. Ce dernier n'est pas candidat, car sa décomposition ne contient pas $[j,h]$. En revanche, $[j,*]$ est un candidat. De plus, s est émergente dans $[j,*]$ relativement au contexte minimal suivant : $M(s,\mathcal{B}([j,h]),\mathcal{B}([a,e]))$. Ainsi, selon le théorème 3, s est $[j,*]$-générale.



3. L'élément suivant de la liste est $[a,e]$. Afin de trouver les contextes candidats, nous regardons les parents du dernier candidat $[j,*]$. Celui-ci n'a qu'un unique parent $[*,*]$, qui n'est pas candidat.



4. L'élément suivant de la liste est le dernier. Par conséquent, l'algorithme s'arrête. En effet, il n'est plus possible de décomposer la liste en deux parties non-vides.

Lors du parcours de $\mathcal{L}_s$, nous n'avons trouvé qu'un seul contexte où s est générale. Il s'agit de $[j,*]$.

4.4.3 Stratégies de sélection

Nous avons dans la section 4.3 défini deux stratégies de sélection : par le contexte maximal où un motif m est général ou par la valeur d'intérêt maximale de m dans un contexte. L'algorithme 5 peut aisément être adapté afin d'intégrer ces stratégies.

4.4.3.1 Sélection par le contexte maximal

Dans l'algorithme 6 nous montrons comment extraire, pour un motif m, le motif contextuel maximal par le contexte généré par m. Le principe est le suivant. Au cours du parcours de la liste de fréquences $\mathcal{L}_m$, la variable C_{max} stocke le dernier contexte où m a été général. A la fin du processus, le motif contextuel (C_{max},m) est retourné. Il s'agit du motif contextuel maximal par le contexte.

Algorithm 6 Génération des motifs contextuels maximaux par le contexte

Entrées: une mesure de fréquences M, une hiérarchie de contextes $\mathcal{H}$, un seuil minimum σ.
Sorties: l'ensemble de tous les motifs contextuels générés par m.

$C_{max} \leftarrow null$
$\mathcal{C} \leftarrow \{c_1\}$
pour tout $i \in \{1, ..., n\}$ **faire**
  $\mathcal{C}' \leftarrow \emptyset$
  pour tout $c \in \mathcal{C}$ **faire**
    si $c_i \in decomp(c)$ **alors**
      si $i = |decomp(c)|$ **alors**
        si $M(m, \mathcal{B}(c_i), \mathcal{B}(c_{i+1})) \geq \sigma$ **alors**
          $c_{max} \leftarrow c$
        finsi
        $\mathcal{P} \leftarrow$ { parents de c dans $\mathcal{H}$ }
        $\mathcal{C}' \leftarrow \mathcal{C}' \cup \mathcal{P}$
      sinon
        $\mathcal{C}' \leftarrow \mathcal{C}' \cup \{c\}$
      finsi
    finsi
  fin pour
  $\mathcal{C} \leftarrow \mathcal{C}'$
fin pour
retourne (C_{max}, m)

4.4.3.2 Sélection par la valeur d'intérêt maximale

L'algorithme 7, pour un motif m, extrait le motif contextuel maximal par la valeur d'intérêt généré par m. Au cours du parcours de la liste de fréquences $\mathcal{L}_m$, la variable V_{max} conserve la valeur maximale rencontrée jusque-là pour un motif contextuel généré par m. La variable C_{max} stocke quant à elle le dernier contexte où cette valeur a été mesurée. A la fin du processus, le motif contextuel (C_{max}, m) est retourné, il s'agit du motif contextuel maximal par la valeur d'intérêt. Notons toutefois qu'il peut exister plusieurs motifs contextuels maximaux par la valeur d'intérêt. En effet, plusieurs motifs contextuels générés par m peuvent avoir la même valeur d'intérêt. Dans ce cas, l'algorithme 7 conserve uniquement celui qui maximise le contexte.

4.4.4 Algorithme général

Les algorithmes présentés précédemment résolvent des sous-problèmes de l'extraction de motifs contextuels. L'algorithme 8, en revanche, présente le processus général en deux étapes :

1. La première consiste en l'extraction de tous les motifs fréquents dans au moins un contexte minimal. Nous avons déjà vu dans le chapitre précédent comment cette étape peut être réalisée. De plus, cette étape s'accompagne de la construction de la liste de fréquences de chacun des motifs.
2. La deuxième étape correspond à la génération des motifs contextuels pour chaque motif extrait précédemment. Cette étape peut varier en fonction des stratégies de sélection choisies par l'utilisateur et ainsi faire appel aux différents algorithmes dans cette section :

Algorithm 7 Génération des motifs contextuels maximaux par la valeur d'intérêt

Entrées: une mesure de fréquences M, une hiérarchie de contextes $\mathcal{H}$, un seuil minimum σ.
Sorties: l'ensemble de tous les motifs contextuels générés par m.

$C_{max} \leftarrow null$
$V_{max} \leftarrow \sigma$
$\mathcal{C} \leftarrow \{c_1\}$
pour tout $i \in \{1, ..., n\}$ **faire**
  $\mathcal{C}' \leftarrow \emptyset$
  pour tout $c \in \mathcal{C}$ **faire**
    si $c_i \in decomp(c)$ **alors**
      si $i = |decomp(c)|$ **alors**
        si $M(m, \mathcal{B}(c_i), \mathcal{B}(c_{i+1})) \geq V_{max}$ **alors**
          $c_{max} \leftarrow c$
        finsi
        $\mathcal{P} \leftarrow$ { parents de c dans $\mathcal{H}$ }
        $\mathcal{C}' \leftarrow \mathcal{C}' \cup \mathcal{P}$
      sinon
        $\mathcal{C}' \leftarrow \mathcal{C}' \cup \{c\}$
      finsi
    finsi
  fin pour
  $\mathcal{C} \leftarrow \mathcal{C}'$
fin pour
retourne (C_{max}, m)

extraction de tous les motifs contextuels (algorithme 5), des motifs contextuels maximaux par le contexte (algorithme 6) ou des motifs contextuels maximaux par la valeur d'intérêt (algorithme 7).

Algorithm 8 `CoPaM`

Entrées: une base contextuelle de séquences $\mathcal{CB}$, une mesure d'intérêt M, un seuil minimum d'intérêt σ, un seuil minimum de fréquence ρ, une hiérarchie de contextes $\mathcal{H}$.

`/* Extraction des motifs fréquents requis */`
Extraire les couples $(m, \mathcal{L}_m)$ tels que m est fréquent dans au moins un contexte minimal

`/* Génération des motifs contextuels */`
pour tout $(m, \mathcal{L}_m)$ **faire**
  générer et afficher les motifs contextuels (c, m)
fin pour

4.5 Expérimentations

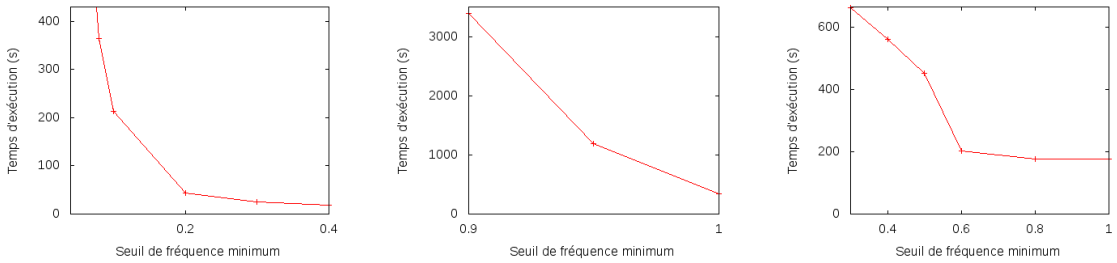
Nous évaluons dans un premier temps l'efficacité des algorithmes proposés dans la section 4.4 pour l'extraction de motifs contextuels. Les jeux de données utilisés sont les mêmes que dans le chapitre 3. Bien que nous ayons démontré dans la section 4.3 que l'approche fonctionne

pour plusieurs mesures d'intérêt (émergence, gain d'information, lift et fréquence), nous nous concentrons dans ces expérimentations sur la mesure d'émergence. En effet, l'utilisation d'une mesure par rapport à une autre change de manière négligeable le comportement des algorithmes.

La première étape de l'algorithme `CoPaM` consiste à extraire tous les motifs m fréquents dans au moins un contexte minimal et à les associer à leur liste de fréquences, i.e., la liste des contextes minimaux de $\mathcal{H}$ triées par la fréquence de m décroissante. Cette étape est similaire à celle utilisée dans le chapitre 3. En effet, il n'existe que deux différences par rapport à celle-ci :

- Nous avons ici besoin de connaître la fréquence de chaque motif m dans chaque contexte minimal, alors que pour l'extraction des motifs fréquents contextuels, nous n'utilisions que les contextes minimaux où m était fréquent (i.e., $\mathcal{F}_m$). Par conséquent, dans ces expérimentations, nous exploitons l'adaptation de `PrefixSpan` sans l'optimisation liée aux bases projetées (*Cf.* section 3.4.1 du chapitre précédent) de manière à pouvoir compter la fréquence des motifs dans chaque contexte minimal.
- La seconde différence consiste à construire la liste de fréquences pour chaque motif. Cette étape étant triviale, nous ne la détaillons pas.

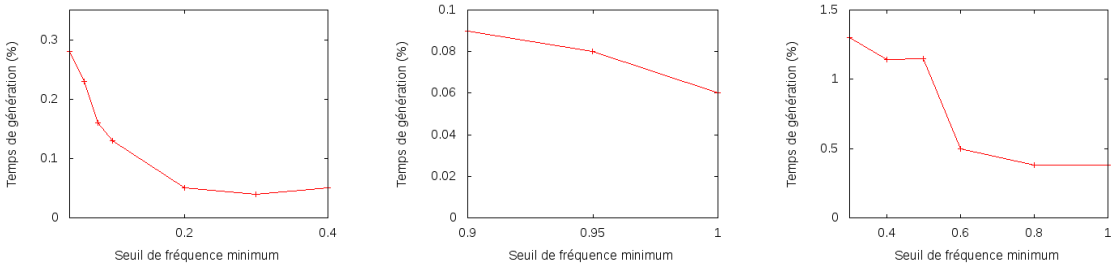
La figure 4.7 présente le temps d'exécution de l'algorithme `CoPaM` afin d'extraire les motifs contextuels émergents en fonction du seuil minimum de fréquence et lorsqu'aucun seuil minimum d'émergence n'est fixé (i.e., tous les motifs contextuels émergents sont extraits). Comparés à l'algorithme `CFPM` présenté et évalué dans le chapitre précédent, `CoPaM` est plus coûteux. Ce coût supplémentaire s'explique principalement par les différences qui existent entre les étapes d'extraction des motifs fréquents dans les contextes minimaux.



(a) *Amazon* (b) *Puces à ADN* (c) *Consommation énergétique*

FIGURE 4.7 – Temps d'exécution de l'algorithme `CoPaM` en fonction du seuil minimum de fréquence.

La figure 4.8 montre la proportion du temps d'exécution de l'algorithme `CoPaM` due à la génération des motifs contextuels d'intérêt. Comme pour l'extraction de motifs fréquents contextuels (*Cf.* chapitre précédent), l'étape de génération des motifs contextuels a un coût généralement négligeable dans le processus global.



(a) *Amazon* (b) *Puces à ADN* (c) *Consommation énergétique*

FIGURE 4.8 – Proportion du temps de génération des motifs contextuels dans l'algorithme `CoPaM` en fonction du seuil minimum de fréquence.

4.6 Discussion

Dans ce chapitre, nous avons dépassé les limitations inhérentes aux motifs fréquents contextuels proposés dans le chapitre 3, en généralisant les notions associées aux diverses mesures d'intérêt proposées dans la littérature. Par là même, nous avons adapté les motifs d'intérêt classiques (motifs émergents, motifs discriminants, etc.), habituellement définis dans un cadre « *plat* » où les différentes classes de données sont disjointes et sans relation hiérarchique au cadre des données contextuelles.

Dans la section 4.1, nous avons décrit la notion de mesure d'intérêt dans l'extraction de motifs puis, dans la section 4.2, nous l'avons adaptée pour l'extraction de motifs contextuels. Or, les définitions proposées rendent la tâche d'extraction particulièrement difficile. Nous avons donc montré dans la section 4.3 que les propriétés partagées par certaines mesures d'intérêt (émergence, gain d'information, etc.) offrent un cadre plus accessible. En dévoilant puis en exploitant ces propriétés, nous avons été en mesure de proposer une approche d'extraction efficace dont nous avons donné les algorithmes dans la section 4.4. Finalement, l'efficacité de l'approche proposée a été montrée par les expérimentations présentées dans la section 4.5 sur trois jeux de données.

La méthode d'extraction de motifs contextuels décrite dans ce chapitre considère uniquement les mesures d'intérêt qui vérifient les propriétés isolées (*Cf.* section 4.3). Toutefois, d'autres mesures pourraient être intégrées à condition de trouver des méthodes adaptées.

Souvent, les motifs d'intérêt classiques sont utilisés dans des tâches de classification. Concernant les motifs contextuels, une question se pose alors : « *Comment exploiter les motifs contextuels dans une tâche de classification ?* ». Nous répondons à cette question dans le chapitre suivant.

Chapitre 5

Classification et motifs contextuels

Sommaire

Introduction

Le problème de la classification supervisée est aujourd'hui un des problèmes majeurs en fouille de données. De manière générale, la classification est le processus qui vise à affecter une classe à un objet. Dans notre cas d'étude, il s'agira par exemple d'attribuer un âge parmi les valeurs *jeune* ou *âgé* à un habitant, étant donné sa séquence d'activités. Dans le cadre supervisé, la méthode de classification est induite d'un ensemble d'exemples labellisés, i.e., dont on connaît la classe *a priori*. Depuis la fin des années 1990, le problème de classification supervisée a été abordé sous un jour nouveau, en tirant bénéfice du pouvoir discriminant de certains motifs.

Dans ce manuscrit, nous avons jusque là traité le problème de la prise en compte d'informations contextuelles lors de l'extraction de motifs fréquents (*Cf.* chapitre 3) ou répondant à une contrainte d'intérêt minimum (*Cf.* chapitre 4). Les motifs contextuels peuvent, par exemple, révéler comment les activités des habitants dépendent du contexte (l'*âge* de l'habitant ou la *saison*). Dès lors, une question s'impose : « *Peut-on exploiter les motifs contextuels pour enrichir le processus de classification ?* ».

En effet, de même que des informations contextuelles peuvent être prises en compte lors de l'extraction de motifs, le processus de classification peut également bénéficier d'informations additionnelles. Pour nous en convaincre, considérons l'habitant auquel nous souhaitons associer une classe (*jeune* ou *âgé*) d'après sa séquence d'activités s. Dans le cadre des données contextuelles, ce problème consiste à associer à s le contexte correspondant aux jeunes $[j, *]$ ou aux âgés $[a, *]$. Le tableau 5.1(a) présente notre base d'apprentissage qui est ici une base contextuelle. Nous considérons dans ce tableau un motif m ainsi que les objets de la base qui le supportent. Nous posons la question : « *Quelle classe affecter à la séquence s, sachant que s supporte m ?* ».

Nous observons que m est supporté par 4 séquences sur 8 dans la classe correspondant aux *jeunes* habitants et par 3 séquences sur 6 dans celle correspondant aux habitants *âgés*. Le problème de classification de s consiste donc à lui associer un âge, en sachant que s supporte le motif m. La fréquence de m étant identique pour les deux classes *jeune* et *âgé*, il est difficile de trancher en faveur de l'une ou l'autre des classes.

Supposons à présent qu'une information supplémentaire soit disponible : la séquence s a été enregistrée en été. Cette information contextuelle peut être exploitée comme un guide pour la classification puisqu'il s'agit maintenant d'associer s à l'un des deux contextes $[j, e]$ ou $[a, e]$, plus spécifiques que $[j, *]$ et $[a, *]$. Nous obtenons la base d'apprentissage réduite à ces deux contextes, décrite dans le tableau 5.1(b). Le problème de décision devient alors plus aisé : aucune séquence dans la classe *âgé* ne supporte m, tandis que 3 sur 5 le supportent dans la classe *jeune*. Cette dernière est donc naturellement choisie. Ainsi, le motif m qui n'était initialement pas utile pour la classification s'est révélé particulièrement précieux dès lors que nous avons considéré les informations contextuelles.

Cet exemple montre que tenir compte des informations contextuelles disponibles sur l'objet à classer permet d'identifier de nouveaux motifs discriminants et d'enrichir ainsi le processus de classification.

Néanmoins, l'adaptation des principes de la classification basée sur les motifs dans les données

(a) Problème classique de classification basée sur un motif.

id	Age	Saison	m
s_1	jeune	été	×
s_2	jeune	été	
s_3	jeune	été	×
s_4	jeune	été	
s_5	jeune	été	×
s_6	jeune	hiver	
s_7	jeune	hiver	×
s_8	jeune	hiver	
s_9	âgé	été	
s_{10}	âgé	été	
s_{11}	âgé	été	
s_{12}	âgé	hiver	×
s_{13}	âgé	hiver	×
s_{14}	âgé	hiver	×

(b) Prise en compte de l'information *été*.

id	Age	Saison	m
s_1	jeune	été	×
s_2	jeune	été	
s_3	jeune	été	×
s_4	jeune	été	
s_5	jeune	été	×
s_9	âgé	été	
s_{10}	âgé	été	
s_{11}	âgé	été	

TABLE 5.1 – Une base contextuelle de séquences.

contextuelles soulève certaines difficultés. En particulier, tous les contextes possibles pour décrire l'objet doivent être pris en compte. Dans l'exemple précédent, la classification classique (i.e., sans prise en compte du contexte) nécessite d'extraire des motifs discriminants dans deux classes *jeune* et *âgé* (i.e., $dom(Age)$). En revanche, si l'on considère les informations contextuelles, il faudra effectuer cette même tâche pour chaque valeur possible sur la dimension *Saison* : *été*, *hiver* et $*$ (i.e., aucune information n'est disponible). Ainsi, nous devrons considérer au total $dom(Age) \times dom'(Saison) = 2 \times 3 = 6$ contextes : $[j, *]$, $[a, *]$, $[j, e]$, $[a, e]$, $[j, h]$ et $[a, h]$. Le problème d'extraction des motifs pour la classification est donc plus difficile lorsque l'on souhaite intégrer les informations contextuelles.

Une idée intuitive pour extraire les motifs nécessaires à la classification consiste à adopter les motifs contextuels définis dans le chapitre précédent. Néanmoins, ces motifs ne peuvent être exploités immédiatement. Ceux-ci visent, en effet, à obtenir les motifs qui différencient un contexte de l'ensemble du reste de la hiérarchie. Pour la classification, nous sommes en revanche uniquement intéressés par les motifs qui différencient les contextes dans lesquels un objet peut être classé connaissant les informations contextuelles associées. Dans l'exemple donné plus haut, nous recherchons les motifs qui différencient le contexte $[j, e]$ du contexte $[a, e]$ uniquement et non de tout le reste de la hiérarchie.

Dans un premier temps, nous modélisons le problème de classification en utilisant les données contextuelles définies et exploitées dans les chapitres 3 et 4. Dès lors, nous nous appuyons sur cette modélisation pour aborder le problème de l'extraction de motifs contextuels pour la classification puis proposons une méthode de classification basée sur les motifs contextuels. Cette

première contribution ne dépend pas du type de motifs considérés (itemsets, motifs séquentiels, etc.). Nous nous intéressons par la suite à l'application de cette méthode dans les données séquentielles et observons finalement que le problème de classification peut facilement s'appliquer à un problème de prédiction. Par exemple, la question « *La consommation d'électricité sera-t-elle élevée ou basse dans une heure ?* » peut se traduire par « *Quel label de classe associer à la consommation d'électricité dans une heure : basse ou élevée ?* ». En suivant ce constat, nous montrons que les motifs contextuels dans les données séquentielles se révèlent alors très utiles.

La suite de ce chapitre s'organise comme suit. La section 5.1 présente un panorama des travaux existant pour la classification supervisée basée sur les motifs et discute les caractéristiques des différentes catégories d'approches. Dans la section 5.2, nous définissons formellement le problème de classification basée sur les motifs contextuels et soulignons les différences qui existent entre les motifs contextuels définis dans le chapitre 4 et les motifs nécessaires pour résoudre le problème de classification. Ces motifs sont ensuite exploités en adaptant une méthode existante de classification basée sur les motifs émergents. Nous nous intéressons dans la section 5.3 au cas des données séquentielles et montrons comment la classification basée sur les motifs contextuels peut également être utilisée pour résoudre un problème de prédiction dans une séquence. Dans la section 5.4, l'approche proposée est évaluée sur différents jeux de données réels. Enfin, dans la section 5.5, nous discutons les contributions présentées dans ce chapitre.

5.1 Classification basée sur les motifs

La classification supervisée basée sur les motifs a fait l'objet de nombreux travaux dans les quinze dernières années. Intuitivement, ils exploitent le pouvoir discriminant de certains motifs, mesuré par le biais de mesures d'intérêt, pour classer de nouveaux objets. Les propositions dans ce domaine reposent généralement sur les étapes suivantes [BNZ09] :

1. Des motifs vérifiant certaines contraintes sont extraits de la base d'apprentissage. Selon les approches, il peut s'agir d'extraire l'ensemble des motifs fréquents sur la base ou encore des motifs discriminants d'une classe de données.
2. Une sélection est réalisée sur l'ensemble des motifs extraits de manière à obtenir un ensemble de motifs adéquat selon trois critères principaux : *(1)* l'ensemble des motifs doit être représentatif des données d'apprentissage (i.e., couvrir une majorité des données d'apprentissage), *(2)* chaque motif doit être considéré comme significatif (i.e., respecter une contrainte de valeur minimum sur une mesure d'intérêt donnée) et *(3)* il doit être concis et sans redondance.
3. Enfin, l'ensemble des motifs est utilisé pour construire le classifieur.

Bien que les méthodes de classification basée sur les motifs suivent toutes ces étapes, elles peuvent être catégorisées en trois familles d'approches que nous décrivons ci-dessous.

Classification associative La classification associative repose sur des règles de la forme $m \rightarrow C$, où m est un motif et C un label de classe, sélectionnées en fonction de leur valeur pour une

mesure d'intérêt donnée. De nombreuses stratégies de sélection existent pour sélectionner le meilleur ensemble de règles parmi celles extraites [CG09]. Une fois les règles extraites et filtrées, le modèle de classification peut être construit. Deux familles d'approches co-existent dans la littérature.

