Fouille de données par extraction de motifs graduels : contextualisation et enrichissement

Amal Oudni

THÈSE DE DOCTORAT DE
l'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

Informatique

École doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Amal OUDNI

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Fouille de données par extraction de motifs graduels : contextualisation et enrichissement

devant le jury composé de :

Bernd Amann	LIP6 - UPMC	Examinateur
Sadok Ben Yahia	URPAH - Université des Sciences de Tunis	Examinateur
Anne Laurent	LIRMM - Montpellier 2	Rapportrice
Marie-Jeanne Lesot	LIP6 - UPMC	Directrice de thèse
Olivier Pivert	ENSSAT - Rennes 1	Rapporteur
Maria Rifqi	LEMMA - Paris 2	Directrice de thèse

Il ne faut pas penser à l'objectif à atteindre,
il faut seulement penser à avancer.
C'est ainsi, à force d'avancer,
qu'on atteint ou qu'on dépasse ses objectifs
sans même s'en apercevoir.

Bernard Werber

Résumé

Les travaux de cette thèse s'inscrivent dans le cadre de l'extraction de connaissances et de la fouille de données appliquée à des bases de données numériques ou floues afin d'extraire des résumés linguistiques sous la forme de motifs graduels exprimant des corrélations de covariations des valeurs des attributs, de la forme « plus la température augmente, plus la pression augmente ». Notre objectif est de les contextualiser et de les enrichir en proposant différents types de compléments d'information afin d'augmenter leur qualité et leur apporter une meilleure interprétation.

Nous proposons quatre formes de nouveaux motifs : nous avons tout d'abord étudié les motifs dits « *renforcés* », qui effectuent, dans le cas de données floues, une contextualisation par intégration d'attributs complémentaires, ajoutant des clauses introduites linguistiquement par l'expression « d'autant plus que ». Ils peuvent être illustrés par l'exemple « plus la température diminue, plus le volume de l'air diminue, d'autant plus que sa densité augmente ». Ce renforcement est interprété comme validité accrue des motifs graduels. Nous nous sommes également intéressées à la transposition de la notion de renforcement aux règles d'association classiques en discutant de leurs interprétations possibles et nous montrons leur apport limité.

Nous proposons ensuite de traiter le problème des *motifs graduels contradictoires* rencontré par exemple lors de l'extraction simultanée des deux motifs « plus la température augmente, plus l'humidité augmente » et « plus la température augmente, plus l'humidité diminue ». Pour gérer ces contradictions, nous proposons une définition contrainte du support d'un motif graduel, qui, en particulier, ne dépend pas uniquement du motif considéré, mais aussi de ses contradicteurs potentiels. Nous proposons également deux méthodes d'extraction, respectivement basées sur un filtrage a posteriori et sur l'intégration de la contrainte du nouveau support dans le processus de génération.

Nous introduisons également les *motifs graduels caractérisés*, définis par l'ajout d'une clause linguistiquement introduite par l'expression « surtout si » comme par exemple « plus la température diminue, plus l'humidité diminue, surtout si la température varie dans $[0, 10]$ °C» : la clause additionnelle précise des plages de valeurs sur lesquelles la validité des motifs est accrue. Nous formalisons la qualité de cet enrichissement comme un compromis entre deux contraintes imposées à l'intervalle identifié, portant sur sa taille et sa validité, ainsi qu'une extension tenant compte de la densité des données. Nous proposons une méthode d'extraction automatique basée sur des outils de morphologie mathématique et la définition d'un filtre approprié.

Nous définissons aussi les *motifs graduels accélérés*, qui qualifient les corrélations entre les valeurs d'attributs et contextualisent les motifs graduels par l'expression linguistique « rapidement », comme par exemple « plus la température augmente, plus l'humidité augmente rapidement ». Nous traduisons cet effet comme une contrainte de convexité que nous modélisons comme une contrainte de covariation supplémentaire, qui s'exprime dans le même formalisme que les contraintes d'ordre des motifs classiques. Nous proposons et étudions deux méthodes d'extraction, par filtrage a posteriori et intégration dans le processus de génération.

Pour chacune des quatre contextualisation proposées, nous étudions et formalisons la sémantique et l'interprétation souhaitées. Nous proposons ensuite des mesures de qualité pour évaluer la validité des motifs proposés. Nous proposons et implémentons des algorithmes efficaces d'extraction automatique des motifs qui maximisent les critères de qualité proposés. Enfin, nous réalisons une étude expérimentale, à la fois sur des données jouets pour étudier et analyser le comportement des approches proposées, et sur des données réelles pour montrer la pertinence des approches et l'intérêt des motifs extraits. Les expérimentations réalisées pour chaque approche permettent de valider l'apport des différentes formes de motifs proposées, ainsi que leur interprétation associée.

Mots-clés: Extraction de connaissances, fouille de données, résumés linguistiques, règles d'association, motifs graduels, interprétabilité, contextualisation, renforcement, caractérisation, motifs contradictoires, morphologie mathématique, accélération.