- La première consiste à choisir la « *meilleure règle* » parmi l'ensemble des règles qui s'appliquent sur l'objet à classer (i.e., dont l'antécédent est supporté par cet objet). La classe affectée à l'objet est alors le conséquent de cette règle. Les approches divergent sur la notion de « *meilleure règle* ». Par exemple, l'approche pionnière de classification associative `CBA` [LHM98] sélectionne la règle de plus haute confiance.
- La seconde famille de méthodes considère en revanche l'ensemble des règles qui s'appliquent sur l'objet puis calcule un score agrégé pour chaque classe. La classe finalement affectée à l'objet est celle qui obtient le meilleur score. La méthode `CMAR` [LHP01] s'appuie, par exemple, sur un score pondéré basé sur le χ^2.

Construction d'un modèle dans l'espace des motifs Un autre type d'approches de classification basée sur les motifs repose sur le principe suivant : les motifs extraits et sélectionnés forment un nouvel espace d'attributs dans lequel sont décrites les données. L'extraction de motifs fréquents dans les données est suivie d'une étape de sélection. Notons que le nombre de motifs extraits dans la première étape est souvent très grand et peut entraîner les problèmes suivants. Souvent, la proportion de motifs réellement intéressants pour la classification (discriminants et non redondants) est minime parmi les motifs extraits. Dans ce cas, une grande partie des motifs deviennent inutiles et doivent être supprimés. De plus, les algorithmes classiques de sélection d'attributs ne peuvent faire face à cette explosion du nombre de motifs fréquents extraits [CYHY08]. Pour ces raisons, de nouvelles approches ont vu le jour afin d'intégrer, partiellement ou entièrement, l'étape de sélection des motifs pendant l'étape d'extraction [CYHY08, NGDR09].

La construction du modèle de classification est ensuite elle-même composée de deux phases :

- Les données initiales doivent être transformées et représentées dans l'espace des motifs (i.e., dans cet espace, chaque motif devient un attribut).
- Une méthode de classification quelconque (SVM [CV95], C4.5 [Qui93], etc.) est employée dans cette nouvelle base d'apprentissage.

La construction de modèles de classification dans un espace formé par des motifs offre une certaine flexibilité puisque tout algorithme d'apprentissage pour la classification peut être utilisé. La difficulté de la tâche repose alors uniquement sur les étapes d'extraction et de sélection des motifs dans les données. Néanmoins, l'aisance d'interprétabilité fournie par les motifs est perdue puisque les modèles construits sont généralement eux-mêmes difficiles à déchiffrer.

Approches basées sur les motifs émergents Une troisième catégorie d'approches concerne les classifieurs à base de motifs émergents. Les motifs émergents définis dans [DL99] sont, intuitivement, les motifs dont la fréquence change significativement d'une classe à l'autre. Définis selon la mesure d'émergence déjà employée dans le chapitre 4, leur pouvoir de discriminance a été exploité pour la classification.

[RF07] fournit une étude des approches de classification à base de motifs émergents et en décrit la méthodologie générale. Dans un premier temps, à partir d'un jeu de données d'appren-

tissage constitué de n classes, les motifs émergents sont extraits pour chaque classe, relativement à l'ensemble du reste des données. L'ensemble des motifs obtenu est ensuite post-traité, afin de sélectionner les « *meilleurs* » motifs émergents. Puis, la classification d'un nouvel objet consiste à calculer un score pour chaque classe, basé sur les motifs supportés par celui-ci. La classe assignée à l'objet est celle où il obtient le score le plus élevé.

Les premiers travaux ont produit l'algorithme `CAEP` [DZWL99] (*Classification by Aggregating Emerging Patterns*) que nous décrivons plus en détails dans la section 5.2. Par la suite, d'autres approches ont vu le jour, telles que `JEPC` [LDR01] (*Jumping Emerging Pattern Classifier*) qui exploite uniquement les motifs ayant une émergence infinie dans une classe (i.e., ces motifs ont une fréquence non nulle dans cette classe et nulle ailleurs) ou encore `BCEP` [FR03] (*Bayesian Classification based on Emerging Patterns*) qui exploite les caractéristiques du classifieur naïf de bayes [DH73].

Notons que la classification basée sur les motifs émergents peut être vue comme un cas particulier de la classification associative où la mesure d'intérêt choisie pour construire les règles est la mesure d'émergence. Cependant, deux différences fondamentales peuvent être soulignées. La classification associative s'appuie sur les motifs fréquents dans l'ensemble des données pour construire les règles de classification. Cette étape peut poser problème lorsque les classes sont inégalement réparties. Ainsi, une classe en minorité par rapport aux autres générera peu ou pas de règles. Dans le cas de la classification basée sur les motifs émergents, les motifs sont extraits indépendamment dans chaque classe, sans tenir compte de leur taille. De plus, la mesure d'émergence, pour un motif donné, dépend uniquement de sa fréquence dans les différentes classes. La plupart des mesures d'intérêt utilisées dans la classification associative sont en revanche sensibles à la répartition des données dans les classes. Ces mesures peuvent avoir pour effet de favoriser les motifs présents dans les classes majoritaires, alors même que leur fréquence y est faible. Cet aspect a notamment été discuté dans [Gay09].

Ces deux caractéristiques concernent l'inégalité de répartition des données dans les classes. Or, dans le cadre des données contextuelles qui nous intéresse ici, nous avons déjà souligné cet aspect et la nécessité d'en tenir compte. À notre connaissance, aucune des approches existantes n'intègre des informations contextuelles. Afin de combler cette absence, nous proposons une méthode de classification basée sur des motifs contextuels émergents afin d'intégrer les informations contextuelles associées aux données à classer et nous abstraire des problèmes posés par l'inégalité de répartition, inhérente aux données contextuelles. Par ailleurs, bien que la plupart des approches de classification basée sur les motifs considèrent des motifs ensemblistes (itemsets), ces approches peuvent aisément être adaptées à d'autres motifs tels que les motifs séquentiels ou encore les motifs inter-transactionnels. De plus, pour ce dernier cas, nous montrons dans la section 5.3 qu'un changement de représentation du problème de classification peut aborder un problème de prédiction d'item dans une séquence étendue.

5.2 Intégration des motifs contextuels pour la classification

Dans cette section, nous décrivons formellement le problème de classification lorsque des informations contextuelles sont disponibles sur les objets à classer. Les motifs contextuels définis

dans le chapitre 4 n'étant pas directement exploitables pour résoudre ce problème, nous adaptons leur définition.

5.2.1 Présentation du problème

Dans la suite de ce chapitre, nous utilisons la base contextuelle de séquences ainsi que la hiérarchie de contextes associées à notre cas d'étude, respectivement rappelées dans le tableau 5.2 et la figure 5.1.

id	Age	Saison	Sequence
s_1	jeune	été	$\langle(ad)(b)\rangle$
s_2	jeune	été	$\langle(ab)(b)\rangle$
s_3	jeune	été	$\langle(a)(a)(b)\rangle$
s_4	jeune	été	$\langle(c)(a)(bc)\rangle$
s_5	jeune	été	$\langle(d)(ab)(bcd)\rangle$
s_6	jeune	hiver	$\langle(b)(a)\rangle$
s_7	jeune	hiver	$\langle(a)(b)(a)\rangle$
s_8	jeune	hiver	$\langle(d)(a)(bc)\rangle$
s_9	âgé	été	$\langle(ab)(a)(bd)\rangle$
s_{10}	âgé	été	$\langle(bcd)\rangle$
s_{11}	âgé	été	$\langle(bd)(a)\rangle$
s_{12}	âgé	hiver	$\langle(e)(bcd)(a)\rangle$
s_{13}	âgé	hiver	$\langle(bde)\rangle$
s_{14}	âgé	hiver	$\langle(b)(a)(e)\rangle$

TABLE 5.2 – Une base contextuelle de séquences.

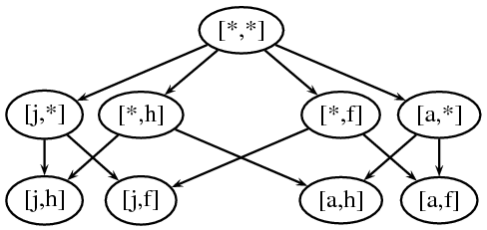


FIGURE 5.1 – La hiérarchie de contextes $\mathcal{H}$.

Afin de définir le problème de classification dans les données contextuelles, nous considérons que nous avons à disposition d'une part des dimensions contextuelles décrivant l'objet à classer et d'autre part une dimension pour laquelle nous cherchons à associer un label de classe. Plus formellement, nous définissons le partitionnemement des dimensions contextuelles ci-dessous.

Définition 51 (Partitionnement des dimensions contextuelles) **:** Soit $\{D_1, D_2, ..., D_n\}$ l'ensemble des dimensions contextuelles. Cet ensemble contient :

- **Une dimension de classe** D^c. Il s'agit de la dimension dont le domaine correspond aux différents labels de classes. Ainsi, l'objectif de la classification est d'associer à un nouvel objet une valeur sur cette dimension. Cette dimension doit nécessairement avoir dans son domaine enrichi une valeur $*$ généralisant toute autre valeur du domaine, i.e., telle que $* \in dom'(D^c)$ et $\forall d^c \in dom'(D^c), d^c \subseteq_{D^c} *$.
- **Un ensemble des dimensions guides** $\mathcal{D}^g = \{D_1^g, ..., D_q^g\}$. Il s'agit des dimensions dont nous connaissons les valeurs pour l'objet à classer et sur lesquelles nous nous appuyons pour estimer la valeur de la dimension de classe.

Exemple 48 : Considérons la base contextuelle de séquences $\mathcal{CB}$ rappelée dans le tableau 5.2. L'ensemble des dimensions contextuelles de $\mathcal{CB}$ est constitué de deux dimensions *Age* et *Saison*. Dans la suite, nous considérons que *Age* est la dimension de classe et *Saison* est une dimension guide. Par conséquent, l'ensemble des dimensions contextuelles contient : une dimension de classe $D^c = Age$ et un ensemble de dimensions guides constitué d'un seul élément $\mathcal{D}^g = \{Saison\}$.

Définition 52 (Contexte guide) **:** Soit un contexte $g = [d^c, d_1^g, ..., d_k^g]$ défini sur $\{D^c\} \cup \mathcal{D}^g$. g est un *contexte guide* si :
- $d^c = *$;
- $\forall i \in \{1, ..., k\}$, $d_i^g \in dom'(D_k^g)$.

Exemple 49 : Le contexte $[*, e]$ est un contexte guide, ainsi que $[*, *]$. En effet, leur valeur sur la dimension de classe *Age* est $*$ et leur valeur sur la dimension guide *Saison* est une valeur quelconque de $dom'(Saison)$. En revanche, $[j, e]$ n'est pas un contexte guide car la valeur sur la dimension de classement *Age* est différente de $*$.

Un contexte guide a pour valeur $*$ sur la dimension de classe et une valeur quelconque sur les autres dimensions. En effet, l'intuition derrière cette définition est la suivante : un contexte guide doit représenter l'information contextuelle disponible sur un objet à classer (i.e., les valeurs sur les dimensions guides) mais aucune information sur la dimension de classe (i.e., la valeur $*$). Notons que le nombre de contextes guides possibles dans une hiérarchie de contextes peut être extrêmement grand. En effet, il correspond à tous les contextes possibles définis sur les dimensions guides de $\mathcal{D}^g$: $\bigcup_{D \in \mathcal{D}^g} |dom'(D)|$. Dans notre exemple, ce nombre est de trois car il existe une seule dimension guide (*Saison*) et son domaine enrichi contient trois valeurs ($|dom'(Saison)| = 3$).

Un contexte guide permet de cerner de manière plus spécifiques les contextes, appelés contextes candidats, dans lesquels un objet peut être classé.

Définition 53 (Contextes candidats) **:** Soit $g = [*, d_1^g, ..., d_k^g]$ un contexte guide. L'ensemble des contextes candidats pour le guide g, noté $Cand(g)$, est défini comme suit :

$$Cand(g) = \{[d^c, d_1^g, ..., d_k^g] | d^c \in dom(D_i^c)\}.$$

Exemple 50 : Pour le contexte guide $[*, e]$, l'ensemble des contextes candidats est $Cand([*, e]) = \{[j, e], [a, e]\}$. De même, l'ensemble des contextes candidats pour le guide $[*, *]$

est $Cand([*,*]) = \{[j,*],[a,*]\}$.

Un contexte candidat pour un guide donné est tel que la valeur $*$ sur la dimension de classe D^c est remplacée par une valeur de son domaine, i.e., par un des labels de classe possibles. Aussi, le nombre de contextes candidats pour un guide g donné est facilement identifiable : $|Cand(g)| = |dom(D^c)|$.

D'après la définition d'un contexte candidat pour un guide g, nous pouvons aisément remarquer que tout contexte candidat pour g est un descendant de g. Par ailleurs, l'ensemble de candidats possède les propriétés suivantes.

Propriété 3 : Soit g un contexte guide dans la hiérarchie de contextes $\mathcal{H}$. L'ensemble de ses contextes candidats vérifie $\forall c, c' \in Cand(g)$, $\mathcal{B}(c) \cap \mathcal{B}(c') = \emptyset$;

Propriété 4 : Soit g un contexte guide dans la hiérarchie de contextes $\mathcal{H}$. L'ensemble de ses contextes candidats vérifie $\bigcup_{c \in Cand(g)} \mathcal{B}(c) = \mathcal{B}(g)$.

Selon la première propriété, les contextes candidats pour le guide g sont deux à deux disjoints. En effet, si deux contextes sont candidats pour un même guide, alors ils se différencient par leur valeur sur la dimension D^c. Or, un objet ne peut être associé qu'à une seule valeur de $dom(D^c)$ et par conséquent ne peut appartenir à deux contextes candidats. La seconde propriété affirme que l'ensemble des bases d'objets des contextes candidats recouvre g. En effet, tous les objets de $\mathcal{B}(g)$ sont associés à une valeur de $dom(D^c)$. Or, chacune de ces valeurs correspond à un contexte candidat pour g. Aussi, tout objet de $\mathcal{B}(g)$ est couvert par un contexte candidat. Nous verrons plus tard comment ces remarques peuvent être exploitées dans le cadre de l'extraction de motifs contextuels pour la classification.

Suivant les définitions précédentes, le problème de classification dans les données contextuelles peut être défini comme suit.

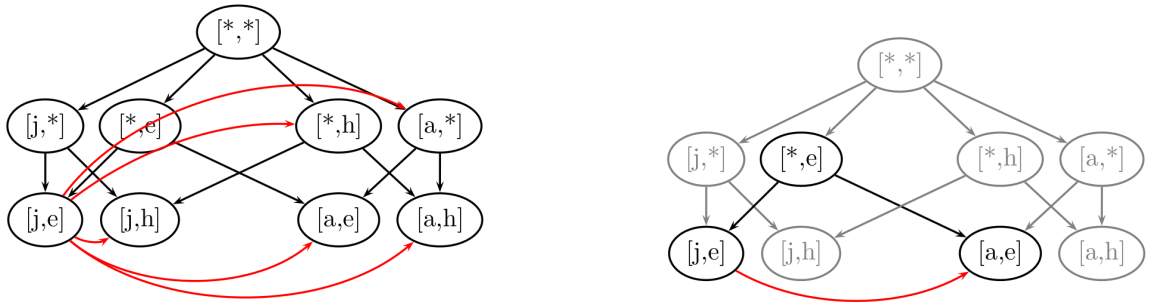
Définition 54 : Étant donné un objet o et un contexte guide g décrivant les informations contextuelles disponibles sur cet objet, le problème de classification consiste à associer à o un contexte parmi les éléments de $Cand(g)$.

Exemple 51 : Considérons le contexte guide $[*,e]$ et une séquence s à classer. Le problème de classification consiste alors à retrouver parmi les candidats $[j,e]$ et $[a,e]$ le contexte qui sera associé à s. Si le contexte guide est $[*,*]$, alors nous n'avons aucune information disponible sur l'objet à classer. Dans ce cas, il s'agira d'associer à s un contexte parmi les candidats correspondants $[j,*]$ et $[a,*]$. Remarquons par ailleurs que ce dernier cas correspond au cas classique de classification supervisée où aucune information contextuelle n'est considérée.

5.2.2 Extraction de motifs contextuels pour la classification

Afin de résoudre le problème de classification, il faut choisir parmi les contextes candidats celui qui sera associé à l'objet à classer. Pour ce faire, en adoptant le principe de la classification

basée sur les motifs, nous utilisons ceux qui différencient chaque contexte candidat par rapport aux autres. Comme nous l'avons vu dans l'exemple précédent, si la séquence à classer a été collectée en *été* (i.e., le contexte guide est $[*, e]$), alors la séquence ne peut être classée que dans les contextes candidats $[j, e]$ ou $[a, e]$. Nous avons donc besoin des motifs qui différencient ces deux contextes uniquement. Or, comme le schématise la figure 5.2(a), l'extraction de motifs contextuels associés à $[j, e]$ fournira les motifs qui différencient $[j, e]$ de $[a, e]$, mais également de tous les autres contextes disjoints de la hiérarchie : $[j, h]$, $[a, h]$, $[*, h]$ et $[a, *]$. L'extraction de motifs contextuels proposée dans le chapitre précédent n'est donc pas directement adaptée.



(a) Recherche des motifs contextuels dans $\mathcal{H}$.

(b) Recherche des motifs contextuels dans la restriction de $\mathcal{H}$.

FIGURE 5.2 – Extraction de motifs contextuels avec ou sans restriction sur la hiérarchie de contextes.

Néanmoins, comme le montre la figure 5.2(b), nous observons que lorsque nous considérons la hiérarchie $\mathcal{H}$ réduite au contexte guide $[*, e]$ et à ses descendants $[j, e]$ et $[a, e]$, les motifs contextuels associés au contexte $[j, e]$ sont les motifs recherchés, i.e., ceux qui différencient $[j, e]$ de $[a, e]$.

En exploitant cette remarque, nous adaptons dans cette section les principes de l'extraction de motifs contextuels vus dans le chapitre 4 aux besoins de la classification. Nous définissons formellement, dans un premier temps, la notion de restriction sur la hiérarchie de contextes $\mathcal{H}$ comme suit.

Définition 55 (Restriction de $\mathcal{H}$ sur un contexte) **:** La restriction de $\mathcal{H}$ sur un contexte c, notée $\mathcal{H}_c$, est la hiérarchie réduite aux composants de c (i.e., c et ses descendants).

Définition 56 (Motifs contextuels pour la classification) **:** Soit $\mathcal{G}$ l'ensemble des contextes guides dans $\mathcal{H}$. Le problème de l'extraction des motifs contextuels pour la classification consiste à extraire, pour chaque guide $g \in \mathcal{G}$, l'ensemble des motifs contextuels (c, m) dans $\mathcal{H}_g$ tels que c est un contexte candidat pour g ($c \in Cand(g)$).

Ainsi défini, le problème d'extraction des motifs contextuels pour la classification peut être résolu en suivant les étapes suivantes :

1. L'ensemble $\mathcal{G}$ des guides de $\mathcal{H}$ est construit.
2. Les motifs recherchés sont, pour chaque guide $g \in \mathcal{G}$, les motifs contextuels associés à un

candidat pour g dans $\mathcal{H}_g$.

Selon cette approche, l'extraction des motifs contextuels pour la classification peut être effectuée par l'algorithme d'extraction de motifs contextuels proposé dans le chapitre 4. Pourtant, deux caractéristiques propres au cadre de la classification basée sur les motifs contextuels peuvent être exploitées. D'abord, nous ne recherchons pas l'ensemble de tous les motifs contextuels dans $\mathcal{H}_g$, mais uniquement ceux qui sont associés à un contexte candidat pour g. Cette remarque nous permettra de mieux cibler les motifs recherchés. Ensuite, nous avons montré que deux contextes candidats pour un même guide sont nécessairement disjoints (*Cf.* propriété 3). Comme nous le montrons ci-dessous, cette remarque va dans la suite nous permettre de mieux cerner le contexte candidat dans lequel un motif peut être général.

Nous considérons un guide g ainsi que la hiérarchie de contextes $\mathcal{H}_g$ et dans cette hiérarchie, la liste de fréquences du motif m notée $\mathcal{L}_m$ (*Cf.* chapitre 4).

D'après le théorème 3 du chapitre 4, la première condition pour que m soit général dans un contexte c (i.e., (c, m) forme un motif contextuel) de $\mathcal{H}_g$ est que la liste $\mathcal{L}_m$ puisse être divisée en deux parties, telles que la partie gauche (contenant les plus hautes fréquences) corresponde à la décomposition de c.

Puisque nous avons observé que les contextes candidats sont disjoints deux à deux, nous obtenons le lemme suivant.

Lemme 5 : Soit m un motif et $\mathcal{L}_m = (c_1, ..., c_n)$ la liste de fréquences de m dans $\mathcal{H}_g$. Il existe au plus un contexte candidat $c \in Cand(g)$ tel que m est c-général (i.e., tel que (c, m) est un motif contextuel dans $\mathcal{H}_g$). De plus, ce contexte vérifie $decomp(c) = \{c_1, ..., c_i\}$, où $i = |decomp(c)|$.

Démonstration : *Nous avons remarqué plus tôt que deux contextes candidats c et c' ont des bases d'objets disjointes, i.e., $\mathcal{B}(c) \cap \mathcal{B}(c') = \emptyset$. Par conséquent, ces contextes ont également des décompositions disjointes, i.e., $decomp(c) \cap decomp(c') = \emptyset$. Or, si m est général à la fois dans c et c', alors il existe deux entiers $1 \leq i < j < n$ tels que $decompc = \{c_1, ..., c_i\}$ et $decomp(c') = \{c_1, ..., c_j\}$. Dans ce cas, $decomp(c) \subset decomp(c')$, ce qui est contradictoire avec la remarque précédente. Ainsi, il existe au plus un contexte candidat c où m est général.*
De plus, $decomp(c) = \{c_1, ..., c_i\}$, où $i = |decomp(c)|$, par simple application du théorème 3 du chapitre 4. □

Le lemme 5 montre d'abord que pour chaque motif m et chaque contexte guide, il existe au plus un contexte candidat c où m est général. Ce lemme offre également un moyen simple de découvrir ce contexte candidat. D'abord, il est le seul dont la décomposition contient le contexte minimal c_1, i.e., le premier élément de $\mathcal{L}_m$. Par ailleurs, les autres contextes candidats étant disjoints de c, leur décomposition ne peut contenir c_1. Le contexte c est donc facilement identifiable : il s'agit du seul contexte candidat dont la décomposition inclut c_1.

Cependant, le lemme 5 ne garantit pas que m soit général dans le contexte c identifié. En effet, en appliquant le théorème 3 du chapitre 4, nous constatons qu'il est également nécessaire pour cela que la valeur d'intérêt du motif soit supérieure au seuil minimum σ fixé :

$$M(m, c_i, c_{i+1}) \geq \sigma, \text{ où } i = |decomp(c)|.$$

Nous avons dorénavant à notre disposition toutes les informations nécessaires à l'extraction des motifs contextuels pour chaque guide donné.

Algorithm 9 `CoPaC` : Contextual Patterns for Classification

Entrées: une base contextuelle de séquences $\mathcal{CB}$, une mesure d'intérêt M, un seuil minimum d'intérêt σ, un seuil minimum de fréquence ρ, une hiérarchie de contextes $\mathcal{H}$.

`/* Extraction des motifs fréquents requis */`
Extraire les couples $(m, \mathcal{L}_m)$ tels que m est fréquent dans au moins un contexte minimal

`/* Construction des guides dans` $\mathcal{H}$ `*/`
$\mathcal{G} \leftarrow$ ensemble des contextes guides de $\mathcal{H}$

`/* Génération des motifs contextuels */`
pour tout $(m, \mathcal{L}_m)$ **faire**
  pour tout $g \in \mathcal{G}$ **faire**
    construire $\mathcal{L}'_m = (c_1, ..., c_n)$ la liste de fréquences de m dans $\mathcal{H}_g$
    `/* On identifie l'unique contexte candidat où` m `peut être général */`
    $c \leftarrow$ le contexte dans $Cand(g)$ tel que $c_1 \subseteq decomp(c)$
    $i \leftarrow |decomp(c)|$
    `/* On teste si` m `est général dans` c `*/`
    si $\{c_1, ..., c_i\} = decomp(c)$ et $M(m, \mathcal{B}(c_i), \mathcal{B}(c_{i+1})) \geq \sigma$ **alors**
      générer et afficher le motif contextuel (c, m)
    finsi
  fin pour
fin pour

L'algorithme 9, appelé `CoPaC`, présente la méthodologie générale employée pour extraire ces motifs. D'abord, les motifs fréquents dans au moins un contexte minimal de $\mathcal{H}$ sont extraits et associés à leur liste de fréquence. Cette phase est identique à celle utilisée dans l'algorithme `CoPaM` (*Cf.* chapitre 4). Puis l'ensemble de tous les contextes guides de la hiérarchie $\mathcal{H}$ est construit et stocké dans la variable $\mathcal{G}$.

Par la suite, pour chaque motif m extrait, l'algorithme génère tous les motifs contextuels recherchés générés par m, en suivant les étapes suivantes :

1. Pour chaque contexte guide $g \in \mathcal{G}$, on construit la liste de fréquences $\mathcal{L}'_m$ contenant uniquement les contextes minimaux de $\mathcal{H}_g$.
2. L'algorithme identifie ensuite l'unique contexte candidat c dans lequel m peut être général en exploitant le lemme 5 : c'est le seul contexte candidat pour g tel que sa décomposition contient le premier élément de c_1.
3. Enfin, si m est c-général, alors le motif contextuel (c, m) est construit et affiché. Le test de c-généralité est effectué au travers du théorème 3 du chapitre 4 : la partie gauche de $\mathcal{L}'_m$ (de taille $|decomp(c)|$) doit être égale à $decomp(c)$ et la valeur de la mesure d'intérêt entre le dernier élément de la partie gauche de $\mathcal{L}'_m$ et le premier élément de sa partie gauche doit dépasser le seuil minimum σ fixé.

5.2.3 Classification basée sur les motifs contextuels

Dans la section précédente, nous avons décrit l'extraction des motifs contextuels pour la classification. Cette approche est applicable pour toute mesure d'intérêt choisie parmi celles vérifiant les propriétés montrées dans le chapitre 4 (émergence, gain d'information, lift, fréquence).

Dans cette section, nous nous concentrons sur l'exploitation de ces motifs contextuels lors de la phase de classification. La méthode employée étant dépendante de la mesure d'intérêt choisie, nous nous focalisons dorénavant sur la mesure d'émergence, i.e., sur les motifs contextuels extraits pour leur mesure d'émergence (i.e., les motifs contextuels émergents). Comme nous l'avons expliqué dans la section 5.1, la raison de ce choix relève principalement de l'insensibilité de cette mesure dans le cas où les données sont inégalement réparties dans les contextes. En effet, la mesure d'émergence ne repose que sur la notion de fréquence et ne dépend pas de la taille des contextes considérés. Notons toutefois que d'autres mesures, également basées sur la fréquence, possèdent cette caractéristique et pourraient par conséquent être utilisées pour effectuer la classification. En particulier, c'est le cas des mesures qui vérifient les propriétés isolées dans le chapitre 4 (Emergence, Gain d'information, Lift, Fréquence).

Comme nous l'avons vu dans la section 5.1, plusieurs méthodes de classification basées sur les motifs émergents existent dans la littérature. Nous adaptons ci-dessous aux motifs contextuels émergents l'approche `CAEP` initialement proposée dans [DZWL99].

Considérons un nouvel objet à classer o et un contexte guide g dans la hiérarchie de contextes. Nous notons $Cand(g) = \{c_1, ..., c_n\}$ l'ensemble des contextes candidats pour g, i.e., l'ensemble des contextes où o peut être classé.

Le principe général de `CAEP` repose sur le fait que pour chaque contexte c_i avec $i \in \{1, ..., n\}$, l'objet o est couvert par un certain nombre de motifs émergents dans c_i. `CAEP` propose que chacun de ces motifs apporte sa contribution à la décision finale, i.e., o doit-il ou non être classé dans c_i ? En effet, l'ensemble de ces motifs couvre un nombre plus grand de cas qu'un motif émergent seul. Trois questions se posent alors :

1. *Comment quantifier la contribution individuelle d'un motif émergent ?*
2. *Comment combiner les contributions de chaque motif émergent dans la classe c_i ?*
3. *Comment, en fonction de la contribution mesurée dans chaque classe, associer une classe à l'objet ?*

La première réponse apportée par `CAEP` tient compte à la fois du taux d'émergence du motif et de sa fréquence. Dans la suite, nous adaptons les notions de CAEP au cas des motifs contextuels extraits dans la section précédente.

Définition 57 (Contribution de α) **:** Soit $\alpha = (c, m)$ un motif contextuel émergent. La contribution de α pour la classe c, notée $Score(\alpha)$, est définie comme :

$$Score(\alpha) = \frac{Em(\alpha)}{Em(\alpha)+1} \times Freq(\alpha)$$

Remarquons tout d'abord que la définition de la contribution d'un motif repose sur deux termes : $\frac{Em(\alpha)}{Em(\alpha)+1}$ et $Freq(\alpha)$. Le premier s'apparente à la probabilité conditionnelle qu'un objet

o soit dans la classe c sachant que o supporte m, tandis que le second quantifie la fraction de c recouverte par m (i.e., la fraction d'objets dans c qui supportent m).

La deuxième question posée concerne l'agrégation des contributions individuelles de chaque motif pour la classe c. `CAEP` propose d'effectuer une somme des contributions individuelles.

Définition 58 (Score agrégé) **:** Soit o un objet à classer, c un contexte candidat et $\mathcal{M}'_c$ l'ensemble des motifs contextuels dans c (i.e., de la forme (c, m)) supportés par o. Le score agrégé de l'objet o pour la classe c, noté $Score(o, c)$, est défini comme suit :

$$Score(o, c) = \sum_{\alpha \in \mathcal{M}'_c} Score(\alpha)$$

Le score agrégé de l'objet o étant évalué pour chaque classe, la troisième question se rapporte au choix de la classe à attribuer à o. Une première idée pourrait consister à assigner à o la classe dans laquelle il a le score agrégé le plus élevé. Cependant, le nombre de motifs émergents peut varier fortement d'une classe à l'autre et une telle stratégie pourrait favoriser les classes possédant plus de motifs contextuels. Afin d'éliminer ce problème, [DZWL99] propose une normalisation des scores agrégés. Pour chaque classe, un *score de base* est défini en s'appuyant sur les objets de la base d'apprentissage.

Définition 59 (Score de base) **:** Le *score de base* du contexte candidat c, noté $BScore(c)$, est défini comme le score de l'objet x de $\mathcal{B}(c)$ tel que 50% des objets de $\mathcal{B}(c)$ ont un score supérieur ou égal à $Score(x, c)$.

Le score de base dans un contexte candidat c est donc défini comme le score médian parmi les scores de chaque objet de $\mathcal{B}(c)$. Nous pouvons désormais définir le score normalisé d'un objet dans un contexte candidat.

Définition 60 (Score normalisé) **:** Le *score normalisé* d'un objet o dans le contexte candidat c, noté $NScore(o, c)$, est défini comme suit :

$$NScore(o, c) = \frac{Score(o,c)}{BScore(c)}$$

Le processus de classification s'appuie ensuite sur le score normalisé [1] d'un objet o dans chacun des contextes candidats pour choisir la classe qui sera associée à o. Le contexte candidat associé à o à l'issue du processus de classification est celui qui obtient le score normalisé le plus élevé.

L'adaptation de `CAEP` ainsi définie permet d'intégrer les motifs contextuels émergents dans le processus de classification. Nous montrons de plus dans la section suivante que cette méthode, dans le cadre des données séquentielles, peut facilement être traduite en un problème de prédiction.

1. Notons que, de l'aveu même des auteurs, l'usage du terme « *normalisé* » constitue un abus puisque les scores normalisés peuvent être supérieurs à 1.

5.3 Vers un cas particulier : la prédiction

Les contributions présentées dans ce mémoire s'appliquent au cadre général de la découverte de motifs dans une base d'objets. Nous portons cependant une attention particulière aux applications impliquant des données séquentielles. En appliquant l'approche de classification proposée dans ce chapitre à de telles données, nous constatons que le problème de classification abordé peut facilement s'apparenter à un problème de prédiction.

5.3.1 Présentation du problème

L'exemple suivant présente de manière intuitive le problème de prédiction traité. Considérons la séquence étendue s suivante :

$$s = \langle (ab\ E_{haut})^1 (bc\ E_{bas})^2 (e\ E_{haut})^3 (ad\ E_{haut})^4 (cd\ E_{bas})^5 (ab\ E_{haut})^6 \rangle$$

Ici, les items a, b, c, d, e sont les activités suivies par l'habitant et les items E_{bas} et E_{haut} représentent la consommation électrique dans l'appartement associé. La question qui nous intéresse est alors la suivante : « *Étant donné les informations contenues dans s aux temps $t = 6, t-1, t-2, ...$, quelle sera la valeur de la consommation électrique au temps $t+1$?* ».

Dans la suite de cette section, nous cherchons à résoudre ce problème de prédiction à l'aide de motifs inter-transactionnels (IT) extraits dans la séquence. La prédiction basée sur les motifs ou les règles d'association inter-transactionnels dans une séquence étendue a déjà été étudiée dans la littérature [LFH00, TLHF03, FDL01, BAV04, DJ05, NR06]. En effet, l'information portée par ces motifs s'avère particulièrement adaptée dans ce cas puisqu'elle peuvent nous permettre d'extraire des règles inter-transactionnelles de la forme : « *Lorsque les activités a et b se produisent au temps t, la valeur de la consommation électrique est élevée au temps $t+1$ avec une probabilité de 80%* ». En exploitant une telle règle, un utilisateur peut prédire une valeur élevée pour l'heure suivante aussitôt que les items a et b sont enregistrés.

Étonnement, bien que de nombreux travaux aient abordé la problématique de l'extraction d'itemsets inter-transactionnels et souligné leur utilité pour la prédiction [LFH00, TLHF03, LW07, WC11], peu d'efforts ont été consacrés au problème de prédiction en lui-même. Aussi, les méthodes de prédiction décrites dans la littérature se limitent souvent au principe suivant [FDL01, BAV04, DJ05] :

- Les règles d'association inter-transactionnelles de la forme $X \rightarrow Y$ sont extraites pour un support et une confiance minimums fixés par l'utilisateur.
- Dans une séquence étendue s, dès lors que l'antécédent X d'une règle extraite est rencontré, alors le conséquent Y de la même règle est prédit.

Cette approche n'est pas adaptée dans notre cas, principalement parce qu'elle ne prend pas en compte le fait que dans l'ensemble de règles extrait, certaines règles peuvent être contradictoires. En effet, appliqué à notre exemple, s'il existe deux règles $X \rightarrow E_{bas}$ et $X' \rightarrow E_{haut}$ qui s'appliquent toutes les deux, une telle approche prédira à la fois l'item E_{bas} et E_{haut}. Or ces deux items sont incompatibles dans notre problème : la consommation électrique ne peut avoir deux valeurs simultanément.

Dans [NR06] néanmoins, le problème posé par les auteurs se rapproche du nôtre : il s'agit de prédire la localisation d'un utilisateur mobile au temps $t+1$. Comme dans le cas qui nous

intéressent, les différentes localisations sont donc également incompatibles entre elles. L'approche proposée consiste à sélectionner toutes les règles qui s'appliquent. Dès lors, pour chacune de ces règles R sélectionnées, la somme Σ_R de sa fréquence et de sa confiance est calculée ($\Sigma_R = Freq_{\mathcal{B}}(R)+conf_{\mathcal{B}}(R)$). La règle ayant la plus haute valeur de Σ est alors choisie et son conséquent est prédit. Cette approche constitue en réalité un cas particulier de la classification associative (*Cf.* section 5.1).

Une telle approche est limitée car elle subit les mêmes inconvénients que la classification associative basée sur le couple de mesures fréquence/confiance. Ainsi, les règles extraites favorisent les événements fréquents au détriment d'événements plus rares pour lesquelles il est difficile de trouver des règles.

Comme nous venons de le voir et bien que la littérature sur le sujet n'en fasse pas état, le problème de prédiction défini dans cette section et le problème de classification basée sur les motifs sont intrinsèquement proches. Nous décrivons par conséquent dans la suite de cette section comment les méthodes présentées jusqu'ici peuvent être exploitées pour résoudre le problème de prédiction dans les données séquentielles.

Le principe général sur lequel repose notre approche est le suivant : prédire une valeur de consommation électrique revient à associer à la séquence un label de classe E_{bas} ou E_{haut}. Le problème général peut alors être considéré comme suit : si ce qui se produit avant l'occurrence de E_{bas} et ce qui se produit avant l'occurrence de E_{haut} forment deux contextes différents, alors nous devons dans un premier temps extraire les motifs contextuels discriminants (par exemple, les motifs contextuels émergents) de chacun de ces contextes. Les questions suivantes se posent alors :

- *Comment construire la base d'apprentissage ?* Afin de résoudre le problème de classification, il est nécessaire de construire la base d'itemsets IT contextuelle correspondante.
- *Quels motifs extraire ?* Bien que les itemsets IT semblent porter toute l'information nécessaire pour répondre à notre problème de prédiction, leur définition initiale n'est pas parfaitement adaptée au problème que nous souhaitons résoudre. Nous devrons par conséquent apporter quelques modifications au formalisme présenté dans le chapitre 2.

5.3.2 Motifs inter-transactionnels pour la prédiction

Dans la base contextuelle d'itemsets IT, chaque itemset IT doit décrire ce qui s'est produit avant l'occurrence d'un item E_{bas} ou E_{haut}. Dans la séquence s rappelée ci-dessous, considérons par exemple l'item E_{bas} apparu dans l'itemset d'estampille $t = 5$ (en gras dans la séquence ci-dessous).

$$s = \langle (ab\ E_{haut})^1 (bc\ E_{bas})^2 \overbrace{(e\ E_{haut})^3 (ad\ E_{haut})^4}^{fenêtre\ précédente} (cd\ \mathbf{E_{bas}})^5 (ab\ E_{haut})^6 \rangle$$

Afin de décrire la fenêtre sur s (de taille $maxSpan = 2$) précédant l'item considéré, nous avons besoin d'une première modification dans le formalisme des itemsets IT. En effet, le formalisme initial décrit des fenêtres qui commencent à un point de référence donné. Dans notre cas, nous avons à l'inverse besoin de construire la fenêtre qui précède un point de référence donné (la

fenêtre précédant l'estampille 5 dans notre exemple). Nous définissons donc la notion de motif IT précédant une estampille.

Définition 61 (Itemset IT précédant l'estampille d_k) **:** Soient $s = \langle I_1^{d_1} I_2^{d_2} ... I_n^{d_n} \rangle$ une séquence étendue, d_k une estampille sur s appelée estampille de référence et $maxSpan$ un seuil d'écart maximal. L'itemset IT précédant d_k sur s est l'ensemble des couples $(i, p - d_k)$ tels que $i \in I_p$ et $d_k - maxSpan \leq p < d_k$.

Exemple 52 : Considérons la séquence s et un seuil d'écart maximal $maxSpan = 2$. L'itemset inter-transactionnel I précédant l'estampille 5 dans s est :

$$I = \{(e, -2)(E_{haut}, -2)(a, -1)(d, -1)(E_{haut}, -1)\}.$$

Si les estampilles correspondent à des heures, cet itemset IT peut être interprété comme suit : « *Les items e et* E_{haut} *apparaissent deux heures avant l'estampille* 5*, puis les items a, d et* E_{haut} *une heure avant* ».

Cette nouvelle définition marque une différence fondamentale entre les motifs décrits dans le formalisme original et celui que nous utilisons. Alors que l'estampille de référence est à l'origine inclus dans l'itemset IT construit, nous souhaitons dans notre cas qu'il soit en dehors de cet itemset IT. Ainsi, nous ne pouvons dans notre cas trouver d'item de la forme $(i, 0)$.

En construisant les itemsets IT précédant chacune des estampilles sur s, nous obtenons la base contextuelle d'itemsets IT présentée dans le tableau 5.3. Dans ce tableau, chaque itemset IT est associé à un item E_{bas} ou E_{haut}. Par conséquent, le premier élément de cette base peut être interprété comme suit : « *Une heure avant une consommation d'électricité basse, les activités a et b ont été réalisées alors que la consommation électrique était élevée* ».

Nous avons à présent la base contextuelle dans laquelle les motifs peuvent être extraits. Cependant, nous pouvons nous interroger sur la nature des motifs à extraire. Les itemsets IT initialement définis imposent qu'un itemset IT extrait contienne au moins un item de la forme $(i, 0)$, où 0 correspond à l'estampille de référence, afin de limiter la redondance dans les motifs extraits. Dans notre cas, cette contrainte n'est pas pertinente pour les deux raisons suivantes :

- Tout d'abord, il n'existe pas de redondance dans les itemsets IT nouvellement définis. En effet, les motifs $\{(e, -2)\}$ et $\{(e, -1)\}$ ont dans notre cas une signification très différente puisqu'ils décrivent un écart différent par rapport à l'estampille de référence fixée.
- De plus, l'estampille de référence dans les itemsets IT que nous avons définis se trouve en dehors de la fenêtre considérée. Il n'existe donc pas d'item de la forme $(i, 0)$.

La découverte des motifs prédictifs recherchés consiste alors à extraire les itemsets IT contextuels émergents pour chaque contexte $[E_{bas}]$ et $[E_{haut}]$.

Exemple 53 : Considérons l'itemset IT $I = \{(a, -1)(E_{haut}, -1)\}$ et la base contextuelle présentée dans le tableau 5.3. La fréquence de I est $Freq_{[E_{bas}]}(I) = 1$ dans le contexte $[E_{bas}]$ et $Freq_{[E_{haut}]}(I) = 0$ dans le contexte $[E_{haut}]$. Par conséquent, I est discriminante d'une valeur d'électricité basse selon la mesure d'émergence :

$$Em(I, \mathcal{B}([E_{bas}]), \mathcal{B}([E_{haut}])) = +\infty.$$

IIT	Classe
$\{(a,-1)(b,-1)(E_{haut},-1)\}$	E_{bas}
$\{(a,-2)(b,-2)(E_{haut},-2)(b,-1)(c,-1)(E_{bas},-1)\}$	E_{haut}
$\{(b,-2)(c,-2)(E_{bas},-2)(e,-1)(E_{haut},-1)\}$	E_{haut}
$\{(e,-2)(E_{haut},-2)(a,-1)(d,-1)(E_{haut},-1)\}$	E_{bas}
$\{(a,-2)(d,-2)(E_{haut},-2)(c,-1)(d,-1)(E_{bas},-1)\}$	E_{haut}

TABLE 5.3 – Base d'itemsets inter-transactionnels labellisés sur s.

De même, considérons l'itemset IT $I' = \{(a,-2)(E_{haut},-2)(c,-1)\}$. La fréquence de I' est $Freq_{[E_{bas}]}(I') = 0$ dans le contexte $[E_{bas}]$ et $Freq_{[E_{haut}]}(I') = \frac{2}{3}$ dans le contexte $[E_{haut}]$. Par conséquent, I' est en revanche discriminant d'une valeur élevée :

$$Em(I', \mathcal{B}([E_{haut}]), \mathcal{B}([E_{bas}])) = +\infty.$$

Par la suite, le processus de classification associé aux motifs contextuels extraits est identique à celui décrit dans la section 5.2.3. Si l'on se réfère uniquement aux deux motifs extraits dans l'exemple précédent, le processus de classification pour la séquence s est le suivant. Afin de prédire la valeur de la consommation électrique à l'estampille 7, on construit d'abord l'itemset IT sur s pour l'estampille de référence 7 :

$$\{(c,-2)(d,-2)(E_{bas},-2)(a,-1)(b,-1)(E_{haut},-1)\}.$$

Nous notons que le motif I associé au contexte E_{bas} est inclus dans cet itemset IT tandis que I' associé au contexte E_{haut} ne l'est pas. Par conséquent, sans détailler les calculs de scores pour chaque contexte, nous savons que le processus prédira une valeur basse d'électricité (i.e., l'itemset IT est classé dans le contexte $[E_{bas}]$).