Abstract

This thesis's works belongs to the framework of knowledge extraction and data mining applied to numerical or fuzzy data in order to extract linguistic summaries in the form of gradual itemsets : the latter express correlation between attribute values of the form « the more the temperature increases, the more the pressure increases ». Our goal is to contextualize and enrich these gradual itemsets by proposing different types of additional information so as to increase their quality and provide a better interpretation.

We propose four types of new itemsets : first of all, *reinforced gradual itemsets*, in the case of fuzzy data, perform a contextualization by integrating additional attributes linguistically introduced by the expression « all the more ». They can be illustrated by the example « the more the temperature decreases, the more the volume of air decreases, all the more its density increases ». Reinforcement is interpreted as increased validity of the gradual itemset. In addition, we study the extension of the concept of reinforcement to association rules, discussing their possible interpretations and showing their limited contribution.

We then propose to process the *contradictory itemsets* that arise for example in the case of simultaneous extraction of « the more the temperature increases, the more the humidity increases » and « the more the temperature increases, the less the humidity decreases ». To manage these contradictions, we define a constrained variant of the gradual itemset support, which, in particular, does not only depend on the considered itemset, but also on its potential contradictors. We also propose two extraction methods : the first one consists in filtering, after all itemsets have been generated, and the second one integrates the filtering process within the generation step.

We introduce *characterized gradual itemsets*, defined by adding a clause linguistically introduced by the expression « especially if » that can be illustrated by a sentence such as « the more the temperature decreases, the more the humidity decreases, especially if the temperature varies in $[0, 10]$ ℃» : the additional clause precise value ranges on which the validity of the itemset is increased. We formalize the quality of this enrichment as a trade-off between two constraints imposed to identified interval, namely a high validity and a high size, as well as an extension taking into account the data density. We propose a method to

automatically extract characterized gradual based on appropriate mathematical morphology tools and the definition of an appropriate filter.

We define also *accelerated gradual itemsets* that quantify the correlations between the attribute values and contextualize the gradual itemset through the linguistic expression « quickly », for example « the more the temperature increases, the more quickly the humidity increases ».

We propose an interpretation as convexity constraint imposed on the relation between the attributes composing a considered gradual itemset that we model as an additional constraint covariation, which is expressed in the same formalism as constraints of classical gradual itemsets. We propose and study two extraction methods, by filtering a posteriori and integrating in the generation process.

For each of the four proposed contextualizations, we study and formalize the semantics and desired interpretation. We then propose quality measures to evaluate the validity of the given enriched itemset. We also propose and implement efficient algorithms for the automatic extraction of itemsets that maximize the proposed quality criteria. Finally, we carry out experimental studies both on artificial data, to study and analyze the behavior of the proposed approaches, and on real data to show the relevance of the proposed approaches and the interest of extracted enriched itemsets. The experimental results for each approach allow to validate the contribution of the different proposed gradual itemsets and their associated interpretation.

Keywords: Knowledge extraction, data mining, linguistic summaries, association rules, gradual itemsets, interpretability, contextualization, contradictory itemsets, mathematical morphology, acceleration effect.

Remerciements

Il me sera très difficile de remercier tout le monde car c'est grâce à l'aide de nombreuses personnes que j'ai pu mener cette thèse à son terme.

Je remercie chaleureusement Bernadette Bouchon-Meunier de m'avoir accueillie il y a de cela 4 ans en stage de Master 2, de m'avoir donné une chance pour faire ma thèse au sein de son équipe et d'avoir dirigé ma thèse pendant trois ans. Puisse-t-elle trouver ici l'expression de mon profond respect et de toute ma reconnaissance.

Je tiens à remercier tout particulièrement Marie-Jeanne Lesot et Maria Rifqi, mes directrices de thèse, qui m'ont encadrée durant cette thèse. Nous avons eu ensemble des discussions très enrichissantes qui m'ont orientée et aidée dans mes travaux de recherche.

Elles ont toujours été disponibles, à l'écoute de mes nombreuses questions, et ont toujours suivi de prêt l'avancée de mes travaux. Leur rigueur et leurs très nombreuses connaissances m'ont permis de progresser et ont répondu à mes préoccupations.

Leur enthousiasme et leur expérience m'ont montré que le monde de la recherche pouvait être un univers passionnant. Marie-Jeanne et Maria ont consacré beaucoup de leur temps à la relecture minutieuse de ce manuscrit et je les remercie pour tous leurs conseils avisés et pour leur disponibilité. Cette thèse leur doit beaucoup. Pour tout cela merci.

Je remercie vivement Anne Laurent et Olivier Pivert d'avoir accepté d'être rapporteurs de cette thèse ainsi que pour tout le temps qu'ils ont consacré à la lecture de ce manuscrit. Leurs suggestions et remarques ont beaucoup contribué à son amélioration.