Intégration de dimensions contextuelles L'exemple que nous avons considéré pour la prédiction ne considère que la dimension *Classe*. Cependant, en suivant les principes de la classification basée sur les motifs contextuels présentée dans ce chapitre, rien n'interdit d'intégrer d'autres dimensions qui seront alors des dimensions guides. De telles dimensions pourront correspondre au jour de la semaine, à la saison, etc. Le tableau 5.4 présente un exemple d'une base contextuelle d'itemsets intégrant ce type d'informations. Ainsi, le premier élément de cette base signifie : « *L'itemset* $\{(a,-1)(b,-1)(E_{haut},-1)\}$ *précède une valeur basse de consommation électrique mesurée un lundi en été* ».

5.4 Expérimentations

Nous évaluons dans cette section l'approche de classification proposée dans ce manuscrit pour les différents jeux de données *Amazon*, *Puces à ADN* et *Consommation énergétique*. La classification basée sur les motifs contextuels nécessite dans un premier temps de définir, parmi les dimensions contextuelles disponibles, la dimension de classe et l'ensemble des dimensions guides (*Cf.* définition 51). C'est ce que nous faisons dans la sous-section suivante pour chacun

IIT	Classe	Saison	Jour
$\{(a,-1)(b,-1)(E_{haut},-1)\}$	E_{bas}	été	lundi
$\{(a,-2)(b,-2)(E_{haut},-2)(b,-1)(c,-1)(E_{bas},-1)\}$	E_{haut}	été	lundi
$\{(b,-2)(c,-2)(E_{bas},-2)(e,-1)(E_{haut},-1)\}$	E_{haut}	hiver	mercredi
$\{(e,-2)(E_{haut},-2)(a,-1)(d,-1)(E_{haut},-1)\}$	E_{bas}	hiver	mercredi
$\{(a,-2)(d,-2)(E_{haut},-2)(c,-1)(d,-1)(E_{bas},-1)\}$	E_{haut}	hiver	jeudi

TABLE 5.4 – Base contextuelle d'itemsets IT.

des jeux de données avant de présenter une évaluation de la classification en termes de rappel et précision.

5.4.1 Partitionnement des dimensions

Amazon Dans les commentaires *Amazon*, nous utilisons les motifs contextuels pour retrouver la note associée à un commentaire (*Good*, *Bad* ou *Neutral*). Par conséquent, le partitionnement des dimensions contextuelles sera le suivant :

- La dimension de classe est la dimension *Note* (i.e., $D^c = Note$) ;
- L'ensemble des dimensions guides est composé des dimensions contextuelles restantes : *Produit* et *Retour* (i.e., $\mathcal{D}^g = \{Produit, Retour\}$).

Puces à ADN L'étude des puces à ADN vise à découvrir le grade de cancer associé à une puce à ADN (i.e., à une tumeur). Le partitionnement des dimensions contextuelles est donc réalisé comme suit :

- La dimension de classe est la dimension *Grade*, (i.e., $D^c = Grade$) ;
- L'ensemble des dimensions guides est un singleton composé de la dimension *Age* (i.e., $\mathcal{D}^g = \{Age\}$).

Consommation énergétique Le jeu de données décrivant la consommation énergétique est un cas particulier car, comme nous l'avons vu dans la section 5.3, l'approche de classification peut être utilisée dans ce cas pour effectuer une tâche de prédiction. Par conséquent, le partitionnement des dimensions contextuelles est réalisé comme suit :

- La dimension de classe est représentée par la valeur de la consommation électrique au temps à prédire que nous notons *Électricité*. Ainsi, $D^c = Électricité$;
- Les dimensions guides correspondent aux dimensions contextuelles $\mathcal{D}^g = \{Jour, Heure\}$. Notons que la dimension contextuelle *Mois* a été supprimée, de manière à conserver des contextes minimaux contenant suffisamment d'objets pour en extraire des motifs.

5.4.2 Résultats expérimentaux

Les expérimentations menées ont pour objectif principal de montrer l'apport de la prise en compte du contexte par le biais des motifs contextuels.

Les tableaux 5.5, 5.6 et 5.7 montrent, pour chacun des jeux de données, les résultats de la classification basée sur les motifs contextuels dans différents contextes plus ou moins généraux.

Notons que ces résultats ont été obtenus en utilisant une validation croisée en considérant 90% de chaque jeu de données pour l'extraction des motifs contextuels et 10% pour effectuer les tests. Les résultats présentés concernent la moyenne obtenue par validation croisée. Nous constatons que les résultats sont systématiquement meilleurs ou égaux, en rappel comme en précision, lorsque le contexte est pris en compte.

Par exemple, observons le cas des données *Amazon*, où sans information contextuelle (i.e., avec le contexte guide $[*, *, *]$) la précision et le rappel sont respectivement 0.67 et 0.58. Intégrer la connaissance selon laquelle le produit commenté est un livre (i.e., le contexte guide est $[*, Books, *]$) permet d'améliorer significativement la qualité du classifieur : la précision et le rappel sont alors respectivement 0.77 et 0.83. En particulier, cette constatation s'applique sur le problème de prédiction dans une séquence étendue que nous avons présenté dans la section 5.3. En effet, la consommation d'électricité est fortement dépendante du contexte. Il existe par exemple une différence importante entre la consommation électrique un jour de semaine (*lundi, mardi, ..., vendredi*) et un jour de week-end (*samedi, dimanche*).

Contexte guide	**Précision**	**Rappel**
$[*, *, *]$	0.67	0.58
$[*, Books, *]$	0.77	0.83
$[*, Books, 75-100]$	0.80	0.88

TABLE 5.5 – Résultats en classification sur le jeu de données *Amazon*.

Contexte guide	**Précision**	**Rappel**
$[*, *]$	0.59	0.62
$[*, 50-62]$	0.93	0.86

TABLE 5.6 – Résultats en classification sur le jeu de données *Puces à ADN*.

Contexte guide	**Précision**	**Rappel**
$[*, *, *]$	0.42	0.64
$[*, lundi, *]$	0.72	0.84
$[*, lundi, 14h-20h]$	0.85	0.89

TABLE 5.7 – Résultats en prédiction sur le jeu de données *Consommation énergétique*.

5.5 Discussion

Dans ce chapitre, nous avons proposé une approche originale de classification supervisée basée sur les motifs contextuels ayant de nombreuses applications pratiques pour divers domaines d'activité. La section 5.1 a donné un aperçu du problème général de la classification basée sur les

motifs. Après avoir défini un nouveau problème de classification basé sur les motifs contextuels et mis en lumière ses spécificités, nous avons proposé une méthode et les algorithmes nécessaires dans la section 5.2 qui regroupe deux contributions importantes. D'une part, nous avons abordé le problème de l'extraction des motifs contextuels significatifs pour la classification au travers de l'algorithme `CoPaC`, en adaptant les notions liées aux motifs contextuels présentée dans le chapitre 4. D'autre part, nous avons traité la question de la classification en elle-même en ajustant une méthode de classification existante basée sur les motifs émergents.

Un aspect particulièrement intéressant abordé dans la section 5.3 concerne les données séquentielles. Nous avons en effet montré que, dans une certaine mesure, la classification basée sur les motifs inter-transactionnels dans les données séquentielles peut être traduite comme un problème de prédiction. En ajustant le formalisme lié à l'extraction de motifs IT dans une séquence étendue, nous avons proposé une méthode de prédiction basée sur les motifs IT contextuels qui permet là aussi de tenir compte d'informations contextuelles.

Les expérimentations décrites dans la section 5.4 effectuées sur divers jeux de données réelles ont souligné l'intérêt de l'approche proposée. Les travaux présentés dans ce chapitre peuvent cependant être approfondis suivant différents axes de recherche. Par exemple, l'algorithme `CoPaC` de découverte de motifs contextuels pour la classification ne considère qu'un seul critère : l'intérêt des motifs extraits selon la mesure choisie. Cependant, de nombreux travaux mettent en avant d'autres critères importants tels que la concision et l'absence de redondance ou encore la couverture des données de l'ensemble de motifs extraits. Une piste intéressante d'évolution pour l'algorithme `CoPaC` consiste à tenir compte de ces différents critères de manière à extraire directement un ensemble de motifs adéquat pour la classification.

Comme nous avons montré dans ce chapitre l'intérêt des motifs contextuels pour la classification, nous allons montrer dans le chapitre suivant qu'ils sont tout aussi efficaces pour résoudre des problèmes de détection d'anomalies.

Chapitre 6

Détection d'anomalies et motifs contextuels

Sommaire

Introduction

La détection d'anomalies consiste, de manière intuitive, en la découverte de fragments de données qui ne sont pas conformes au comportement attendu. Issue de différents domaines, on la retrouve sous différents noms aussi variés que la détection d'anomalies, d'outliers, d'intrusions, d'aberrations, d'exceptions, de nouveautés, etc. Une telle diversité s'explique par le champ extrêmement large d'applications où ces phénomènes se traduisent en connaissances précieuses.

Considérons, dans notre cas d'étude, le problème de la détection d'anomalies dans la consommation électrique, i.e., la découverte de périodes sur lesquelles la consommation électrique dans un logement est anormale ou inattendue. Une détection automatisée et précoce d'une telle anomalie peut mettre en évidence le dysfonctionnement de l'équipement installé et prévenir des problèmes plus sévères (panne, incendie, gaspillage d'énergie, etc.).

Il s'agit d'un problème difficile notamment lié à l'aspect conditionnel des anomalies. En effet, reconnaître une anomalie dans ces données ne consiste pas simplement à détecter une valeur anormale dans une séquence, mais à détecter une valeur qui est anormale en fonction des circonstances où elle apparaît. En effet, une valeur anormale de consommation électrique dépend de divers facteurs. Tout d'abord, les valeurs suivantes et précédentes de la consommation électrique jouent un rôle l'interprétation de la consommation. Par exemple, un pic brutal de la consommation, observé alors que celle-ci était jusqu'ici modérée peut être interprété comme une anomalie. Notre cas d'étude comporte également d'autres informations ayant un effet sur la consommation électrique : la température extérieure ou les activités réalisées par les habitants. Ces événements, qui ne décrivent pas la consommation électrique, jouent cependant un rôle dans la valeur de celle-ci. Enfin, la consommation électrique peut de même être liée à des informations contextuelles telles que celles que nous avons manipulées dans les chapitres précédents (jour de la semaine, saison, etc.).

Ainsi, détecter une valeur anormale de consommation électrique nécessite d'intégrer les corrélations qui existent sur des informations de formes diverses et provenant de sources multiples. Nombre d'approches classiques de détection d'anomalies n'intègrent pas cette complexité. La découverte de motifs fréquents dans les données séquentielles traite en revanche l'ensemble de ces données pour y découvrir les corrélations existantes. De plus, les motifs fréquents contextuels présentés dans le chapitre 3 offrent la possibilité de tenir compte des informations contextuelles associées à une valeur de consommation énergétique. Le problème abordé revient donc à traiter une séquence étendue telle que :

$$s = \langle (ab\ E_{moy})^1 (bc\ E_{bas})^2 (e\ E_{haut})^3 (ad\ \mathbf{E_{haut}})^{\mathbf{4}} (cd\ E_{bas})^5 (ab\ E_{haut})^6 \rangle.$$

Dans cette séquence, les items $a, b, ..., e$ sont des activités et les items de la forme E_{haut}, E_{moy} ou E_{bas} décrivent l'état de la consommation électrique (respectivement haute, moyenne ou basse) pour chaque itemset de s. Ces derniers items se rapportent ici à des valeurs initialement numériques qui ont été discrétisées. Dans ce chapitre, la question que nous posons est « *La valeur de la consommation électrique à l'estampille* 4 *est-elle normale ?* ». En d'autres termes, il s'agit de décider si la présence de l'item E_{haut} est anormale dans l'itemset d'estampille 4. Afin de répondre à cette question, nous nous appuyons sur les motifs fréquents qui peuvent être

extraits d'un ensemble de données historisées. Le principe général de l'approche présentée dans ce chapitre repose sur les étapes suivantes :

1. L'extraction de motifs IT fréquents permet d'obtenir des informations sur les corrélations qui existent entre les différents items représentant la consommation électrique, les activités ou d'autres informations. De plus, si des informations contextuelles sont disponibles (jour de la semaine, saison, etc.), l'extraction d'itemsets IT fréquents contextuels permet de les intégrer.
2. Afin de décider si la présence d'un item est anormale ou non, nous sélectionnons deux types de motifs parmi ceux extraits. Les *motifs concordants* décrivent des circonstances similaires à celles observées. Ces motifs, intuitivement, tendent à valider l'item observé. À l'inverse, les *motifs discordants* décrivent des circonstances similaires à la situation observée, mais pour une valeur différente de la consommation électrique. Ces motifs soulignent qu'étant donné les circonstances observées, la valeur de consommation électrique n'est pas celle attendue.
3. Les motifs concordants et discordants sont utilisés pour fournir une mesure agrégée de la conformité de l'item avec l'ensemble des motifs extraits et ainsi déclarer l'item comme une anomalie si la valeur de cette mesure est trop basse.

L'approche générale proposée se rapproche donc, dans une certaine mesure, de l'approche de classification étudiée dans le chapitre précédent. Elle s'appuie sur les motifs extraits (qui ici sont des motifs fréquents) pour calculer une contribution agrégée de l'ensemble des motifs.

La suite de cette section s'organise de la manière suivante. Nous présentons dans la section 6.1 un aperçu des grandes familles de méthodes employées pour la détection d'anomalies en nous intéressant plus particulièrement aux données séquentielles. La section 6.2 développe l'approche que nous utilisons pour effectuer la détection d'anomalies à l'aide des motifs contextuels fréquents extraits. Nous présentons dans la section 6.3 les expérimentations menées sur un jeu de données réelles enrichi d'anomalies simulées. Enfin, nous discutons dans la section 6.4 l'approche proposée ainsi que les différentes pistes d'améliorations possibles.

6.1 Détection d'anomalies

La détection d'anomalies, étudiée d'un point de vue statistique depuis le $19^{ème}$ siècle [Edg87], s'est depuis étendue à des domaines d'application diversifiés. Un éventail considérable de méthodes, souvent développées de manière indépendante pour répondre à des problèmes spécifiques, a ainsi vu le jour et en faire une synthèse n'est pas une tâche aisée. Nous isolons cependant trois axes principaux sur lesquels les méthodes définies se différencient : le type des anomalies recherchées, le cadre d'apprentissage et les techniques employées pour résoudre le problème posé.

6.1.1 Type d'anomalies

Une première différence parmi les approches existantes concerne les domaines d'application auxquels sont liés la notion même d'anomalie. [CBK09a] met, par exemple, en avant le cas du domaine médical, où une légère fluctuation de la température corporelle pourra être considérée

comme une anomalie, tandis qu'une fluctuation similaire ou plus marquée sera considérée comme normale dans l'étude des marchés financiers. Pourtant, ces exemples se rapportent à des données similaires dans leur forme : des séries temporelles numériques.

Le domaine d'application visé influence également le type de données manipulées qui varie d'un ensemble de points à des structures plus complexes telles que des séquences [SCA06] ou des graphes [NC03], par exemple.

Revenons aux données qui nous intéressent dans ce manuscrit, les données séquentielles, utilisées pour diverses applications : données biologiques [LMFK99, SCA06], données de capteurs pour la surveillance d'équipements [FYM05, Sri05, BSO09], séquences de commandes utilisateurs dans un système informatique [HFS98, ESL01], etc. Là encore, plusieurs définitions d'une anomalie ont vu le jour. Ainsi, [CBK09b] dégage trois formulations distinctes du problème de détection d'anomalies dans de telles données :

- Détecter les séquences anormales parmi un ensemble de séquences donné [SCA06, BSO09]. Appliqué à notre cas d'étude, cette formulation concernera par exemple la détection d'un habitant dont le comportement est différent de celui des autres habitants de l'immeuble.
- Isoler, dans une unique longue séquence, une sous-séquence jugée anormale [KLLH07, BLF+07]. Il s'agira, par exemple, dans la séquence d'activités d'un habitant enregistrée sur une année de détecter les périodes pendant lesquelles un phénomène inattendu s'est produit.
- Détecter si la fréquence d'apparition d'un motif donné (i.e., une sous-séquence) dans une longue séquence est significativement différente de celle attendue [KLC02, GAS05]. Dans ce cas, un analyste pourra s'intéresser à un fragment de séquence spécifique décrivant, par exemple, une surconsommation d'électricité. Ce fragment requête sera alors déclaré anormal si sa fréquence dans la séquence d'activités d'un habitant donné est supérieure à sa fréquence attendue (obtenue par exemple en étudiant la moyenne des fréquences chez les autres habitants).

Le problème que nous avons présenté pour introduire ce chapitre se situe dans le cadre de la deuxième formulation. Il s'agit en effet de détecter dans une séquence d'activités les valeurs anormales pour une dimension d'analyse donnée.

6.1.2 Données d'apprentissage disponibles

Un deuxième axe de comparaison concerne les exemples disponibles pour l'apprentissage et, plus précisément, la présence ou l'absence d'exemples labellisés comme « *normal* » ou « *anormal* ». Nous isolons trois cadres d'apprentissage.

Cadre supervisé Le cadre supervisé suppose que l'ensemble des exemples disponibles pour l'apprentissage contient à la fois des exemples labellisés comme normaux et anormaux. Dans ce cas, le problème de détection d'anomalies peut généralement être abordé comme un problème de classification à deux classes [1]. Cependant, certains problèmes restent inhérents au cadre supervisé. Notamment, les données sont souvent inégalement réparties entre les classes [JAK01]. En effet, les exemples de données anormales sont généralement moins nombreux que ceux liés à

1. Notons malgré tout qu'il peut exister plus de deux classes si plusieurs types d'anomalies doivent être détectés. Ces anomalies correspondent alors à des classes distinctes.

des données normales. Ce type d'approches ne gèrent pas les anomalies inconnues qui ne sont pas apprises lors de la phase d'apprentissage [LL05]. Finalement, dans la pratique, construire un jeu de données adéquat pour l'apprentissage supervisé d'un modèle de détection d'anomalies peut s'avérer difficile et coûteux [CBK09a].

Cadre semi-supervisé Dans le cadre semi-supervisé [FYM05], l'ensemble d'exemples de la base d'apprentissage contient uniquement des données labellisées comme normales. Ces approches sont plus facilement exploitables que celles liées au cadre supervisé puisqu'elles ne nécessitent pas de construire un modèle pour les données anormales. Le principe général est qu'une anomalie est détectée si elle s'écarte du modèle construit pour les données normales. Bien qu'elles soient rares, remarquons que certaines approches s'appuient au contraire sur des données d'apprentissage uniquement anormales. C'est notamment le cas d'approches qui s'inspirent du système immunitaire animal [DFH96, DM02]. Comme pour le cadre supervisé, une critique de ces approches semi-supervisées relève de l'obtention d'un ensemble de données normales labellisées car, pour de nombreux domaines d'application, il est difficile de s'assurer qu'un ensemble de données est parfaitement normal [LL05].

Cadre non-supervisé Une approche s'appuyant sur un cadre d'apprentissage non-supervisé ne requiert pas de données labellisées pour effectuer l'apprentissage. L'hypothèse généralement faite est alors que les anomalies sont bien moins fréquentes que les données normales dans les exemples disponibles [LL05]. Le problème que nous abordons dans ce chapitre se situe dans ce cadre puisque nous ne supposons pas que les données disponibles pour l'apprentissage sont labellisées.

Un panorama plus détaillé des problèmes de détection d'anomalies est donné dans [CBK09a]. De plus, [CBK09b] offre une étude spécifique à la détection d'anomalies dans les données séquentielles.

6.2 Motifs fréquents contextuels pour la détection d'anomalies

Cette section décrit l'approche de détection d'anomalies dans une séquence étendue. Celle-ci s'appuie sur les motifs IT contextuels fréquents extraits dans une séquence étendue [2]. Nous ne détaillons pas ici le principe de l'extraction de tels motifs puisqu'il s'agit du même que celui présenté dans le chapitre 3. Nous nous concentrons donc uniquement sur l'exploitation de ces motifs pour la détection d'anomalies.

6.2.1 Définitions préliminaires

Revenons à la séquence s présentée en introduction de ce chapitre :

$$s = \langle (ab\ E_{moy})^1 (bc\ E_{bas})^2 (e\ E_{haut})^3 (ad\ E_{haut})^4 (cd\ E_{bas})^5 (ab\ E_{haut})^6 \rangle.$$

Les items de la forme E_{bas}, E_{moy} ou E_{haut} ont des spécificités prises en compte dans la suite. Nous définissons par conséquent la notion de A-item.

2. Les motifs IT considérés ici sont ceux présentés dans le chapitre 2 et non ceux dédiés à la prédiction décrits dans le chapitre 5.

Définition 62 (A-item) **:** Soit A une dimension, appelée dimension d'analyse. Un A-item est un item de la forme A_v où $v \in dom(A)$.

Exemple 54 : Dans notre exemple précédent, la dimension E, décrivant la consommation électrique, est une dimension d'analyse. Le domaine de E est composé de trois valeurs *bas*, *moy* et *haut*, i.e., $dom(E) = \{bas, moy, haut\}$.

Pour une dimension d'analyse A donnée, les A-items ont deux spécificités :

- Un et un seul A-item est présent dans chaque itemset d'une séquence étendue. Ces A-items décrivent l'état de la dimension d'analyse A dans chaque itemset de la séquence étendue. En effet, si deux A-items se trouvaient dans le même itemset cela impliquerait que la dimension d'analyse correspondante possède deux valeurs au même moment (i.e., par exemple la consommation électrique est à la fois haute et basse au même moment).
- Les valeurs sur la dimension A peuvent être ordonnées. C'est par exemple le cas lorsque les valeurs du domaine de A sont des valeurs numériques discrétisées. Ainsi, dans notre exemple, nous considérons que la valeur *bas* est inférieure à la valeur *moy*, qui est elle-même inférieure à la valeur *haut*.

L'approche que nous proposons repose sur le fait que certains motifs IT fréquents sont conformes aux circonstances qui entourent l'occurrence d'un A-item. Afin de manipuler ces concepts, nous définissons ci-dessous la notion de recouvrement.

Définition 63 (Recouvrement) **:** Soient I un itemset inter-transactionnel, $S = \langle I_1^{d_1} I_2^{d_2} ... I_n^{d_n} \rangle$ une séquence étendue et I' l'itemset inter-transactionnel sur S à l'estampille d_k (*Cf.* chapitre 2, définition 22). I *recouvre* S à l'estampille d_k si $I \subseteq I'$.

Exemple 55 : Considérons l'itemset IT $I = \{(e, 0)(E_{haut}, 1)(d, 2)\}$ et la séquence s suivante :

$$s = \langle (ab\ E_{haut})^1 (bc\ E_{bas})^2 (e\ E_{haut})^3 (ad\ E_{haut})^4 (cd\ E_{bas})^5 (ab\ E_{haut})^6 \rangle.$$

L'itemset IT I recouvre s à l'estampille 3 (pour une fenêtre de taille 3). En effet, construisons l'itemset IT I' sur s à l'estampille 3 :

$$I' = \{(e, 0)(E_{haut}, 0)(a, 1)(d, 1)(E_{haut}, 1)(c, 2)(d, 2)(E_{bas})\}.$$

Nous constatons que $I \subseteq I'$, i.e., I recouvre s au point de référence 3.

La notion de recouvrement, pour une estampille donnée, correspond à l'inclusion d'un itemset IT dans une séquence étendue, complétée d'une contrainte sur la position des itemsets couverts. Néanmoins, cette contrainte n'est pas suffisante. Nous souhaitons en effet mesurer à quel point l'occurrence d'un A-item dans une séquence étendue peut être jugée comme conforme à l'ensemble des motifs fréquents extraits. Pour ce faire, nous recherchons les motifs qui recouvrent la séquence et l'itemset donnés ainsi que le A-item considéré. De tels motifs, appelés motifs concordants, sont formellement définis comme suit.

Définition 64 (Motif (A, d_k)-concordant) **:** Soient I un itemset IT, $s = \langle I_1^{d_1} I_2^{d_2} ... I_n^{d_n} \rangle$ une séquence étendue et A une dimension d'analyse sur s.

I est un *motif* (A, d_k)-*concordant* sur s s'il existe un entier $d_{k'}$ tel que :

1. I recouvre s à l'estampille $d_{k'}$;
2. il existe un A-item étendu $i' = (A_v, d_k - d_{k'})$ tel que $i' \in I$.

Exemple 56 : Reprenons la séquence s et l'itemset IT I de l'exemple précédent.
I est un motif $(E, 4)$-concordant. En effet, il existe l'estampille 3 telle que I recouvre s (*Cf.* exemple précédent) et de plus un E-item étendu $(E_{haut}, 1)$ est inclus dans I.
L'existence de ce motif concordant peut être interprétée comme suit : « *Il est fréquent qu'une valeur d'électricité haute E_{haut} soit entourée de l'item e une heure avant et de l'item d une heure après* ». Par conséquent, selon ce motif concordant, l'occurrence de E_{haut} est validée par une partie des circonstances qui l'entourent (i.e., la partie décrite par les items du motif concordant).

Selon cette définition, un motif (A, d_k)-concordant recouvre la séquence étendue de telle manière que le A-item de l'itemset d'estampille d_k est également recouvert. L'occurrence de cet A-item est donc, dans une certaine mesure, validée par le motif concordant correspondant. D'autres motifs, appelés motifs discordants, montrent qu'étant donné les circonstances entourant l'occurrence du A-item considéré une autre valeur sur la dimension A peut être attendue.

Définition 65 (Motif (A, d_k)-discordant) **:** Soient un itemset inter-transactionnel I, une séquence étendue $s = \langle I_1^{d_1} I_2^{d_2} ... I_n^{d_n} \rangle$ et une dimension d'analyse A sur s.
I est un *motif* (A, d_k)-*discordant* si il existe un entier $d_{k'}$ tel que :

1. I contient un A-item étendu $(A_v, d_k - d_{k'})$ tel que $A_v \neq A_{v'}$ où $A_{v'}$ est le A-item de l'itemset I_{d_k} ;
2. $I - \{A_v\}$ recouvre s au point $d_{k'}$

v' est appelée la valeur observée sur s et v la valeur discordante de I.

Exemple 57 : Considérons un itemset IT $I' = \{(e, 0)(E_{bas}, 1)(d, 2)\}$. I' est un motif $(E, 4)$-discordant. En effet, il existe l'estampille de référence 3 telle que I' contient un E-item étendu $(E_{bas}, 1)$ et E_{bas} est différent du E-item sur s à l'estampille 4. De plus, si l'on retire $(E_{bas}, 1)$ de I', alors l'itemset IT obtenu recouvre s à l'estampille 3.
L'existence de ce motif discordant peut être interprétée comme suit :« *Il est fréquent qu'une consommation électrique basse E_{bas} soit entourée de l'item e une heure avant et de l'item d une heure après.)* ». Par conséquent, selon ce motif discordant, une valeur basse de consommation d'électricité est attendue lorsque les items e et d sont rencontrés. Cette valeur étant différente de celle observée dans la séquence (E_{haut}), le motif est qualifié de discordant.

Un motif (A, d_k)-discordant se différencie d'un motif (A, d_k)-concordant uniquement par le A-item à l'estampille d_k considérée. L'idée principale qui domine la définition d'un motif discordant est que les items qui entourent le A-item discordant, i.e., qui décrivent les circonstances dans lesquelles le A-item apparaît, recouvrent la situation observée dans s.

ID	Itemset IT	Frequence
m_1	$\{(a,0)(E_{haut},0)\}$	0.6
m_2	$\{(c,0)(E_{haut},1)(E_{haut},2)\}$	0.5
m_3	$\{(e,0)(d,1)(E_{haut},1)(E_{bas},2)\}$	0.4
m_4	$\{(E_{haut},0)(b,2)\}$	0.4

TABLE 6.1 – $\mathcal{M}^c$ l'ensemble de motifs $(E,4)$-concordants sur s.

6.2.2 Score de conformité

D'après les définitions précédentes, l'occurrence d'un A-item à une estampille donnée peut être associée à deux types de motifs. D'une part, les motifs concordants tendent à valider cet A-item car les circonstances qu'ils décrivent sont identiques à celles dans lesquelles le A-item est observé. D'autre part, les motifs discordants montrent que, dans ces mêmes circonstances, il est fréquent de trouver un autre A-item que celui observé. Nous déterminons si le A-item observé est normal ou non en quantifiant le poids des motifs concordants et des motifs discordants. Ainsi, le A-item sera considéré normal si le poids global des motifs concordants est supérieur à celui des motifs discordants et anormal dans le cas inverse.

Nous quantifions, ci-dessous, le poids de l'ensemble des motifs concordants et de l'ensemble des motifs discordants au travers de scores.

Score de concordance. Le calcul d'un score de concordance de la dimension d'analyse A, pour une estampille d_k, consiste à évaluer à quel point l'état de cette dimension est conforme aux motifs qui caractérisent le comportement fréquent. Nous considérons pour ce faire la contribution de chacun des motifs concordants trouvés. Notons, néanmoins, que deux motifs concordants différents n'offrent pas nécessairement la même contribution dans la décision finale. Par exemple, observons le tableau 6.1 qui présente l'ensemble de motifs concordants $(E,4)$-concordants sur s. Les motifs m_3 et m_4 ont une fréquence identique. Pourtant, les circonstances décrites par le motif m_3 impliquent quatre items contre seulement deux pour le motif m_4. Le score défini accordera, par conséquent, plus d'importance au motif m_3. De même, bien que les motifs m_1 et m_4 aient la même taille, la fréquence de m_1 est supérieure à celle de m_4. On attribuera donc à m_1 un poids plus important. Nous considérons ces aspects pour définir le poids d'un motif concordant.

Définition 66 : Le *poids d'un motif concordant* m, noté $poids(m)$, est défini comme :

$$poids(p) = |p| \times Freq_{\mathcal{B}}(p).$$

Exemple 58 : Le poids du motif m_1 est $poids(m_1) = |m_1| \times Freq_{\mathcal{B}}(m_1) = 2 \times 0.6 = 1.2$.

De manière à considérer le poids de chaque motif concordant, un score global de concordance est calculé en agrégeant les contributions individuelles de chacun des motifs concordants.

Définition 67 (Score de concordance) **:** Soit $\mathcal{M}^c$ l'ensemble de tous les motifs (A, d_k)-concordants. Le **score de concordance** de la dimension A à l'estampille d_k est :

$$score_{conc}(A, d_k) = \sum_{m \in \mathcal{M}^c} poids(m).$$

Exemple 59 : Considérons le tableau 6.1, qui présente $\mathcal{M}^c$ l'ensemble de motifs concordants $(E, 4)$-concordants sur s. Le score de concordance de la dimension d'analyse E à l'estampille 4 peut être calculé comme suit :

$$\begin{aligned} score_{conc}(E, 4) &= \sum_{m \in \mathcal{M}^c} poids(m) \\ &= poids(m_1) + ... + poids(m_4) \\ &= 1.2 + 1.5 + 1.6 + 0.8 \\ &= 5.1 \end{aligned}$$

Score de discordance. De même que le score de concordance fournit une mesure de la contribution de l'ensemble des motifs concordants, nous étudions la contribution des motifs discordants pour la prise de décision finale. Afin d'évaluer la contribution individuelle d'un motif, nous tenons compte des mêmes critères que dans le cas des motifs concordants, i.e., leur taille et leur fréquence.

Cependant, un autre aspect intervient pour les motifs discordants. En effet, nous avons précédemment observé que les A-items pouvaient être ordonnés. La notion de discordance implique par conséquent qu'il existe un écart entre la valeur associée au A-item observé dans la séquence et celui décrit par le motif discordant. Par exemple, considérons le domaine $dom(E) = \{bas, moy, haut\}$. Si l'item observé est E_{bas} et l'item discordant est E_{moy}, l'écart est moins important que si l'item discordant est E_{haut}. Cet aspect est particulièrement important lorsque les valeurs sur la dimension E des valeurs numériques discrétisées (données de capteurs, etc.).

Afin de tenir compte des différences d'écart entre deux A-items, nous définissons ci-dessous le degré de discordance.

Définition 68 (Degré de discordance) **:** Soient m un motif (A, i)-discordant sur la séquence s et $d : dom(A) \times dom(A) \rightarrow \mathbb{R}$ une mesure de dissimilarité sur les valeurs v et v' du domaine de A. Considérons v' la valeur discordante sur m et v la valeur observée sur s, le degré de discordance de m, noté $DegDisc(m)$, est alors défini comme :

$$DegDisc(m) = d(v, v').$$

Exemple 60 : Considérons le domaine de la dimension E, $dom(E) = \{bas, moy, haut\}$ et d une mesure de dissimilarité telle que $d(haut, moy) = 0.5$ et $d(haut, bas) = 1$. Les deux motifs $(E, 4)$-discordants m_5 et m_6 présentés dans le tableau 6.2 ne sont pas associés à la même valeur discordante. En effet, m_5 a pour valeur discordante moy tandis m_6 a pour valeur discordante bas. La valeur observée étant $haut$, nous pouvons calculer le degré de discordance de chacun de ces motifs.

ID	Itemset IT	Fréquence
m_5	$\{(d,0)(E_{moy},0)\}$	0.4
m_6	$\{(E_{bas},0)(d,1)(E_{bas},1)\}$	0.3

TABLE 6.2 – $\mathcal{M}^d$ l'ensemble de motifs $(E,4)$-discordants sur s.

Le degré de discordance du motif discordant m_5 est $DegDisc(m_5) = d(haut, moy) = 0.5$. En revanche, le degré de discordance du motif discordant m_6 est $DegDisc(m_6) = d(haut, bas) = 1$.

Le poids d'un motif discordant est défini en tenant compte à la fois de la taille et de la fréquence d'un motif, mais également du degré de discordance d'un motif discordant.

Définition 69 : Le *poids d'un motif discordant* m, noté $poids(m)$, est défini comme :

$$poids(p) = |p| \times Freq_{\mathcal{B}}(p) \times DegDisc(m).$$

Exemple 61 : Le poids du motif m_5 est $poids(m_5) = |m_5| \times Freq_{\mathcal{B}}(m_5) \times DegDisc(m_5) = 2 \times 0.4 \times 1 = 0.8$.

De même que le score de concordance agrège l'ensemble des contributions individuelles des motifs concordants, le score de discordance tient compte de tous les motifs discordants.

Définition 70 : Soit $\mathcal{M}^d$ l'ensemble de tous les motifs (A, d_k)-discordants. Le **score de discordance** de la dimension A au point de référence d_k est :

$$score_{disc}(A, d_k) = \sum_{m \in \mathcal{M}^d} poids(m).$$

Exemple 62 : Considérons le tableau 6.1, qui présente $\mathcal{M}^d$ l'ensemble de motifs $(E,4)$-discordants sur s. Le score de discordance correspondant peut être calculé comme suit :

$$\begin{aligned} score_{disc}(E,4) &= \sum_{m \in \mathcal{M}^d} poids(m) \\ &= poids(m_5) + poids(m_6) \\ &= 0.8 + 0.45 \\ &= 1.25 \end{aligned}$$

Score de conformité. Le score de conformité global basé sur les deux scores précédents, défini entre -1 et 1, doit répondre aux besoins suivants :

- si $score(A, d_k)$ est proche de 1, la valeur de A à l'estampille d_k est considérée comme normale,
- si $score(A, d_k)$ est proche de -1, la valeur de A à l'estampille d_k est considérée comme anormale,

– si $score(A, d_k)$ est proche de 0, la valeur de A à l'estampille d_k est considérée comme incertaine.

Définition 71 (Score de conformité) **:** Soient A une dimension d'analyse et s une séquence. Le **score de conformité** de A à l'estampille d_k sur s, noté $score(A, d_k)$, est défini par :

$$score(A, i) = \frac{score_{conc}(A, i) - score_{disc}(A, i)}{max(score_{conc}(A, i), score_{disc}(A, i))}$$

Exemple 63 : D'après les scores de concordance et de discordance calculés dans les exemples précédents, nous calculons le score de conformité sur E à l'estampille 4 :

$$\begin{aligned} score(E, 4) &= \frac{score_{conc}(E, 4) - score_{disc}(E, 4)}{max(score_{conc}(E, 4), score_{disc}(E, 4))} \\ &= \frac{5.1 - 1.25}{5.1} \\ &= 0.75 \end{aligned}$$

Utiliser la fréquence globale des motifs dans la base peut mener à un problème : les items rares sont moins bien représentés et sont par là-même considérés comme des anomalies. Cependant, le fait de considérer à la fois les motifs concordants et les motifs discordants permet, s'il n'existe pas de motifs concordants ou discordants, de considérer la conformité de tels items comme incertaine (i.e., avec un score proche de 0) et non comme anormaux (i.e., avec un score proche de -1). Par conséquent, l'approche proposée permet de différencier le cas d'un item anormal (i.e., le poids des motifs discordants est bien supérieur à celui des motifs concordants) d'un cas incertain (i.e., trop peu de motifs existent pour juger de la conformité du A-item rencontré).

6.2.3 Lissage des scores

Un problème important pour la découverte d'anomalies est celui des fausses alarmes, en particulier lorsque les données sont bruitées. Ainsi, il est possible qu'une dimension ait un score de conformité faible dans un itemset qui ne correspond pas à un problème réel. Dans ce cas, le score remonte très rapidement dans les itemsets suivants.

Afin de répondre à ce problème, il peut être nécessaire de tenir compte des scores de conformité sur cette dimension dans les itemsets précédents. Nous proposons par conséquent de fixer une taille de fenêtre de lissage, notée w. Le score lissé pour A au point de référence d_k est la moyenne des scores obtenus par A sur les itemsets précédents dans la fenêtre formée par w.

Définition 72 (Score lissé) **:** Soient w une taille de fenêtre, $s = \langle I_1^{d_1} I_2^{d_2} ... I_n^{d_n} \rangle$ une séquence étendue et d_k une estampille sur s. Considérons deux entiers $1 \leq p \leq q \leq n$, tels que :

– $p = max(1, k - w)$

– $q = min(n, k + w)$

Le *score lissé* de A à l'estampille d_k, noté $lscore(A, d_k)$, est défini comme suit :

$$lscore(A, d_k) = \frac{\sum_{i \in \{p,...,q\}} score(A, i)}{q - p + 1}$$

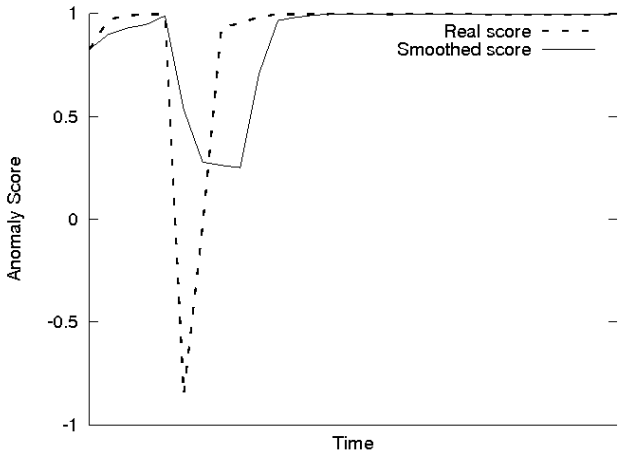


FIGURE 6.1 – Influence du lissage sur le score de conformité (avec $w = 3$).

Exemple 64 : Considérons la séquence étendue s et l'estampille $k = 5$ sur s. Pour une taille de fenêtre w fixée à 2, nous montrons ci-dessous la fenêtre sur s sur laquelle est calculé le score lissé.

$$s = \langle (ab\ E_{moy})^1 (bc\ E_{bas})^2 \overbrace{(e\ E_{haut})^3 (ad\ E_{haut})^4 \underbrace{(\mathbf{cd\ E_{bas}})^{\mathbf{5}}}_{estampille\ 5} (ab\ E_{haut})^6}^{fenêtre\ considérée\ pour\ le\ lissage} \rangle$$

Ainsi, $p = max(1, k - w) = max(1, 3) = 3$ et $q = min(n, k + w) = min(6, 7) = 6$. Par conséquent, le score lissé de E à l'estampille $k = 5$ est égal à :

$$lscore(E, 5) = \frac{\sum\limits_{i \in \{3,\ldots,6\}} score(E, i)}{4}.$$

La figure 6.1 montre l'évolution du score de conformité sur une dimension avec et sans lissage (pour $w = 3$). Sans lissage, nous pouvons constater une chute brusque du score qui pourrait être interprétée comme une anomalie. Néanmoins, la hausse immédiate du score qui suit révèle qu'il ne s'agit pas d'une réelle anomalie. Le lissage permet dans ce cas de conserver un score élevé et d'éviter de générer des fausses alarmes dues au bruit.

6.2.4 Algorithme

D'après les définitions proposées dans le début de cette section, nous proposons l'algorithme 10, qui présente le cœur de notre approche : le calcul du score de conformité sur une séquence étendue s, pour une dimension d'analyse et une estampille données.

Cet algorithme considère une séquence étendue s ainsi qu'une dimension d'analyse A et une estampille k sur s. De plus, l'ensemble de motifs fréquents contextuels $\mathcal{M}$ caractérise les comportements fréquents dans s. Si l'itemset d'estampille k où l'on teste la conformité de A est

Algorithm 10 `CoSCo` : *Conformity Score Computing*

Entrées: une séquence étendue s, une dimension d'analyse A, une estampille k sur s et un ensemble de motifs fréquents contextuels $\mathcal{M}$.

Sorties: le score de conformité $score(A, k)$.

```
/* Initialisation des scores */
```
$S_{conc} \leftarrow 0$
$S_{disc} \leftarrow 0$

```
/* Calcul des scores */
```
pour tout $m \in \mathcal{M}$ **faire**
  si m est un motif (A, k)-concordant **alors**
    $S_{conc} \leftarrow S_{conc} + poids(m)$
  finsi
  si m est un motif (A, k)-discordant **alors**
    $S_{disc} \leftarrow S_{disc} + poids(m)$
  finsi
fin pour
retourne $\dfrac{S_{conc} - S_{disc}}{max(S_{conc}, S_{disc})}$.

associé à un contexte c, alors l'ensemble des motifs considéré est l'ensemble des motifs généraux dans c.

Le principe de l'algorithme 10 est simple. Il teste chacun des motifs de $\mathcal{M}$ et vérifie s'il s'agit d'un motif concordant (resp. d'un motif discordant). Le cas échéant, le poids du motif est calculé et ajouté au poids de concordance (S_{conc}) ou au poids de discordance (S_{disc}). Enfin, l'algorithme calcule et retourne le score de conformité global d'après la définition 71.

Cet algorithme ne décrit pas une approche de détection d'anomalies dans une séquence étendue, mais uniquement le calcul du score de conformité d'une dimension à un moment donné. Pour manipuler le processus global de détection d'anomalies dans la séquence étendue, nous proposons l'algorithme 11.

Celui-ci prend en entrée une séquence étendue s ainsi qu'une dimension d'analyse A et un ensemble de motifs fréquents contextuels $\mathcal{M}$. De plus, il considère également un seuil de conformité minimum $minConf$. Tout A-item dont le score de conformité sera inférieur à ce seuil sera par la suite déclaré comme anormal.

L'algorithme 11 effectue deux parcours des estampilles sur la séquence s. Dans le premier, il calcule pour chaque estampille le score de conformité de la dimension A puis stocke le résultat dans un tableau indexé par l'estampille. Cependant, comme nous l'avons vu dans la section 6.2.3, ces scores ne sont pas suffisants pour effectuer la détection d'anomalies. Un deuxième parcours des estampilles sur s permet de calculer les scores lissés en fonction de la fenêtre w, en exploitant les scores calculés lors de la première passe. Les estampilles dont le score lissé ne satisfait pas le seuil de conformité minimum $minConf$ correspondent à des anomalies et sont donc stockées dans l'ensemble $\mathcal{E}$. Finalement, l'ensemble $\mathcal{E}$ qui contient toutes les anomalies est retourné.

Algorithm 11 `DAESeq` : Detecting Anomalies in Extended Sequences

Entrées: une séquence étendue $s = \langle I_1^{d_1} I_2^{d_2} ... I_n^{d_n} \rangle$, une dimension d'analyse A, un ensemble de motifs fréquents contextuels $\mathcal{M}$, un seuil de conformité minimum $minConf$ et une taille de fenêtre w.

Sorties: l'ensemble $\mathcal{E}$ des couples (i, v), où i est une estampille sur s sur laquelle une anomalie a été détectée et v est le score de conformité associé à cette anomalie.