J'assure de ma gratitude Bernd Amann et Sadok Ben Yahia, qui ont bien voulu participer à ce jury et s'intéresser à mon travail.

Je tiens à remercier également mon encadrant de stage Marc Damez-Fontaine pour m'avoir fait confiance et m'avoir permis d'intégrer l'équipe LFI en me proposant un sujet de recherche très intéressant. Je lui suis profondément reconnaissante.

Je remercie également Christophe Marsala, directeur de LFI, de s'être intéressé à mon travail, me demandant régulièrement des nouvelles sur l'avancement de mes recherches.

Je tiens aussi à mentionner le plaisir que j'ai eu à travailler au sein de LFI, et je remercie ici tous ses membres pour leur accueil et leur bonne humeur.

Nombreuses sont les personnes ayant contribué de près ou de loin à l'aboutissement de cette thèse. Je tiens simplement à leur exprimer mon amitié et mes remerciements. En particulier, toutes les personnes que j'ai pu rencontrer au sein de LFI, dans le cadre du travail ou simplement pour les bons moments partagés ensemble. Entre autres : Adrien, Bénédicte, Fabon, Gilles, Maël, Marcin, Nicolas, Pierre-Xavier, Sabrina, Sahar, Wenyi, Xavier et toutes les personnes que j'ai oublié de citer mais qui se reconnaîtront, certainement.

Je remercie chaleureusement Patricia Giron pour sa gentillesse, et pour m'avoir encouragée tout au long de mon stage et de ma thèse.

Je remercie également tous les thésards de l'équipe ACASA pour les nombreux bons moments passés ensemble. Entre autres : Alexandre, Amine, Bin, Lise-Marie, Mariane, Mihnéa, Suzanne, Zied.

Je voudrais aussi remercier tout particulièrement Ghislaine pour sa sympathie et le traitement efficace des missions, et les ingénieurs système pour leur efficacité et réactivité dans la résolution des problèmes systèmes.

Je tiens à remercier tout particulièrement Marine et Nathanaëlle pour leurs relectures et corrections de mon manuscrit. Elles ont toujours répondu présentes, même dans les cas les plus désespérés. Merci.

Cette thèse a vu le jour grâce aussi au soutien financier de l'organisme ANR, auquel j'adresse ici tous mes remerciements.

Mes remerciements vont également à tous mes amis qui m'ont permis d'oublier momentanément le travail de ma thèse dans des soirées, repas, sorties ou autres.

J'adresse également tous mes remerciements à Fayçal qui a su me soutenir et m'encourager pendant toute la durée de ma thèse et plus particulièrement durant les derniers mois de rédaction qui n'ont pas toujours été des plus agréables.

Je ne peux finir sans adresser tous mes profonds remerciements à mes parents qui m'ont toujours soutenue pour tous les choix que j'ai faits dans ma vie. Je remercie mes frères et sœurs pour les moments de joie partagés. Un merci particulier pour mes deux frères Brahim et Mohamed, pour m'avoir soutenue et encouragée durant toutes ces années. Ils ont su être à l'écoute et ont toujours accordé un grand intérêt à mes travaux.

À mes parents

Table des matières

Introduction générale

Contexte et motivations

Les moyens informatiques modernes permettent de produire et de stocker d'énormes masses de données numériques, dans de très nombreux domaines tels que la bio-informatique, le web mining ou l'économie. Ces données renferment un certain nombre de connaissances qui décrivent des dépendances ou des corrélations, implicites et utiles.

La mise à disposition de ces grandes quantités de données met en évidence la difficulté à les interpréter et à les analyser. Ceci a favorisé dès le début des années 90 l'essor d'une nouvelle discipline scientifique appelée Extraction de Connaissances dans les bases de données, ECD (Lubinsky, 1989; Piateski & Frawley, 1991; Han et al., 1992; Fayyad et al., 1996c) par la communauté d'intelligence artificielle et Fouille de Données (Anwar et al., 1992; Michalski et al., 1992; Stonebraker et al., 1993; Holsheimer et al., 1995) par la communauté des bases de données. Son objectif est la proposition et la mise au point de techniques d'analyse de données pour l'extraction automatique de connaissances nouvelles, utiles et valides, à partir de grandes quantités de données.

Pour répondre à cet objectif, de multiples méthodes d'extraction d'information apportant des solutions adaptées et permettant de traiter ces masses de données ont vu le jour, regroupées sous le terme générique de Fouille de Données (Berry & Linoff, 2004).

La fouille de données n'est qu'une étape d'un processus d'ECD plus large qui s'étend de la préparation des données jusqu'à la visualisation des résultats. L'ECD constitue le contexte général de notre travail. Nous insistons plus particulièrement sur l'étape de fouille de données, car c'est à ce niveau que se situent nos contributions.

L'un des objectifs fréquemment recherchés en fouille de données est la facilité à interpréter les connaissances extraites. Parmi les nombreux schémas proposés, les règles d'association : elles s'appliquent à un ensemble de données binaires qui représentent la présence ou l'absence d'un item (attribut) dans une instance, où une instance est une donnée représentée par une ligne dans la base de données et constituée d'un ensemble d'items et consistent à établir un lien de co-occurrence des valeurs d'attributs. Elles peuvent être interprétées sous forme d'implication conditionnelle : la présence d'un ensemble de valeurs implique la présence d'autres valeurs. Un exemple classique de l'utilité de ces règles est le panier de la ménagère qui décrit un ensemble d'achats effectué au supermarché ; les règles d'association permettent de décou-

vrir des régularités dans l'ensemble de transactions comme par exemple : « *si on achète du fromage alors on achète du pain* ».

Une règle d'association ne tente pas de décrire globalement les données mais décrit un sous-ensemble réduit de données. Agrawal et al. (1993) ont proposé l'algorithme complet Apriori pour extraire automatiquement ces règles pour de grandes quantités de données.

L'extension des règles d'association à des données numériques a soulevé de nombreux problèmes liés à la fois à leur interprétation et à leur extraction automatique. On distingue plusieurs extensions selon deux axes principaux. Le premier axe concerne le type de données considéré, selon qu'il s'agisse de données quantitatives (valeurs d'attributs numériques) (Agrawal et al., 1996; Chan & Au, 1997; Ben Yahia & Jaoua, 2000; Kuok et al., 1998; Chen et al., 2000; Delgado et al., 2003) ou de données floues (Hong et al., 2003; Fiot et al., 2007). Le second axe concerne le type de corrections exprimées, selon qu'il s'agisse de corrélations entre les attributs ou de corrélations entre les variations des attributs (Hüllermeier, 2002; Berzal et al., 2007; Di Jorio et al., 2008; Laurent et al., 2009). Les schémas exprimant des corrélations entre les variations d'attributs résument des tendances globales des données et sont désignés par le terme *motifs graduels*. Notre thèse est centrée sur ces motifs graduels.

Ces motifs graduels ont pour but l'extraction de tendances internes, exprimées comme des corrélations de covariation entre les valeurs d'attributs. Ils peuvent être illustrés linguistiquement par l'exemple de la phrase « plus la vitesse est élevée et plus on freine fort » ou de façon générale, schématiquement par « plus/moins A, plus/moins B » où A et B sont des attributs. Cette forme constitue une représentation d'information qui résume et caractérise l'ensemble de données de manière globale. Elle est simple et compréhensible, elle répond donc à l'objectif de la fouille de données.

Les motifs graduels diffèrent des règles d'association classiques à la fois par le type de base de données à partir duquel ils sont extraits et par la corrélation qu'ils décrivent. En effet, d'abord les motifs graduels ne sont pas appliqués à des données binaires mais à des données numériques ou floues et ils n'expriment pas une corrélation entre items mais des corrélations de co-variations des valeurs d'attributs. De plus, contrairement aux règles d'association classiques où la corrélation exprimée s'applique à chaque objet individuellement, les corrélations exprimées par les motifs graduels s'appliquent à l'ensemble des données. Ils expriment alors une tendance globale à travers les données et une corrélation sur les variations des attributs en dehors du cadre de l'implication logique. Récemment, de nombreuses interprétations et approches d'extraction automatiques de ces motifs ont été proposées (Hüllermeier, 2002; Berzal et al., 2007; Molina et al., 2007; Di Jorio et al., 2008; Laurent et al., 2009).

L'objectif de la thèse est d'enrichir ces motifs graduels, pour rendre plus précises et mieux interprétables les informations extraites, tout en tenant compte de différents types de contextualisation afin de faciliter leur interprétation.

Objectifs et contributions

Cette thèse se place dans le cadre de la fouille de données appliquée à des bases de données numériques, pour l'extraction de motifs graduels exprimant des corrélations de co-variations des valeurs des attributs. Notre objectif est de contextualiser et d'enrichir les motifs graduels en proposant différents types de compléments d'information. Les nouveaux contextes que nous proposons apportent une précision sur l'information exprimée par le motif, qui la rend plus vraie, compréhensible et facilement interprétable. Dans le cadre de notre travail, nous avons considéré les problématiques liées à l'interprétation des motifs graduels. Afin de répondre à ces problématiques, nos contributions sont les suivantes :

Proposition de nouveaux critères et d'une nouvelle extension pour le renforcement par un nouvel attribut : un enrichissement des motifs graduels par un nouvel attribut a été proposé dans le cas de données floues (Bouchon-Meunier et al., 2010), afin d'extraire un nouveau type de motifs dits *motifs graduels renforcés*. Le complément d'information présenté par ce nouvel attribut permet d'associer au motif un contexte plus important qui augmente la validité du motif graduel. Les motifs graduels renforcés sont exprimés linguistiquement par une clause introduite par « *d'autant plus que* » (Bouchon-Meunier et al., 2010), comme par exemple « *plus on est proche du mur, plus on freine fort, d'autant plus que la vitesse est élevée* ».