$\mathcal{E} \leftarrow \emptyset$
`/* 1ère passe sur` s `: calcul des scores de conformité */`
pour tout $i \in \{1, ..., n\}$ **faire**
  $Score[i] \leftarrow CoSCo(s, A, i, \mathcal{M})$
fin pour
`/* 2ème passe sur` s `: calcul des scores lissés */`
pour tout $i \in \{1, ..., n\}$ **faire**
  $conf \leftarrow lscore(A, i)$
  si $conf < minConf$ **alors**
    $\mathcal{E} \leftarrow \mathcal{E} \cup (i, conf)$
  finsi
fin pour
retourne $\mathcal{E}$.

6.3 Expérimentations

Comme nous l'avons vu tout au long de ce chapitre, l'approche proposée vise à détecter des anomalies dans une séquence étendue. Par conséquent, l'évaluation de l'approche nécessite un jeu de données associé à une séquence étendue dans laquelle des anomalies peuvent être étudiées. Les jeux de données *Amazon*, *Puces à ADN* et *Consommation énergétique* que nous avons utilisés jusque là ne peuvent être utilisés pour évaluer la détection d'anomalies.

Nous utilisons donc un nouveau jeu de données issus de la collaboration qui a engendré ce travail de thèse entre l'entreprise *Tecnalia* et l'*Université Montpellier 2*. Ces données sont issues du suivi de trains via des capteurs embarqués. Notre objectif est de détecter les anomalies dans le comportement de certains composants de manière à alerter les experts et pouvoir entreprendre des opérations de maintenance préventive. Par exemple, il s'agira de détecter les anomalies sur une roue à partir des données relevées par des capteurs localisés sur celle-ci et des données décrivant d'autres informations (la température des autres roues ou du moteur, la vitesse du train, etc.).

6.3.1 Description des données

Le système de surveillance des trains utilisé dans nos travaux s'appuie sur une grande quantité de capteurs répartis sur les trains étudiés. Ces capteurs relèvent des informations de températures sur les principaux composants (roues, moteurs, etc.) ainsi que la vitesse générale du train au cours du temps. Nous ne pouvons, pour des raisons de confidentialité, rentrer plus en détails sur la description des données utilisées. Les informations collectées par les capteurs au cours d'un

trajet sont ensuite transformées en une séquence étendue de la forme suivante :

$$s = \langle ...(R1_{bas}R2_{moy}R3_{bas}R4_{bas}A_{moy}V_{bas})^{12}(R1_{moy}R2_{haut}R3_{bas}R4_{moy}...)^{13}...\rangle.$$

La signification de chaque item est la suivante :
- Un item de la forme $R1_{bas}$ décrit une température basse sur la roue $R1$;
- Un item de la forme A_{moy} décrit une température moyenne sur le composant A ;
- Un item de la forme V_{bas} décrit une vitesse du train basse.

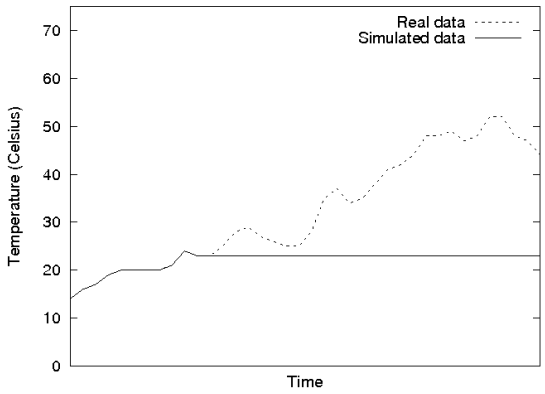
Le problème de détection abordé dans ces expérimentations est la découverte d'anomalies dans l'évolution de la température d'une roue.

Dimensions contextuelles Chaque itemset de la séquence étendue est également associé à des informations contextuelles :
- la *température extérieure* qui a été discrétisée en trois valeurs distinctes (*basse*, *moyenne* et *élevée*) ;
- le type de *fragment* dans le trajet. En effet, nous divisons un trajet (i.e., l'intervalle de temps séparant deux arrêts d'un train) en trois parties. La première, appelée *fragment de début*, correspond à la première phase d'accélération du train (jusqu'à ce que la vitesse se stabilise ou diminue). La deuxième phase est le *fragment de milieu* qui contient tout ce qui se produit entre la phase d'accélération initiale et la dernière phase de décélération. Le dernier fragment du trajet est donc le *fragment de fin* qui correspond à la dernière décélération du train qui mène à son arrêt complet. Ainsi, nous prenons en compte, par exemple, le fait qu'un train au démarrage a un comportement différent de celui de fin de trajet.

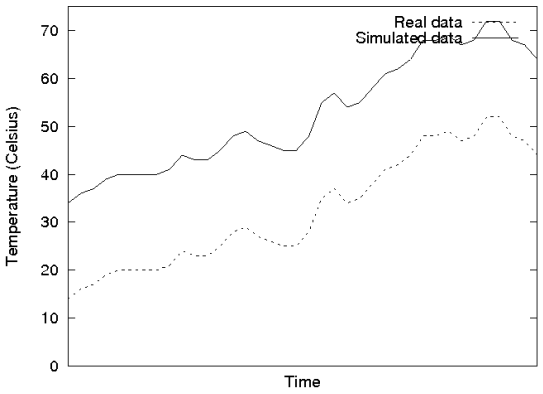
6.3.2 Simulation des anomalies

Les données historisées sur lesquelles nous travaillons n'ont pas été labellisées. Par conséquent, afin d'évaluer notre approche, nous devons simuler des anomalies dans les données existantes. Dans ce but, nous avons corrompu les données disponibles en simulant trois types d'anomalies qui, d'après les experts du domaine, correspondent à des phénomènes fréquemment rencontrés dans les données réelles :
- **Anomalie par blocage de valeurs.** Cette anomalie apparaît lorsque la valeur relevée par un capteur ne se rafraîchit plus au cours du temps. Ce type d'anomalies est généralement lié à un problème de capteur de transmission de l'information. Le comportement du composant est en revanche rarement en cause. Un exemple d'une telle anomalie simulée est donné dans la figure 6.2(a).
- **Anomalie par décalage de valeurs.** Dans ce cas, la valeur relevée par le capteur est décalée par rapport à la valeur réelle. On ajoute ou soustrait donc une constante aux valeurs mesurées réelles relevées par les capteurs. Ce type d'anomalies peut, par exemple, décrire une surchauffe anormale sur un composant. Un exemple d'une anomalie simulée par décalage de valeurs est donné dans la figure 6.2(b).
- **Anomalie par valeurs aléatoires.** Finalement, le dernier type d'anomalies simulé correspond à un comportement aberrant décrit par un capteur. Dans ce cas, les valeurs relevées par le capteurs sont remplacées par des valeurs aléatoires. Ce type d'anomalies correspond

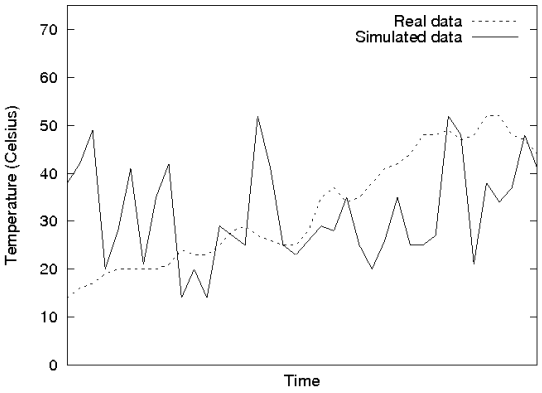
généralement, comme pour les anomalies par blocage de valeurs, à un problème de capteurs ou de transmission de l'information et non à un problème de composant. Un exemple d'anomalie simulée par valeurs aléatoires est donné dans la figure 6.2(c).



(a) Anomalie par blocage de valeurs.



(b) Anomalie par décalage de valeurs.



(c) Anomalie par valeurs aléatoires.

FIGURE 6.2 – Exemples des différents types d'anomalies simulées.

Nous avons construit un jeu de données contenant à la fois des fragments de trajets anormaux (i.e., par une anomalie simulée) et des fragments de trajets supposés normaux. La séquence étendue de départ a donc été segmentée de manière à conserver 100 trajets, eux-mêmes subdivisés en 300 fragments. La moitié de ces fragments (150) ont ensuite été corrompus de manière à créer des anomalies (50 par blocage, 50 par décalage et 50 par valeurs aléatoires). De plus, les anomalies ont été également réparties entre les différents contextes formés par les informations contextuelles (température extérieure, type de fragment dans le trajet).

L'évaluation de la détection d'anomalies consiste alors à appliquer l'algorithme `DAESeq` (*Cf.* algorithme 11 dans la section 6.2) sur chacun de ces fragments de données (qui sont des sous-séquences étendues de la séquence initiale) pour vérifier que les anomalies sont correctement détectées lorsqu'elles existent et que la méthode ne génère pas un nombre excessif de fausses alarmes.

6.3.3 Résultats expérimentaux

Nous décrivons dans cette section les résultats obtenus en appliquant l'algorithme `DAESeq` sur chacune des sous-séquences étendues obtenues dans la sous-section précédente. Afin de compter le nombre d'anomalies détectées, nous fixons un seuil de score de conformité maximum à -0.5 : si un score de conformité inférieur à -0.5 est mesuré dans un itemset d'un fragment de trajet, alors celui-ci est considéré comme anormal. Nous obtenons les résultats présentés dans le tableau 6.3. Les colonnes « *Normal (`DAESeq`)* » et « *Anormal (`DAESeq`)* » comptabilisent les fragments qui ont été déclarés respectivement normaux ou anormaux par `DAESeq`. Les lignes « *Normal* » et « *Anormal* » comptabilisent les fragments réellement normaux ou anormaux.

Par exemple, 138 fragments réellement normaux ont effectivement été classés comme normaux par `DAESeq`, contre 12 qui ont été classés comme anormaux. Nous constatons que les résultats sont au-dessus de 90% à la fois en termes de précision et de rappel, montrant ainsi que notre approche limite à la fois le nombre de fausse alarmes et le nombre d'anomalies non détectées.

	Normal (`DAESeq`)	Anormal (`DAESeq`)	Précision	Rappel
Normal	138	12	0.93	0.92
Anormal	10	140	0.92	0.93

TABLE 6.3 – Résultats de la détection d'anomalies sur l'ensemble du jeu de données.

Le tableau 6.4 présente les résultats obtenus pour chaque type d'anomalies simulé. Nous notons que les anomalies par valeurs aléatoires sont très facilement détectées. Ceci est dû au fait que ces anomalies sont très différentes du comportement attendu. Les anomalies par blocage ou par décalage, bien que générant plus d'erreurs, sont également efficacement détectées.

(a) Anomalies par blocage

	Normal (`DAESeq`)	Anormal (`DAESeq`)
Anomalies simulées	5	45

(b) Anomalies par décalage

	Normal (`DAESeq`)	Anormal (`DAESeq`)
Anomalies simulées	4	46

(c) Anomalies par valeurs aléatoires

	Normal (`DAESeq`)	Anormal (`DAESeq`)
Anomalies simulées	1	49

TABLE 6.4 – Résultats par type d'anomalie.

Ainsi, notre approche donne des résultats de qualité sur le jeu d'anomalies simulées construit.

Cependant, nous sommes conscients que ces résultats ne garantissent pas encore la qualité générale de l'algorithme `DAESeq`. En effet, les anomalies simulées, bien qu'elles correspondent à des cas rencontrés dans la réalité, s'éloignent suffisamment du comportement attendu pour être facilement détectables. Il sera donc nécessaire, dans le futur, de proposer de nouvelles expérimentations afin de nous comparer à d'autres méthodes de détection d'anomalies.

6.4 Discussion

Nous nous sommes intéressés dans ce chapitre à un problème encore peu traité dans la littérature : la détection d'anomalies basée sur les motifs fréquents dans une séquence étendue. Le problème abordé peut être résumé comme suit : « *Étant donné une séquence étendue et des items décrivant l'état d'une variable au cours du temps (des A-items), comment découvrir les itemsets sur s où l'état de A est anormal ?* ». Il s'agit d'un problème difficile ayant de nombreuses perspectives applicatives.

Nous avons montré dans la section 6.1 qu'il existait déjà de nombreuses approches pour résoudre ce problème ou des variantes dans la littérature.

L'approche que nous avons proposée dans la section 6.2 s'appuie uniquement sur les motifs fréquents (contextuels ou non) pour détecter une anomalie, avec la définition de deux types de motifs : les motifs concordants et discordants. Cette méthode possède trois qualités principales. Tout d'abord, non seulement l'approche permet de détecter les anomalies automatiquement, mais de plus l'expert peut consulter l'ensemble des motifs concordants et discordants découverts pour mieux comprendre pourquoi un comportement a été considéré comme anormal. De plus, comme l'approche proposée repose sur les motifs fréquents, elle s'avère pertinente dans un cadre d'apprentissage non-supervisé en tirant parti de l'hypothèse, souvent vérifiée dans la réalité, que les anomalies sont moins fréquentes que les données normales [LL05, CBK09a]. Finalement, le score de conformité fourni par notre approche s'avère particulièrement utile pour faciliter l'analyse des anomalies détectées dans une interface appropriée. Par exemple, la figure 6.3 montre un prototype d'application de visualisation des anomalies détectées. La première permet à l'expert de visualiser l'évolution du score de conformité d'un ensemble de composants pendant un trajet. Dans cet exemple, l'étude des scores de conformité de quatre roues révèle que deux d'entre elles, représentées par les courbes verte et rouge, sont sujettes à un comportement anormal (le score chute à -1) [3].

Toutefois, malgré les qualités de la méthode que nous avons présentée, celle-ci reste préliminaire et peut aisément être améliorée. Elle souffre en effet de défauts qu'il sera intéressant de combler dans le futur. Tout d'abord, le calcul des scores de conformité est défini de manière empirique, sans garantie théorique de sa pertinence. Aussi, une piste intéressante de recherche consistera à étudier d'autres mesures d'intérêt que la fréquence dans le calcul des scores de conformités, afin de centrer l'approche sur les propriétés statistiques des corrélations découvertes entre les items. Notamment, il pourrait être intéressant d'extraire et d'exploiter des règles d'association positives du type « *Lorsque les items a et b apparaissent au temps t, alors l'item*

3. Par ailleurs, la réalité de cette anomalie a été validée par un expert. Il ne s'agit donc pas d'une fausse alarme.

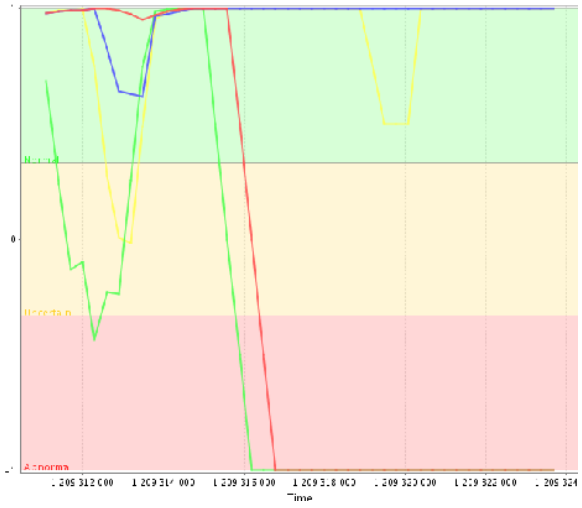


FIGURE 6.3 – Prototype de visualisation des anomalies détectées.

E_{bas} apparaît au temps $t+1$ », mais également des règles négatives du type « *Lorsque les items b et d apparaissent respectivement aux temps t et $t+2$, alors l'item E_{haut} n'apparaît pas au temps $t+1$* ». La combinaison de telles règles, étudiées pour la classification associative dans [AZ04], pourrait s'avérer judicieuse pour la détection d'anomalies.

Par ailleurs, nous avons souligné dans la section 6.3 que les expérimentations menées ne peuvent garantir de manière certaine la qualité de notre approche. Par conséquent, des travaux futurs consisteront d'une part à comparer cette approche aux approches existantes et d'autre part, à effectuer des expérimentations sur d'autres jeux de données comportant des anomalies plus fines que celles simulées dans ce chapitre.

Un autre aspect que nous n'avons pas abordé dans ce chapitre se rapporte à l'étude de l'évolution des scores de conformité au cours du temps. Supposons par exemple que nous étudions les anomalies dans la séquence d'activités d'un habitant au cours d'une longue période (une année, par exemple). L'étude des tendances dans le score d'anomalies peut alors se révéler intéressante. Il pourra par exemple être montré que, bien que le score de conformité reste globalement supérieur au seuil minimum fixé par l'utilisateur, il décroît de manière régulière de semaine en semaine. Ce point de vue plus général sur les scores de conformité pourrait, par exemple, permettre de prédire des anomalies futures par extrapolation des scores de conformité.

Chapitre 7

Bilan, Perspectives et Conclusions

Dans de nombreux domaines, la masse d'informations collectées a augmenté de façon exponentielle au cours de ces vingt dernières années, amplifiant les besoins des experts en termes d'aide à l'accès et à l'analyse de ces multiples sources de connaissances.

Afin de remédier à ce problème, de nombreuses méthodes, techniques et outils ont été proposés par les communautés scientifiques et industrielles issues de domaines variés tels que les statistiques, l'apprentissage automatique, les bases de données ou encore l'interaction homme-machine.

Dans cette thèse, nous nous sommes intéressés à l'Extraction de Connaissances dans les Données, qui consiste à identifier des schémas, motifs d'intérêts utilisables par les experts. Plus particulièrement, nous nous sommes focalisés sur les données séquentielles décrivant des événements ordonnés dans le temps et sur un type de motif particulier prenant en compte des informations additionnelles décrivant le contexte, c'est-à-dire les circonstances liées aux données : les motifs contextuels. À notre connaissance, aucun travaux de la littérature ne permettait de traiter efficacement cette notion de contexte pour l'extraction de motifs.

L'objectif de ce mémoire a consisté à proposer un cadre formel pour exploiter la notion de contexte et des algorithmes exploitant les propriétés associées à cette formalisation. Chaque proposition a été validée par des expérimentations sur des jeux de données réelles aussi bien d'un point de vue sémantique que de passage à l'échelle. Nous avons montré l'intérêt des motifs contextuels pour deux tâches, la classification et la détection d'anomalies, ayant diverses applications pratiques. Ce mémoire s'inscrivant dans le cadre d'une thèse CIFRE financée par le centre de recherche *Tecnalia*, les propositions faites dans ce manuscrit s'intègrent également dans des applications industrielles qui n'ont pas toutes, pour des raisons de confidentialité, été décrites (essentiellement, la maintenance d'équipement et la consommation énergétique).

Toutefois, si le travail réalisé permet de répondre à ces problématiques et offre aux experts des outils efficaces pour manipuler des données séquentielles en prenant en compte les aspects contextuels, certains problèmes demeurent non résolus. Nous présentons donc dans la section 7.1, pour conclure ce manuscrit, un bilan des travaux réalisés puis un certain nombre de perspectives, dans la section 7.2, issues plus ou moins directement des propositions formulées.

7.1 Travail réalisé

Dans le cadre de cette thèse, nous nous sommes intéressés à l'extraction de motifs contextuels et à leur utilisation dans les tâches de classification et de détection d'anomalies. Nous dressons dans cette section un bilan de ce travail.

7.1.1 Motifs fréquents et données séquentielles

Dans le chapitre 2 de ce manuscrit, qui se veut généraliste et introductif, nous nous sommes attachés à définir la notion de motif fréquent que nous avons ensuite instanciée aux données séquentielles. Ces définitions ont, ensuite, été utilisées tout au long de la thèse.

Nous avons également dressé un panorama des travaux existants permettant d'extraire de tels motifs en nous focalisant sur trois types de motifs particuliers : les itemsets fréquents, les motifs séquentiels et les itemsets inter-transactionnels. Ces deux derniers s'avèrent en fait complémentaires et adaptés à de nombreux cas d'application, comme cela a été montré dans les autres chapitres de la thèse.

7.1.2 Extraction de motifs fréquents contextuels

Les motifs présentés dans le chapitre 2 ne tiennent pas compte des informations circonstancielles qui, pourtant, peuvent influer sur ce qui se produit dans les données. Dans le chapitre 3, nous avons donc étudié l'influence des informations contextuelles sur l'extraction de motifs fréquents. Après avoir motivé leur intérêt, nous avons défini formellement la notion de motif contextuel fréquent, trouvé et démontré des propriétés utiles pour proposer une méthode efficace afin de les extraire, puis nous avons donné les algorithmes associés. Nos expérimentations ont souligné l'intérêt de ces motifs sur trois jeux de données réelles (des données textuelles issues de commentaires sur le site *Amazon*, des puces à ADN associées à différents types de cancer du sein et des données de consommation énergétique). Nous avons également validé l'efficacité des algorithmes proposés.

Dans ce chapitre, nous avons tenu à proposer des définitions qui se veulent générales et applicables à divers types de motifs même si nous les avons illustrées uniquement sur le cas des motifs séquentiels. En effet, ces définitions seraient tout aussi valable sur le cas des itemsets fréquents, des sous-graphes fréquents dans une base de graphes, etc. Par ailleurs, nous tenons à souligner une nouvelle fois que même si les motifs contextuels peuvent être rapprochés des motifs séquentiels multidimensionnels souvent décrits dans la littérature, ils sont différents dans leur sémantique et sont donc complémentaires, en présentant des intérêts différents pour des experts.

7.1.3 Extraction de motifs contextuels d'intérêt

Les motifs contextuels fréquents tels que décrits dans le chapitre 3, ne sont sélectionnés qu'en s'appuyant sur la notion de fréquence. Dans le chapitre 4, nous avons étendu ces motifs à d'autres mesures d'intérêt. Nous avons étudié les propriétés théoriques de certaines mesures et les avons exploité pour proposer une méthode d'extraction et les algorithmes associés. Une nouvelle fois, nous avons expérimenté cette méthode sur des jeux de données réelles pour valider l'apport sémantique des motifs sélectionnés selon une mesure d'intérêt. Nous nous sommes également assurés de l'efficacité des algorithmes en terme de passage à l'échelle.

Finalement, si les nombreuses mesures d'intérêt existant dans la littérature n'ont pas toutes été étudiées dans cette thèse, nous pensons que d'autres mesures pourraient être intégrées en exploitant des propriétés différentes de celles mises en avant dans le chapitre 4.