Nous nous intéressons à l'interprétation de tels motifs renforcés, à leurs critères de qualité, ainsi qu'à l'algorithme qui permet leur extraction : nous réalisons une étude approfondie des critères de qualité proposés, puis nous proposons de nouveaux critères. Nous les étudions expérimentalement sur des données réelles. Nous examinons également la transposition de ce type d'enrichissement et de ces critères de qualité au cas des règles d'association classiques.

Traitement du problème de motifs graduels contradictoires : bien que des algorithmes d'extraction de motifs graduels très efficaces aient été proposés dans de récents travaux (Di Jorio et al. 2008; 2009), ceux-ci fournissent fréquemment des résultats, dans lesquels se pose le problème de la contradiction. Ces algorithmes peuvent en effet générer des motifs contradictoires tels que « plus A, plus B » et, simultanément, « plus A, moins B ». Ces motifs graduels contradictoires nuisent à la lisibilité et l'interprétabilité de la connaissance extraite. La connaissance utile que présente chaque motif se trouve alors affaiblie. Nous proposons une formalisation de cette notion de contradiction de ces motifs et une nouvelle définition du support qui pénalise les motifs contradictoires en imposant des contraintes supplémentaires dépendant non seulement du motif considéré mais aussi de ses contradicteurs potentiels.

Caractérisation par un intervalle de valeurs : afin d'augmenter encore la facilité d'interprétation des motifs graduels générés en grand nombre, nous proposons de contextualiser les motifs graduels avec une nouvelle information d'ordre sémantique. Nous identifions automatiquement des intervalles de valeurs que nous appelons « intervalles d'intérêt ». Ces intervalles sont introduits par une clause de caractérisation exprimée linguistiquement par l'expression « *surtout si* ». Les motifs graduels caractérisés peuvent être illustrés par l'exemple « plus on est proche du mur, plus on freine fort, surtout si la distance au mur

est dans $[0, 50]$m ». Cette nouvelle clause permet d'apporter une précision supplémentaire au motif graduel qui le rend plus vrai. Elle fournit au motif une validité accrue qui facilite son interprétation. Nous proposons un nouveau critère de qualité pour mesurer la qualité de tels motifs, ainsi qu'une méthode d' optimisation basée sur des outils de morphologie mathématique.

Qualification et précision du mode de corrélation graduelle : nous proposons d'enrichir la découverte de corrélations entre valeurs d'attributs exprimant des dépendances graduelles, en quantifiant et précisant le mode de ces dépendances graduelles. Nous introduisons la notion d'accélération par rapport aux autres attributs. Pour exprimer cette information, nous introduisons l'expression linguistique « rapidement » et nous formalisons l'extraction d'un nouveau type de motifs appelés « motifs graduels accélérés », qui peuvent être illustrés par l'exemple « *plus on est jeune, plus on apprend rapidement* ». Nous définissons un nouveau critère de qualité pour évaluer cette nouvelle information et proposons de le combiner au support graduel classique afin de mieux évaluer la qualité de tels motifs, ainsi qu'à deux algorithmes qui permettent leur extraction automatique.

Organisation du mémoire

La structure de la thèse correspond aux contributions mentionnées ci-dessus :

Le chapitre 1 présente le cadre général dans lequel s'inscrivent nos travaux, en décrivant les travaux existants sur les règles d'association et leurs extensions aux données numériques ou floues, en se concentrant plus particulièrement sur l'extraction de connaissances par motifs graduels. Les chapitres 2, 3, 4 et 5 présentent ensuite respectivement les différentes formes de contextualisation que nous proposons : le chapitre 2 est dédié à l'étude complémentaire de l'enrichissement par renforcement. Le chapitre 3 s'intéresse au problème des motifs graduels contradictoires, le chapitre 4 à la caractérisation des motifs graduels par identification d'intervalles d'intérêt et le chapitre 5 à la contextualisation des motifs graduels par des clauses d'accélération. Tous les chapitres proposent également une étude expérimentale des approches proposées en utilisant des données réelles.

Enfin, dans le chapitre 6, nous concluons et présentons les perspectives soulevées par notre travail.

L'annexe A.2 présente la base de données artificielles et la base de données réelles météorologiques utilisées pour tester les méthodes proposées dans chacun des chapitres.

1

La fouille de motifs

Sommaire

Introduction

L'extraction de connaissances peut prendre de multiples formes, permettant de délivrer à des experts divers types de connaissances. À cet égard, les règles d'association, leurs variantes étendues plus particulièrement et les motifs graduels sont des modèles fréquemment fournis aux utilisateurs finaux.

Ces types de connaissances consistent à mettre en évidence des schémas récurrents dans les données et résument les tendances internes dans un ensemble de données de diverses manières. Ils se distinguent par le type de corrélation exprimée et par la nature des données à partir desquelles ils sont extraits. Ainsi, les motifs extraits à partir de données binaires diffèrent de par la sémantique et la technique d'extraction de ceux extraits à partir de données

quantitatives ou encore de données floues. De même, les motifs décrivant une corrélation entre les attributs se différencient de ceux décrivant une corrélation entre les variations des attributs.

Dans ce qui suit, nous nous intéressons tout d'abord, dans la section 1.1.1, aux règles d'association classiques, à savoir les règles extraites à partir de données binaires et qui expriment des corrélations entre attributs. Cette section présente l'ensemble des définitions préliminaires aux travaux présentés dans cette thèse, ainsi que les principaux algorithmes d'extraction de motifs sur lesquels reposent la plupart des travaux qui les suivent, notamment l'extraction de motifs et de règles graduels. Nous nous focalisons ensuite, dans les sections 1.1.2 et 1.1.3 sur les méthodes qui prennent en compte l'aspect quantitatif des données et qui s'intéressent à la recherche de covariation de valeurs en discutant le cas numérique et le cas flou.

Dans la deuxième partie de ce chapitre, nous détaillons l'état de l'art concernant les motifs et règles graduels. Ces derniers s'appuient sur la notion de corrélation entre valeurs d'attributs et considèrent des données numériques ou/et floues. Nous exposons les interprétations qui leur sont associées puis les algorithmes qui permettent leur extraction automatique. Pour chaque interprétation, nous rappelons les critères de qualité proposés pour l'évaluation de ces motifs et règles graduels. Nous comparons ensuite toutes les approches d'extraction des motifs graduels existantes.

Notre thèse s'appuie principalement sur l'interprétation et l'algorithme GRITE, GRadual ITemset Extraction, proposés par Di Jorio et al. (2009). Pour cette raison, la section entière (5.4.1) est dédiée à cette approche.

1.1 Règles d'association classiques

L'extraction de règles d'association est un domaine de l'extraction de connaissances dans les bases de données qui se définit comme un procédé permettant de trouver des motifs valides, utiles et compréhensibles dans les données (Fayyad et al., 1996b). Historiquement, les règles d'association ont été proposées afin de répondre à la problématique du panier de la ménagère pour une tâche d'analyse de données de supermarchés. Par la suite, les règles d'association ont été étendues dans plusieurs directions, comme la prise en compte des données numériques. Nous présentons d'abord les règles d'association classiques extraites à partir de données binaires dans la section 1.1.1, puis les règles d'association quantitatives extraites à partir de données numériques dans la section 1.1.2 et enfin, les règles d'association floues extraites à partir de données floues dans la section 1.1.3.

Nous ne détaillons pas les multiples applications auxquelles les règles d'association et leurs variantes ont donné lieu (Koperski & Han, 1995; Özden et al., 1998; Savasere et al., 1998).

1.1.1 Cas binaire

Cette section est consacrée aux règles d'association dans le cas classique, c'est-à-dire le cas de données binaires. Nous présentons leur définition avant de décrire les critères de qualité puis les algorithmes d'extraction.

Définitions

Le problème d'extraction de règles d'association introduit par Agrawal et al. (1993) a pour but de découvrir des relations significatives entre attributs binaires (présence ou absence de l'attribut). Un exemple de règle d'association extraite d'une base de données de ventes de supermarché est « si on achète des céréales, alors on achète du lait ». Cette règle indique que les clients qui achètent des céréales ont également tendance à acheter du lait.

De façon classique, les règles d'association s'appliquent à un ensemble de données dites transactionnelles : chaque transaction contient une liste d'*items*. Dans l'exemple des ventes de supermarché, les items correspondent aux produits achetés et la transaction à un ticket de caisse. Cette base transactionnelle est représentée par une base de données binaires où les attributs correspondent aux items possibles. Ils prennent pour valeur 1 ou 0, indiquant respectivement la présence ou l'absence de cet item dans la transaction correspondante.

Définition 1.1 (Motif-itemset)**.** Un *motif*, aussi appelé *itemset*, est un sous-ensemble non vide de $\mathcal{I}$ où $\mathcal{I}$ représente l'ensemble des items.

À un motif M, on associe sa longueur, k, définie comme le nombre d'items qu'il contient. Un motif de longueur k est noté $k-$motif.

Définition 1.2 (Règle d'association)**.** Une *règle d'association*, notée $M_1 \rightarrow M_2$, est constituée de deux motifs disjoints non vides liés par une relation de causalité, M_1 et M_2 ; M_1 est appelé la *prémisse* de la règle et M_2 le *conséquent* de la règle.