7.1.4 Classification et détection d'anomalies basées sur les motifs contextuels

Si l'apport des motifs contextuels, qu'ils soient sélectionnés selon leur fréquence ou leur intérêt, a été finement décrit dans les chapitres précédents, nous avons abordé leurs potentiels en terme d'applications dans le chapitre 5 pour la classification et dans la chapitre 6 pour la détection d'anomalies. Nous avons montré, pour ces problèmes, l'intérêt d'intégrer les informations contextuelles pour améliorer les performances des méthodes classiques. Dans les deux cas, nous avons formalisé l'intégration des motifs contextuels dans les problèmes généraux, proposé des méthodes adaptées et décrit les algorithmes associés. Les expérimentations soulignent l'intérêt des propositions en termes de qualité.

Les approches proposées peuvent être améliorées sur différents points : par exemple, l'extraction de motifs contextuels pour la classification ne garantit pas, pour l'heure, que l'ensemble de motifs soit exempt de redondance. Dans le cadre de la détection d'anomalies, l'intégration de mesures d'intérêt autres que la fréquence pour considérer plus finement les corrélations décrites par les motifs pourrait également s'avérer utile.

7.1.5 Synthèse

S'inscrivant parmi les travaux classiques du domaine de l'extraction de connaissances, les contributions de ce manuscrit ont consisté à décrire un cadre formel pour exploiter la notion de motifs contextuels, à proposer des méthodes et des algorithmes exploitant les propriétés associées à ces formalisations et à valider chaque proposition sur des jeux de données réelles. Nous avons montré l'intérêt de ces motifs pour deux tâches, la classification et la détection d'anomalies, correspondant à des motivations industrielles issues du financement de cette thèse (maintenance d'équipement, consommation énergétique, etc.).

Si l'on se place dans le cadre de l'exemple que nous avons déroulé tout au long de ce manuscrit sur la consommation électrique dans un immeuble, nos contributions liées à la prise en compte du contexte offrent des solutions pour permettre à un expert de répondre aux questions suivantes :

- *Quels sont les comportements fréquents représentatifs d'une période de l'année (par exemple, l'été) ?*
- *Quels sont les motifs significativement plus fréquents dans une saison en particulier (par exemple, quels sont les motifs que l'on retrouve plus fréquemment en été qu'en hiver ?)*
- *Comment utiliser les motifs pour classer le comportement d'un habitant dans un contexte précis (par exemple, classer l'habitant dans une catégorie d'âge en fonction de son comportement) ?*
- *Comment prédire la consommation électrique en prenant en compte le contexte ?*
- *Comment utiliser les comportements décrits dans les motifs pour détecter des anomalies dans la consommation électrique ?*

Cette liste de questions montre que la prise en compte du contexte offre de nouveaux outils particulièrement utiles aux experts. Toutefois, comme tous travaux de recherche, ces contributions soulèvent de nombreuses perspectives décrites dans la section suivante.

7.2 Perspectives

Si nous avons dressé, dans la section précédente, une liste de perspectives immédiates à chaque contribution, nous décrivons dans cette section des perspectives à plus long terme qui montrent tout le pouvoir applicatif des motifs contextuels.

7.2.1 Motifs clos contextuels

L'extraction de motifs fréquents contextuels, tout comme l'extraction de motifs fréquents traditionnels, génère souvent un grand nombre de motifs présentant une redondance. Par exemple, dans [YHA03], les auteurs mettent en évidence cette redondance dans l'extraction de motifs séquentiels en considérant l'exemple d'une base de séquences ne contenant qu'une seule séquence $S = \langle (a_1)(a_2)...(a_{100}) \rangle$. L'extraction de motifs séquentiels dans une telle base générera $2^{100} - 1$ séquences fréquentes, soit toutes les sous-séquences de S. Ces motifs séquentiels ont tous une fréquence de 1 dans la base et sont par conséquent, excepté pour la plus longue, redondants.

Afin de limiter le nombre de motifs extraits, une solution a été proposée par le biais de l'extraction de motifs fréquents clos [1]. Un motif clos est tel qu'aucun de ses super-motifs n'a un support identique. L'extraction de motifs clos (itemsets fréquents clos, motifs séquentiels clos ou itemsets inter-transactionnels clos, etc.) présente deux avantages :

- elle fournit une représentation condensée exacte de l'ensemble complet des motifs fréquents ;
- elle offre des propriétés d'élagage supplémentaires dans l'espace de recherche qui permettent une extraction plus efficace.

Par conséquent, nous posons la question suivante : « *Est-il possible d'obtenir une représentation condensée exacte de l'ensemble des motifs fréquents contextuels dans un contexte au travers des motifs clos ?* ». Nous répondons par l'affirmative à cette question. En effet, la propriété suivante montre que l'ensemble des motifs fréquents à la fois généraux et clos dans un contexte est une représentation condensée exacte de l'ensemble des motifs généraux de ce contexte (i.e., on peut retrouver les contextes où un motif non-clos est général à partir de ses motifs clos).

Propriété 5 (Motifs généraux clos) **:** Soit c un contexte et m un motif c-général et clos dans c. S'il existe un motif m' tel que m est un super-motif de m' et $Supp_c(m) = Supp_c(m')$, alors m' est c-général.

Démonstration : *En effet, si m est un super-motif de m' et $Supp_c(m) = Supp_c(m')$, alors chaque objet de $\mathcal{B}(c)$ qui supporte m supporte aussi m' et réciproquement. Par conséquent, pour tout contexte c' descendant de c, nous avons $Supp_{c'}(m) = Supp_{c'}(m')$. Or, comme m est fréquent dans les descendants de c (par définition de la c-généralité), m' est également fréquent dans les descendants de c, i.e., m' est c-général.* □

En tirant parti de cette propriété, nous pourrions extraire uniquement les motifs fréquents contextuels clos, i.e., les motifs fréquents contextuels (c, m) tels que m est c-général et clos dans

1. Notons que dans la littérature francophone, deux traductions équivalentes du terme anglais *closed* peuvent être rencontrées : *clos* ou *fermé*.

c. Ceci réduirait d'une part le nombre de motifs contextuels extraits et pourrait également offrir d'intéressantes propriétés pour rendre plus efficace l'extraction des motifs fréquents dans les contextes minimaux, première phase indispensable à l'extraction des motifs fréquents contextuels. Rappelons que la phase d'extraction de ces motifs fréquents dans les contextes minimaux est la phase la plus coûteuse de l'algorithme d'extraction des motifs fréquents contextuels `CFPM` (*Cf.* section 3.5 du chapitre 3). Par conséquent, profiter de l'efficacité de l'extraction des motifs clos dans cette étape aurait un impact immédiat et appréciable sur l'algorithme `CFPM`.

Nous venons de montrer que l'exploitation des motifs clos pouvait se révéler utile dans l'extraction des motifs fréquents contextuels. Qu'en est-il des motifs contextuels d'intérêt définis dans le chapitre 4 ? Là aussi, d'intéressantes pistes de recherche peuvent être trouvées dans des travaux récents. Ces résultats ([SCR05] pour les itemsets clos, [PC09] pour les motifs séquentiels clos), une fois traduits dans les données contextuelles, assurent que les motifs clos d'un contexte y maximisent certaines valeurs d'intérêt, parmi lesquelles l'émergence, le gain d'information, le lift, la fréquence, i.e., toutes les mesures pour lesquelles l'algorithme `CoPaM` peut extraire des motifs contextuels. Ces pistes pourraient servir dans le futur à mieux cerner les contextes dans lesquels un motif peut être général et ainsi améliorer également l'extraction des motifs contextuels d'intérêt.

7.2.2 Variations autour des motifs pour une exploration par les experts guidée par le contexte

Une première perspective intéressante vise à étudier les différentes variantes d'un motif par le prisme des informations contextuelles afin de fournir aux experts des outils pour naviguer et s'approprier les motifs extraits. En effet, une limite courante des méthodes d'extraction de motifs est le nombre de motifs qu'elles génèrent, trop élevé pour être exploité par un expert. Généralement, les mesures d'intérêt sont utilisées pour filtrer les motifs, mais la notion de contexte pourrait également permettre ce filtrage.

Par exemple, considérons les motifs séquentiels contextuels suivants :
- $s = ([jeune, *], \langle(a)\rangle)$
- $s_1 = ([jeune, été], \langle(a)(b)\rangle)$
- $s_2 = ([jeune, hiver], \langle(a)(c)\rangle)$

Les motifs fréquents contextuels s_1 et s_2 peuvent être vus comme deux variantes (« *estivale* » et « *hivernale* ») d'un même motif s, s étant l'ancêtre commun de s_1 et s_2. L'exploitation de cette remarque permettrait de regrouper les motifs, d'identifier des « *variantes* » dans ces regroupements et offrirait une aide précieuse pour l'interprétation des experts. On peut envisager des interfaces de navigation dans les motifs, éventuellement semblables à celles proposées dans [SPB+10], qui prendraient en compte la hiérarchie de contextes et la notion de variation. Ainsi, l'expert pourrait explorer les motifs extraits en commençant par les plus généraux (i.e., associés à des contextes généraux) pour ensuite aller vers des motifs plus spécifiques.

7.2.3 Extraction incrémentale de motifs contextuels dans les flots de données

La découverte de motifs fréquents dans les flots de données est une problématique de recherche difficile. Un exemple de flot de données correspond à l'ensemble des données météo-

rologiques recueillies sur un territoire sur une dizaine d'années par une centaine de capteurs. Toutes les données du flot ne sont pas informatives pour un utilisateur météorologue et le flot est généralement trop volumineux pour que toutes les données soient conservées, ce qui rend complexe l'exploitation décisionnelle de ces données. Pourtant, les données issues du flot sont précieuses et les utilisateurs peuvent y rechercher les informations suivantes :

- *Quels sont les motifs représentatifs de l'ensemble du flot ?* Ces motifs sont fréquents sur l'ensemble du flot.
- *Quels sont les motifs représentatifs d'une période donnée ?* Ces motifs décrivent une période donnée par l'utilisateur (par exemple, les motifs représentatifs de l'été 2011).
- *Quels sont les motifs saisonniers ?* Ces motifs deviennent fréquents de manière saisonnière sur le flot (par exemple, tous les étés ou tous les week-ends).

À notre connaissance, il n'existe pas d'approche d'extraction de motifs qui soit capable d'extraire l'ensemble de ces connaissances. En revanche, les motifs fréquents contextuels permettent de représenter toute l'information souhaitée. Considérons par exemple que le flot de données soit segmenté en jours (chaque jour est un ensemble de données à part entière). Chaque jour peut être décrit par des informations contextuelles sur plusieurs dimensions, telles que le *jour de la semaine*, la *saison* et l'*année*. Ces dimensions contextuelles forment alors une hiérarchie de contextes représentant l'ensemble du flot. Ainsi, les motifs contextuels extraits permettent de répondre aux questions précédemment listées :

- Les motifs représentatifs de l'ensemble du flot sont les motifs $[*; *; *]$-généraux. Ce sont les motifs extrêmement généraux qui ont été fréquents chaque jour.
- Les motifs représentatifs de l'été 2011 sont les motifs $[*; été; 2011]$-généraux. Ces motifs ont été fréquents chaque jour pendant l'été 2011.
- Les motifs saisonniers sont par exemple les motifs $[week - end; *; *]$-généraux. Ces motifs sont fréquents tous les jours de week-end.

Finalement, les motifs fréquents contextuels extraits sur le flot ont un pouvoir d'expression supérieur aux motifs traditionnels. De plus, leur extraction de manière incrémentale offre des propriétés intéressantes. En effet, l'extraction des nouveaux motifs chaque jour peut être effectué de la manière suivante :

1. Lorsque les données d'une journée arrivent, i.e., un batch, nous extrayons les motifs fréquents dans ce batch uniquement.
2. Pour chaque motif extrait dans le nouveau batch, il s'agit de retrouver le contexte de la hiérarchie où il est associé en consultant les contextes auxquels il était déjà associé avant l'arrivée du nouveau batch.

Par ailleurs, l'extraction de motifs dans les flots de données nécessite par nature de définir une « *fonction d'oubli* » des motifs extraits (sinon, le nombre de motifs extraits et stockés s'avérerait potentiellement infini). Là encore, les motifs fréquents contextuels fournissent un moyen simple et naturel d'« *oublier* » les motifs trop anciens ou trop spécifiques. En effet, il s'agira simplement de supprimer les motifs associés à ces contextes. Par exemple, conserver les motifs associés au contexte (minimal) « *jeudi 12 juillet 1994* » ne sera probablement plus pertinent après quelques années. Ce contexte et les motifs associés pourront être supprimés pour ne conserver que des contextes plus généraux du type « *été 1994* ». Ainsi, la prise en compte du contexte « *temporel* » décrivant les données permettra à l'expert de naviguer dans le flot et de le maintenir via les

fonctions d'oubli.

7.2.4 Informations sur la hiérarchie

Les motifs contextuels peuvent également apporter des informations intéressantes sur la hiérarchie des contextes à laquelle ils se rapportent. Des approches telles que celle de [DJBF+08] utilisent les motifs traditionnels extraits à partir de textes pour réorganiser les concepts d'une hiérarchie et rajouter de nouveaux concepts. Les motifs contextuels permettent d'aller un cran plus loin dans cette évolution des hiérarchies.

Par exemple, considérons deux contextes c et c'. Si une large proportion des motifs généraux dans c sont également généraux dans c' (et inversement), alors nous pouvons supposer que ces deux contextes sont « *proches* ». Dans ce cas, s'il n'existe pas déjà, un nouveau contexte peut être créé dans la hiérarchie formant un parent commun de c et c'. De même, si un contexte c ne contient pas ou contient très peu de motifs généraux, ce contexte pourra être supprimé de la hiérarchie.

Finalement, l'exploitation des motifs contextuels selon ce type d'approche pourrait s'avérer très efficace pour construire et maintenir des ontologies qui sont des objets plus évolués que les hiérarchies et qui ont de nombreuses applications dans le contexte du Web sémantique (par exemple, pour l'annotation sémantique de ressources et la structuration de bases de connaissances).

7.2.5 Parallélisation du processus d'extraction de motifs contextuels

La programmation parallèle, reposant sur l'idée que les problèmes complexes peuvent souvent être divisés en problèmes plus simples, consiste à implémenter des algorithmes traitant les sous-problèmes de manière simultanée, en réalisant un maximum d'opérations en parallèle dans un temps réduit.

Les expérimentations effectuées dans ce manuscrit ont montré que le temps d'extraction des motifs contextuels est avant tout lié à l'extraction des motifs fréquents dans les contextes minimaux (le temps de génération des motifs contextuels est négligeable en comparaison). Paralléliser ce processus permettrait un gain certain de performances. En effet, la tâche à effectuer (trouver tous les motifs fréquents) serait décomposée en de multiples sous-tâches exécutées en parallèle (chaque sous-tâche s'occupant indépendamment d'extraire les séquences fréquentes dans un contexte minimal).

Cette perspective, bien que très applicative, prend tout son sens lorsque l'on considère des domaines d'applications avec de très gros volumes de données et l'essor des architectures multi-cœurs de plus en plus performantes, pour lesquelles seules les applications parallélisées pourront espérer gagner en performance.

7.2.6 Synthèse

Pour conclure sur les perspectives de ce travail de thèse, cinq axes ont été dégagés. Le premier se rapporte à l'amélioration des approches d'extraction des motifs contextuels (fréquents ou d'intérêt) en définissant la notion de motif contextuel clos. Le deuxième concerne la mise à disposition des motifs pour les experts en s'appuyant sur une navigation via la hiérarchie des

contextes, qui pourrait être implémentée sous la forme d'une interface innovante. Le troisième concerne le type de données fouillées, en l'occurrence des flots de données, pour lesquels le pouvoir d'expression des motifs contextuels s'avérerait nettement supérieur à celui des motifs traditionnels. Le quatrième concerne l'évolution des hiérarchies associées aux données et pourrait, par exemple, s'avérer très utile dans le contexte de la maintenance des ontologies. Pour finir, la parallélisation des algorithmes de recherche des motifs contextuels pourraient accentuer le passage à l'échelle des méthodes proposées. Cette liste de perspectives pourrait être étendue car de nombreuses autres voies, aussi bien théoriques qu'applicatives, restent à explorer.

Publications dans le cadre de cette thèse

Revue internationale avec comité de relecture

- *Anomaly detection in monitoring sensor data for preventive maintenance* J. Rabatel, S. Bringay and P. Poncelet. Expert Systems With Applications 38 (2011), pp. 7003-7015.

Revue nationale avec comité de relecture

- *Aide au diagnostic de pannes guidée par l'extraction de motifs séquentiels.* J. Rabatel, S. Bringay, P. Poncelet and M. Teisseire. Revue RNTI, Numéro Spécial Fouille de données complexes, Cépadues, 2010, pp. 87-112.

Conférences et ateliers internationaux avec comité de lecture

- *Contextual Sequential Pattern Mining.* J. Rabatel, S. Bringay and P. Poncelet. Domain Driven Data Mining Workshop (DDDM) in conjunction with IEEE ICDM 2010, Sydney, Australia. December, 2010, pp. 981-988.

- *Fuzzy Anomaly Detection in Monitoring Sensor Data.* J. Rabatel, S. Bringay and P. Poncelet. 18th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2010), Barcelona, Spain. July, 2010, pp. 1-8.

- *SOMAD : SensOr Mining for Anomaly Detection in Railway Data.* J. Rabatel, S. Bringay and P. Poncelet. In Proceedings of the 9th Industrial Conference on Data Mining (ICDM 2009), Leipzig, Germany. July, 2009, pp. 191-205.

Conférences et ateliers nationaux avec comité de lecture

- *Extraction de motifs séquentiels contextuels.* J. Rabatel and S. Bringay. Actes des 11ièmes Journées Francophones Extraction et Gestion des Connaissances (EGC 2011), Brest, France. January, 2011, pp. 11-22.

- *Prédiction du grade d'un cancer du sein par la découverte de motifs séquentiels contextuels dans des puces à ADN* J. Rabatel, M. Fabrègue, S. Bringay, P. Poncelet and M. Teisseire. Atelier Extraction de Connaissances et Santé, in conjunction with the 11ièmes Journées Francophones Extraction et Gestion des Connaissances (EGC 2011), Brest, France. January, 2011, pp. 37-48.

- *Aide à la décision pour la maintenance ferroviaire préventive.* J. Rabatel, S. Bringay and P. Poncelet. Actes des 10ièmes Journées Francophones Extraction et Gestion des Connaissances (EGC 2010), Hammamet, Tunisie. January, 2010, pp. 363-368.

- *Visualisation des motifs séquentiels extraits à partir d'un corpus en Ancien Français.* J. Rabatel, Y. Lin, Y. Pitarch, H. Saneifar, C. Serp, M. Roche, A. Laurent. Actes des 8ièmes Journées Francophones Extraction et Gestion des Connaissances (EGC 2008), Nice, France. January, 2008, pp. 237-238.

Annexe A

Preuve du lemme 3

Soit $decomp(C) = \{c_1, ..., c_n\}$ tel que pour tous les entiers $1 \leq i < j \leq n$, $Freq_{c_i}(m) \geq Freq_{c_j}(m)$ (i.e., $decomp(C)$ est un ensemble ordonné en fonction de la fréquence croissante de m dans ses éléments). Par conséquent, nous obtenons $f_{min} = Freq_{c_n}(m)$. Nous cherchons dans un premier temps à prouver l'équation :

$$f_{min} \leq Freq_C(m) \tag{A.1}$$

Étudions le cas où $n = 2$, i.e., la décomposition de C contient deux contextes seulement : c_1 et c_2. D'après la construction de $decomp(C)$, $f_{min} = Freq_{c_2}(m)$ (i.e., $Freq_{c_2}(m) \leq Freq_{c_1}(m)$). Nous pouvons donc réécrire l'équation A.1 :

$$\begin{aligned}
f_{min} &\leq Freq_C(m) \\
Freq_{c_2}(m) &\leq Freq_C(m) \\
\frac{Supp_{c_2}(m)}{|c_2|} &\leq \frac{\sum_{i=1}^{n} Supp_{c_i}(m)}{\sum_{i=1}^{n} |c_i|} \\
\frac{Supp_{c_2}(m)}{|c_2|} &\leq \frac{Supp_{c_1}(m) + Supp_{c_2}(m)}{|c_1| + |c_2|}
\end{aligned}$$

Pour la suite, nous considérons $a = Supp_{c_2}(m)$, $a' = |c_2|$, $b = Supp_{c_1}(m)$ et $b' = |c_1|$:

$$\begin{aligned}
\frac{a}{a'} &\leq \frac{a+b}{a'+b'} \\
a\left(a'+b'\right) &\leq a'\left(a+b\right) \\
aa' + ab' &\leq aa' + a'b \\
ab' &\leq a'b \\
\frac{a}{a'} &\leq \frac{b}{b'} \\
\frac{Supp_{c_2}(m)}{|c_2|} &\leq \frac{Supp_{c_1}(m)}{|c_1|} \\
Freq_{c_2}(m) &\leq Freq_{c_1}(m)
\end{aligned} \tag{A.2}$$

À ce stade, nous avons donc montré que lorsque $decomp(C)$ contient seulement deux éléments l'équation A.1 est vérifiée puisque $Freq_{c_2}(m) \leq Freq_{c_1}(m) \Leftrightarrow Freq_{c_2}(m) \leq Freq_C(m)$. Généralisons à présent ce résultat pour $n > 2$.

La propriété à prouver est donc :

$$\begin{aligned} Freq_{c_n}(m) &\leq \frac{\sum_{i=1}^{n} Supp_{c_i}(m)}{\sum_{i=1}^{n} |c_i|} \\ \frac{Supp_{c_n}(m)}{|c_n|} &\leq \frac{Supp_{c_n}(m) + \sum_{i=1}^{n-1} Supp_{c_i}(m)}{|c_n| + \sum_{i=1}^{n-1} |c_i|} \end{aligned}$$

Nous revenons donc à un cas similaire à l'équation A.2 et en déduisons que la propriété à prouver est à présent :

$$\frac{Supp_{c_n}(m)}{|c_n|} \leq \frac{\sum_{i=1}^{n-1} Supp_{c_i}(m)}{\sum_{i=1}^{n-1} |c_i|}$$

$$Freq_{c_n}(m) \leq \frac{\sum_{i=1}^{n-1} Supp_{c_i}(m)}{\sum_{i=1}^{n-1} |c_i|}$$

Or, si la propriété de départ est vraie pour $n-1$, nous avons :

$$Freq_{c_{n-1}}(m) \leq \frac{\sum_{i=1}^{n-1} Supp_{c_i}(m)}{\sum_{i=1}^{n-1} |c_i|}$$

Comme, par construction de $decomp(C)$, nous savons que $Freq_{c_n}(m) \leq Freq_{c_{n-1}}(m)$, alors nous prouvons bien que :

$$Freq_{c_n}(m) \leq \frac{\sum_{i=1}^{n-1} Supp_{c_i}(m)}{\sum_{i=1}^{n-1} |c_i|}$$

Par conséquent, nous prouvons en exploitant un raisonnement par récurrence que, pour tout $n \geq 2$, $f_{min} \leq Freq_C(m)$. Un raisonnement similaire peut être appliqué pour prouver que $Freq_C(m) \leq f_{max}$. Par conséquent, nous avons la démonstration que :

$$f_{min} \leq Freq_C(m) \leq f_{max}$$

Bibliographie

[AFGY02] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu. Sequential pattern mining using a bitmap representation. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 429–435. ACM, 2002. (Cité en page 31.)

[AIS93] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM, 1993. (Cité en pages 20, 24 et 26.)

[AMS+96] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A.I. Verkamo, et al. Fast discovery of association rules. *Advances in knowledge discovery and data mining*, 12 :307–328, 1996. (Cité en page 27.)

[AS94] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, volume 1215, pages 487–499. Citeseer, 1994. (Cité en pages 20 et 27.)

[AS95] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, ICDE '95, pages 3–14, Washington, DC, USA, 1995. IEEE Computer Society. (Cité en pages 20, 28 et 31.)

[AT11] G. Adomavicius and A. Tuzhilin. Context-aware recommender systems. *Recommender Systems Handbook*, pages 217–253, 2011. (Cité en page 14.)

[AZ04] M.L. Antonie and O.R. Zaïane. An associative classifier based on positive and negative rules. In *Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 64–69. ACM, 2004. (Cité en page 129.)

[BAV04] C. Berberidis, L. Angelis, and I. Vlahavas. Prevent : An algorithm for mining intertransactional patterns for the prediction of rare events. In *Proc. Second Starting AI Researchers' Symposium*, volume 9, 2004. (Cité en page 103.)

[BBR00] J.F. Boulicaut, A. Bykowski, and C. Rigotti. Approximation of frequency queries by means of free-sets. *Principles of Data Mining and Knowledge Discovery*, pages 24–43, 2000. (Cité en page 28.)

[BL06] T. Berners-Lee. Linked data. *International Journal on Semantic Web and Information Systems*, 4(2), 2006. (Cité en page 13.)

[BLF+07] Y. Bu, T.W. Leung, A.W.C. Fu, E. Keogh, J. Pei, and S. Meshkin. Wat : Finding top-k discords in time series database. In *Proceedings of 7th SIAM International Conference on Data Mining*. Citeseer, 2007. (Cité en page 114.)

[BNZ09] B. Bringmann, S. Nijssen, and A. Zimmermann. Pattern-based classification : a unifying perspective. *FROM LOCAL PATTERNS TO GLOBAL MODELS*, page 36, 2009. (Cité en page 92.)

[BP99] S.D. Bay and M.J. Pazzani. Detecting change in categorical data : Mining contrast sets. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 302–306. ACM, 1999. (Cité en page 67.)

[BSO09] S. Budalakoti, A.N. Srivastava, and M.E. Otey. Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety. *Systems, Man, and Cybernetics, Part C : Applications and Reviews, IEEE Transactions on*, 39(1) :101–113, 2009. (Cité en page 114.)

[CBK09a] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection : A survey. *ACM Computing Surveys (CSUR)*, 41(3) :1–58, 2009. (Cité en pages 113, 115 et 128.)

[CBK09b] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection for discrete sequences : a survey. *Knowledge and Data Engineering, IEEE Transactions on*, (99) :1–1, 2009. (Cité en pages 114 et 115.)

[CG02] T. Calders and B. Goethals. Mining all non-derivable frequent itemsets. *Principles of Data Mining and Knowledge Discovery*, pages 1–42, 2002. (Cité en page 28.)

[CG09] S. Chiusano and P. Garza. Selection of high quality rules in associative classification. *Post-Mining of Association Rules : Techniques for Effective*, page 173, 2009. (Cité en page 93.)

[CJP+08] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li. Context-aware query suggestion by mining click-through and session data. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 875–883. ACM, 2008. (Cité en page 14.)

[CLZ+09] H. Cheng, D. Lo, Y. Zhou, X. Wang, and X. Yan. Identifying bug signatures using discriminative graph mining. In *Proceedings of the eighteenth international symposium on Software testing and analysis*, pages 141–152. ACM, 2009. (Cité en page 68.)

[CV95] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3) :273–297, 1995. (Cité en page 93.)

[CYHH07] H. Cheng, X. Yan, J. Han, and C.W. Hsu. Discriminative frequent pattern analysis for effective classification. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 716–725. IEEE, 2007. (Cité en pages 67, 68 et 81.)

[CYHY08] H. Cheng, X. Yan, J. Han, and PS Yu. Direct discriminative pattern mining for effective classification. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 169–178. IEEE, 2008. (Cité en pages 68 et 93.)

[DFH96] P. D'haeseleer, S. Forrest, and P. Helman. An immunological approach to change detection : Algorithms, analysis and implications. In *Security and Privacy, 1996. Proceedings., 1996 IEEE Symposium on*, pages 110–119. IEEE, 1996. (Cité en page 115.)

[DH73] R.O. Duda and P.E. Hart. Pattern classification and scene analysis. 1973. (Cité en page 94.)

[DJ05] J. Deogun and L. Jiang. Prediction mining–an approach to mining association rules for prediction. *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, pages 98–108, 2005. (Cité en page 103.)

[DJBF+08] L. Di-Jorio, S. Bringay, C. Fiot, A. Laurent, and M. Teisseire. Sequential patterns for maintaining ontologies over time. *On the Move to Meaningful Internet Systems : OTM 2008*, pages 1385–1403, 2008. (Cité en page 137.)

[DL99] G. Dong and J. Li. Efficient mining of emerging patterns : Discovering trends and differences. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 43–52. ACM, 1999. (Cité en pages 67 et 93.)

[DM02] D. Dasgupta and N.S. Majumdar. Anomaly detection in multidimensional data using negative selection algorithm. In *Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on*, volume 2, pages 1039–1044. IEEE, 2002. (Cité en page 115.)

[DZ09] K. Deng and O. Zaïane. Contrasting sequence groups by emerging sequences. In *Discovery Science*, pages 377–384. Springer, 2009. (Cité en page 68.)

[DZWL99] G. Dong, X. Zhang, L. Wong, and J. Li. Caep : Classification by aggregating emerging patterns. In *Discovery Science*, pages 737–737. Springer, 1999. (Cité en pages 94, 101 et 102.)

[Edg87] FY Edgeworth. On discordant observations. *Philosophical Magazine*, 23(5) :364–375, 1887. (Cité en page 113.)

[ESL01] E. Eskin, S.J. Stolfo, and W. Lee. Modeling system calls for intrusion detection with dynamic window sizes. In *discex*, page 0165. Published by the IEEE Computer Society, 2001. (Cité en page 114.)

[FDL01] L. Feng, T. Dillon, and J. Liu. Inter-transactional association rules for multi-dimensional contexts for prediction and their application to studying meteorological data. *Data & Knowledge Engineering*, 37(1) :85–115, 2001. (Cité en page 103.)

[FPSSU96] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in knowledge discovery and data mining. 1996. (Cité en page 9.)

[FR03] H. Fan and K. Ramamohanarao. A bayesian approach to use emerging patterns for classification. In *Proceedings of the 14th Australasian database conference-Volume 17*, pages 39–48. Australian Computer Society, Inc., 2003. (Cité en page 94.)

[FYM05] R. Fujimaki, T. Yairi, and K. Machida. An approach to spacecraft anomaly detection problem using kernel feature space. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 401–410. ACM, 2005. (Cité en pages 114 et 115.)

[GAS05] R. Gwadera, M.J. Atallah, and W. Szpankowski. Reliable detection of episodes in event sequences. *Knowledge and Information Systems*, 7(4) :415–437, 2005. (Cité en page 114.)

[Gay09] D. Gay. *Calcul de motifs sous contraintes pour la classification supervisée*. PhD thesis, 2009. (Cité en page 94.)

[GH06] L. Geng and H.J. Hamilton. Interestingness measures for data mining : A survey. *ACM Computing Surveys (CSUR)*, 38(3) :9, 2006. (Cité en pages 42 et 67.)

[GLWX01] G. Grahne, LVS Lakshmanan, X. Wang, and M.H. Xie. On dual mining : from patterns to circumstances, and back. In *Data Engineering, 2001. Proceedings. 17th International Conference on*, pages 195–204. IEEE, 2001. (Cité en pages 40 et 41.)

[Goe02] B. Goethals. *Efficient frequent pattern mining*. PhD thesis, Department of Computer Science University of Helsinki, 2002. (Cité en page 27.)

[GRW08] H. Grosskreutz, S. Rüping, and S. Wrobel. Tight optimistic estimates for fast subgroup discovery. *Machine Learning and Knowledge Discovery in Databases*, pages 440–456, 2008. (Cité en page 67.)

[HFS98] S.A. Hofmeyr, S. Forrest, and A. Somayaji. Intrusion detection using sequences of system calls. *Journal of Computer Security*, 6(3) :151–180, 1998. (Cité en page 114.)

[HHS+00] M. Hirao, H. Hoshino, A. Shinohara, M. Takeda, and S. Arikawa. A practical algorithm to find the best subsequence patterns. In *Discovery Science*, pages 141–154. Springer, 2000. (Cité en page 68.)

[HPY00] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. *ACM SIGMOD Record*, 29(2) :1–12, 2000. (Cité en pages 27 et 31.)

[HTS+08] K. Hashimoto, I. Takigawa, M. Shiga, M. Kanehisa, and H. Mamitsuka. Mining significant tree patterns in carbohydrate sugar chains. *Bioinformatics*, 24(16) :i167, 2008. (Cité en page 68.)

[JAK01] M.V. Joshi, R.C. Agarwal, and V. Kumar. Mining needle in a haystack : classifying rare classes via two-phase rule induction. In *ACM SIGMOD Record*, volume 30, pages 91–102. ACM, 2001. (Cité en page 114.)

[JHZ10] M. Jalali-Heravi and O.R. Zaïane. A study on interestingness measures for associative classifiers. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1039–1046. ACM, 2010. (Cité en pages 42, 67 et 68.)

[JL08] N. Jindal and B. Liu. Opinion spam and analysis. In *Proceedings of the international conference on Web search and web data mining*, pages 219–230. ACM, 2008. (Cité en page 55.)

[Kaa03] E. Kaasinen. User needs for location-aware mobile services. *Personal and ubiquitous computing*, 7(1) :70–79, 2003. (Cité en page 14.)

[KHC97] M. Kamber, J. Han, and J.Y. Chiang. Metarule-guided mining of multi-dimensional association rules using data cubes. In *KDD*, volume 97, pages 207–210, 1997. (Cité en page 40.)

[KK01] M. Kuramochi and G. Karypis. Frequent subgraph discovery. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 313–320. IEEE, 2001. (Cité en page 20.)

[KLC02] E. Keogh, S. Lonardi, and B.Y. Chiu. Finding surprising patterns in a time series database in linear time and space. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 550–556. ACM, 2002. (Cité en page 114.)

[KLLH07] E. Keogh, J. Lin, S.H. Lee, and H.V. Herle. Finding the most unusual time series subsequence : algorithms and applications. *Knowledge and Information Systems*, 11(1) :1–27, 2007. (Cité en page 114.)

[LD05] S.D. Levitt and S.J. Dubner. Freakonomics : A rogue economist explores the hidden side of everything, 2005. (Cité en page 13.)

[LDR01] J. Li, G. Dong, and K. Ramamohanarao. Making use of the most expressive jumping emerging patterns for classification. *Knowledge and Information systems*, 3(2) :131–145, 2001. (Cité en page 94.)

[LFH00] H. Lu, L. Feng, and J. Han. Beyond intratransaction association analysis : mining multidimensional intertransaction association rules. *ACM Transactions on Information Systems (TOIS)*, 18(4) :423–454, 2000. (Cité en pages 35 et 103.)

[LHF98] H. Lu, J. Han, and L. Feng. Stock movement prediction and n-dimensional intertransaction association rules. In *Proceedings of the ACM SIGMOD workshop on research issues in data mining and knowledge discovery*, page 12. Citeseer, 1998. (Cité en page 32.)

[LHM98] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. *Knowledge discovery and data mining*, pages 80–86, 1998. (Cité en page 93.)

[LHP01] W. Li, J. Han, and J. Pei. Cmar : Accurate and efficient classification based on multiple class-association rules. In *icdm*, page 369. Published by the IEEE Computer Society, 2001. (Cité en page 93.)

[LL05] K. Leung and C. Leckie. Unsupervised anomaly detection in network intrusion detection using clusters. In *Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38*, pages 333–342. Australian Computer Society, Inc., 2005. (Cité en pages 115 et 128.)

[LMFK99] G. Liu, T.K. McDaniel, S. Falkow, and S. Karlin. Sequence anomalies in the cag7 gene of the helicobacter pylori pathogenicity island. *Proceedings of the National Academy of Sciences*, 96(12) :7011, 1999. (Cité en page 114.)

[LW07] A.J.T. Lee and C.S. Wang. An efficient algorithm for mining frequent intertransaction patterns. *Information Sciences*, 177(17) :3453–3476, 2007. (Cité en pages 35 et 103.)

[MCP98] F. Masseglia, F. Cathala, and P. Poncelet. The psp approach for mining sequential patterns. *Principles of Data Mining and Knowledge Discovery*, pages 176–184, 1998. (Cité en page 31.)

[MPT04] F. Masseglia, P. Poncelet, and M. Teisseire. Pre-processing time constraints for efficiently mining generalized sequential patterns. In *Proceedings of the 11th International Symposium on Temporal Representation and Reasoning*, pages 87–95. IEEE Computer Society, 2004. (Cité en page 32.)

[MS00] S. Morishita and J. Sese. Transversing itemset lattices with statistical metric pruning. In *Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 226–236. ACM, 2000. (Cité en page 67.)

[MSU+01] T. Miyahara, T. Shoudai, T. Uchida, K. Takahashi, and H. Ueda. Discovery of frequent tree structured patterns in semistructured web documents. *Advances in Knowledge Discovery and Data Mining*, pages 47–52, 2001. (Cité en page 20.)

[MTIV97] H. Mannila, H. Toivonen, and A. Inkeri Verkamo. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1(3) :259–289, 1997. (Cité en pages 20 et 35.)

[MTV94] H. Mannila, H. Toivonen, and A.I. Verkamo. Efficient algorithms for discovering association rules. In *KDD-94 : AAAI workshop on Knowledge Discovery in Databases*, pages 181–192, 1994. (Cité en page 27.)

[NC03] C.C. Noble and D.J. Cook. Graph-based anomaly detection. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636. ACM, 2003. (Cité en page 114.)

[NGDR09] S. Nijssen, T. Guns, and L. De Raedt. Correlated itemset mining in roc space : a constraint programming approach. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 647–656. ACM, 2009. (Cité en pages 67, 68 et 93.)

[NLW09] P.K. Novak, N. Lavrač, and G.I. Webb. Supervised descriptive rule discovery : A unifying survey of contrast set, emerging pattern and subgroup mining. *The Journal of Machine Learning Research*, 10 :377–403, 2009. (Cité en page 67.)

[NR06] V. Nhan and K. Ryu. Future location prediction of moving objects based on movement rules. *Intelligent Control and Automation*, pages 875–881, 2006. (Cité en page 103.)

[ORS98] B. Ozden, S. Ramaswamy, and A. Silberschatz. Cyclic association rules. In *Data Engineering, 1998. Proceedings., 14th International Conference on*, pages 412–421. IEEE, 1998. (Cité en page 35.)

[OZP+97] Z.P. Ogihara, MJ Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New algorithms for fast discovery of association rules. In *In 3rd Intl. Conf. on Knowledge Discovery and Data Mining*. Citeseer, 1997. (Cité en page 27.)

[PBTL99] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices* 1. *Information systems*, 24(1) :25–46, 1999. (Cité en page 28.)

[PC09] M. Plantevit and B. Crémilleux. Condensed representation of sequential patterns according to frequency-based measures. *Advances in Intelligent Data Analysis VIII*, pages 155–166, 2009. (Cité en page 135.)

[PHMA+01] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.C. Hsu. Prefixspan : Mining sequential patterns efficiently by prefix-projected pattern growth. In *icccn*, page 0215. Published by the IEEE Computer Society, 2001. (Cité en pages 31 et 50.)

[PHP+01] H. Pinto, J. Han, J. Pei, K. Wang, Q. Chen, and U. Dayal. Multi-dimensional sequential pattern mining. In *Proceedings of the tenth international conference on Information and knowledge management*. ACM, 2001. (Cité en pages 41 et 64.)

[Pla08] M. Plantevit. *Extraction De Motifs Séquentiels Dans Des Données Multidimensionelles*. PhD thesis, Université de Montpellier 2, France, 2008. (Cité en pages 31, 41 et 64.)

[Qui93] J.R. Quinlan. *C4. 5 : programs for machine learning*. Morgan Kaufmann, 1993. (Cité en page 93.)

[Raï08] C. Raïssi. *Extraction de séquences fréquentes : des bases de données statiques aux flots de données*. PhD thesis, Université de Montpellier 2, France, 2008. (Cité en pages 25 et 31.)

[RCP08] C. Raïssi, T. Calders, and P. Poncelet. Mining conjunctive sequential patterns. *Data Mining and Knowledge Discovery*, 17(1) :77–93, 2008. (Cité en page 31.)

[RF07] K. Ramamohanarao and H. Fan. Patterns based classifiers. *World Wide Web*, 10(1) :71–83, 2007. (Cité en page 93.)

[SA96] R. Srikant and R. Agrawal. Mining sequential patterns : Generalizations and performance improvements. *Advances in Database Technology—EDBT'96*, pages 1–17, 1996. (Cité en pages 31 et 32.)

[Sal10] P. Salle. *Les motifs séquentiels pour les données issues des puces ADN*. PhD thesis, 2010. (Cité en page 57.)

[SCA06] P. Sun, S. Chawla, and B. Arunasalam. Mining for outliers in sequential databases. In *Proceedings of the Sixth SIAM International Conference on Data Mining*, page 94. Society for Industrial Mathematics, 2006. (Cité en page 114.)

[Sch94] H. Schmid. Probabilistic part-of-speech tagging using decision trees. 1994. (Cité en page 55.)

[SCR05] A. Soulet, B. Crémilleux, and F. Rioult. Condensed representation of eps and patterns quantified by frequency-based measures. *Knowledge Discovery in Inductive Databases*, pages 173–189, 2005. (Cité en page 135.)

[SPB+10] Arnaud Sallaberry, Nicolas Pecheur, Sandra Bringay, Mathieu Roche, and Maguelonne Teisseire. Sequencesviewer : Visualisation de séquences ordonnées de gènes ou comment rendre accessible des motifs séquentiels trop nombreux ? In Sadok Ben Yahia and Jean-Marc Petit, editors, *EGC*, volume RNTI-E-19 of *Revue des Nouvelles Technologies de l'Information*, pages 387–392. Cépaduès-Éditions, 2010. (Cité en page 135.)

[Sri05] A.N. Srivastava. Discovering system health anomalies using data mining techniques. In *Proceedings of the 2005 Joint Army Navy NASA Airforce Conference on Propulsion*. Citeseer, 2005. (Cité en page 114.)

[SWL+06] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, et al. Gene expression profiling in breast cancer : understanding the molecular basis of histologic grade to improve prognosis. *JNCI Cancer Spectrum*, 98(4) :262, 2006. (Cité en page 56.)

[SZ05] J. Stefanowski and R. Ziembinski. Mining context based sequential patterns. *Advances in Web Intelligence*, pages 401–407, 2005. (Cité en page 41.)

[TKS02] P.N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 32–41. ACM, 2002. (Cité en pages 42 et 67.)

[TLHF99] A.K.H. Tung, H. Lu, J. Han, and L. Feng. Breaking the barrier of transactions : Mining inter-transaction association rules. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 297–301. ACM, 1999. (Cité en page 32.)

[TLHF03] A.K.H. Tung, H. Lu, J. Han, and L. Feng. Efficient mining of intertransaction association rules. *IEEE Transactions on Knowledge and Data Engineering*, pages 43–56, 2003. (Cité en pages 35 et 103.)

[WC11] C.S. Wang and K.C. Chu. Using a projection-based approach to mine frequent inter-transaction patterns. *Expert Systems with Applications*, 2011. (Cité en pages 35 et 103.)

[WH04] J. Wang and J. Han. Bide : Efficient mining of frequent closed sequences. In *Data Engineering, 2004. Proceedings. 20th International Conference on*, pages 79–90. IEEE, 2004. (Cité en page 31.)

[Wil82] R. Wille. Restructuring lattice theory : an approach based on hierarchies of concepts. *Boston : Dordrecht*, pages 445–470, 1982. (Cité en page 20.)

[Wro97] S. Wrobel. An algorithm for multi-relational discovery of subgroups. *Principles of Data Mining and Knowledge Discovery*, pages 78–87, 1997. (Cité en page 67.)

[YCHY08] X. Yan, H. Cheng, J. Han, and P.S. Yu. Mining significant graph patterns by leap search. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 433–444. ACM, 2008. (Cité en pages 67 et 68.)

[YH02] X. Yan and J. Han. gspan : Graph-based substructure pattern mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, page 721. IEEE Computer Society, 2002. (Cité en page 20.)

[YHA03] X. Yan, J. Han, and R. Afshar. Clospan : Mining closed sequential patterns in large datasets. In *Proceedings of SIAM International Conference on Data Mining*, pages 166–177, 2003. (Cité en pages 31 et 134.)

[Zak01] M.J. Zaki. Spade : An efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1) :31–60, 2001. (Cité en page 31.)

[ZB] A. Zimmermann and B. Bringmann. Ctc-correlating tree patterns for classification. In *Data Mining, Fifth IEEE International Conference on*, pages 4–pp. IEEE. (Cité en page 68.)

[ZH02] M.J. Zaki and C.J. Hsiao. Charm : An efficient algorithm for closed itemset mining. In *2nd SIAM international conference on data mining*, volume 15. Citeseer, 2002. (Cité en page 28.)