Critères de qualité des règles d'association

Pour extraire les règles d'association pertinentes, on se base sur des critères de qualité qui capturent différentes définitions de pertinence. Les plus classiques sont le *support* et la *confiance* (Agrawal et al., 1993; Agrawal & Srikant, 1994), dont nous rappelons les définitions ci-dessous, le tableau 1.1 présente une liste plus complète (Lallich & Teytaud, 2004; Lenca et al., 2004). De nombreuses études comparatives de ces critères, que nous ne détaillons pas ici, ont été menées (Hilderman & Hamilton, 2001 ; Lenca et al. (2003; 2004) ; Lallich & Teytaud, 2004).

Dans la suite, en considérant que A est un item et M un motif et n le nombre total de transactions dans la base, on note respectivement $n(A)$, $n(M)$ et $n(\bar{A})$ le nombre de transactions qui contiennent respectivement A, le nombre de celles qui contiennent tous les attributs composant M, et le nombre de transactions qui ne contiennent pas A ($\bar{A}$ représente l'absence de l'item A).

Définition 1.3 (Support d'un motif)**.** Le *support d'un motif* M est défini par

$$supp(M) = \frac{n(M)}{n} \tag{1.1}$$

Le support d'un motif est donc le rapport de la cardinalité de l'ensemble des transactions qui contiennent tous les items de M par la cardinalité de l'ensemble de toutes les transactions. Il capture la portée du motif, en mesurant sa fréquence d'occurrence.

Nom	Formule
Confiance	$\frac{n(AB)}{n(A)}$
Confiance centrée	$\frac{n(AB)}{n(A)} - n(B)$
Pearl	$n(A)\lvert\frac{n(AB)}{n(A)} - n(B)\rvert$
Piatetsky-Shapiro	$n \times n(A)(\frac{n(AB)}{n(A)} - n(B))$
Loevinger	$\frac{\frac{n(AB)}{n(A)} - n(B)}{n(\bar{B})}$
Zhang	$\frac{n(AB) - n(A)n(B)}{\max\{n(AB)n(\bar{B}); n(B)(n(B)n(A\bar{B})\}}$
Corrélation	$\frac{n(AB) - n(A)n(B)}{\sqrt{n(A)n(\bar{A})n(B)n(\bar{B})}}$
Indice d'implication	$\sqrt{n}\frac{n(A\bar{B}) - n(A)n(\bar{B})}{\sqrt{n(A)n(\bar{B})}}$
Lift	$\frac{n(AB)}{n(A)n(B)}$
Surprise	$\frac{n(AB) - n(A\bar{B})}{n(B)}$
Conviction	$\frac{n(A)n(\bar{B})}{n(A\bar{B})}$
Sebag-Schoenauer	$\frac{n(AB)}{n(A\bar{B})}$
Multiplicateur de cote	$\frac{n(AB)n(\bar{B})}{n(A\bar{B})n(B)}$
J-mesure	$n(AB)\ log\frac{n(AB)}{n(A)n(B)} + n(A\bar{B})\ log\frac{n(A\bar{B})}{n(A)P(\bar{B})}$

Tableau 1.1 – Principales mesures de qualité d'une règle d'association $A \rightarrow B$ (Lallich & Teytaud, 2004).

Définition 1.4 (Support d'une règle)**.** Le *support d'une règle* $R = M_1 \rightarrow M_2$ est la proportion de transactions contenant à la fois la prémisse et le conséquent de la règle par rapport au nombre total de transactions. Il est défini par

$$supp(R) = supp(M_1 \cup M_2) \tag{1.2}$$

La mesure du support est une mesure symétrique. Elle évalue les règles $M_1 \rightarrow M_2$ et $M_2 \rightarrow M_1$ de manière équivalente.

Définition 1.5 (Confiance d'une règle)**.** La *confiance d'une règle* $R = M_1 \rightarrow M_2$ est définie comme

$$conf(R) = \frac{supp(M_1 \cup M_2)}{supp(M_1)} \tag{1.3}$$

Elle indique la proportion de transactions contenant le conséquent parmi celles qui contiennent la prémisse. Elle peut être interprétée comme probabilité conditionnelle $P(M_2|M_1)$ et calculée à partir des supports.

Contrairement à la mesure de support, la mesure de confiance n'est pas symétrique. Elle évalue la qualité d'une règle où une relation de causalité est imposée et elle capture sa précision.

Dans notre travail, nous nous sommes restreinte aux mesures de support et de confiance pour deux raisons principales : premièrement, notre travail est orienté vers la sémantique et l'interprétabilité des motifs et non pas vers les mesures de qualité ; deuxièmement, les approches que nous proposons sont basées sur les algorithmes d'extraction fondés sur ces deux mesures de qualité. Pour ces raisons, nous ne détaillons pas les mesures données dans le tableau 1.1.

Problème d'extraction des règles d'association

Le problème de l'extraction des règles d'association consiste à trouver, à partir d'une base de données, l'ensemble de toutes les règles d'association dont le support et la confiance (ou tout autre critère de qualité choisi par exemple parmi ceux rappelés dans le tableau 1.1) sont supérieurs à des seuils respectivement notés minSupp et minConf fixés par l'utilisateur. Pour ce faire, le processus d'extraction des règles se déroule en deux étapes : d'abord, les motifs fréquents sont extraits, c'est-à-dire les motifs dont le support dépasse le seuil de support minSupp, ensuite les relations de causalité dans ces motifs sont mises en évidence.

L'espace de recherche que les algorithmes doivent explorer est de taille exponentielle suivant le nombre m d'items, puisque le nombre de motifs fréquents potentiels est 2^m. Afin de limiter cet espace de recherche, les algorithmes se basent sur la propriété d'anti-monotonie du support.

Propriété (anti-monotonie du support)**.** *Soit M_1 et M_2 deux motifs. Si $M_1 \subseteq M_2$ alors $supp(M_1) \geq supp(M_2)$.*

Cette propriété est particulièrement importante dans les algorithmes d'extraction de motifs, puisqu'elle permet d'affirmer que si un motif M de taille k est fréquent, alors tout motif $M_1 \subseteq M$ est aussi fréquent, c'est-à-dire tout sous-motif d'un motif fréquent est fréquent. Au contraire, si M est non fréquent, alors tout motif $M_2 \supseteq M$ est non fréquent, c'est-à-dire tout sur-motif d'un motif non fréquent est aussi non fréquent. Cela permet de ne pas tester ou même générer les sur-motifs d'un motif non fréquent.

Il existe deux techniques d'extraction, générer-et-élaguer et diviser-pour-régner, présentées successivement ci-dessous. Dans cette thèse, nous adoptons une approche « générer-et-élaguer » et les algorithmes que nous proposons sont basés sur l'algorithme de génération Apriori (Agrawal et al., 1996) qui est détaillé ci-dessous.

Extraction par la technique générer-et-élaguer

Les algorithmes reposant sur cette technique parcourent en largeur l'espace de recherche par niveau et considèrent un ensemble de motifs d'une taille donnée lors de chaque itération. À chaque niveau k, un ensemble de candidats de taille k est généré et les motifs fréquents sont retenus pour en générer d'autres au niveau suivant par jointure. Les supports des motifs candidats sont calculés et les candidats qui ont le support inférieur à minSupp sont élagués. Cet élagage est justifié par la propriété d'anti-monotonie du support.

Nous présentons dans cette section l'algorithme Apriori qui a été le premier algorithme par niveau proposé pour l'extraction de règles d'association (Agrawal et al., 1996) et qui constitue le principe sur lequel sont basées nos approches, ainsi que quelques variantes.

Algorithme Apriori L'algorithme Apriori est présenté dans l'algorithme 1 en utilisant les notations suivantes :

— C_k : ensemble des k-motifs candidats dont on ne connaît pas encore le support ;

— F_k : ensemble des k-motifs fréquents de taille k.

Les motifs fréquents sont calculés de façon itérative, dans l'ordre ascendant suivant leur taille. À chaque itération, la base de données est parcourue une fois et tous les motifs fréquents de taille k sont générés.

La ligne 1 trouve tous les $1-$motifs fréquents. L'algorithme alterne ensuite la génération des candidats et sélectionne parmi eux ceux étant fréquents dans les lignes 3 à 15 : à l'itération k, l'ensemble F_{k-1} des $(k-1)-$ motifs fréquents correspondant aux motifs de niveau $(k-1)$ est utilisé pour générer l'ensemble C_k des $k-$motifs candidats.

La procédure Apriori-Gen appelée en ligne 4 est présentée dans l'algorithme 2. Elle prend F_{k-1} en entrée et génère C_k comme résultat. L'initialisation de C_k à l'ensemble vide est faite en ligne 1. Ensuite, une jointure est effectuée entre les éléments de F_{k-1} (lignes 2 à 6). Deux motifs p et q de F_{k-1} forment un motif c si et seulement s'ils ont $(k-2)$ attributs (dans le préfixe) en commun, ce qui est exprimé en utilisant l'ordre lexicographique [1] dans la condition de la ligne 4 de l'algorithme Apriori-Gen. Les étapes suivantes (lignes 7 à 11) assurent, après avoir généré un candidat de taille k à partir de deux $(k-1)$-motifs fréquents, que tous les sous-ensembles du nouveau candidat sont fréquents.

Une fois que l'ensemble C_k des motifs candidats a été calculé, la base de transactions est parcourue afin de calculer le support de chaque candidat. Ainsi, parmi les candidats de

1. Un ordre lexicographique est une relation d'ordre sur t^k , où t est un ensemble totalement ordonné et k un entier. La relation d'ordre est définie de la façon suivante : $(x_1, x_2, ..., x_k) \leq (y_1, y_2, ..., y_k)$, si et seulement s'il existe i tel que pour tout $j < i, x_j = y_j$ *et* $x_i < y_i$.

Algorithm 1 Apriori $(\mathcal{DB}, minSupp)$

Entrées : $\mathcal{DB}$ base de données transactionnelle, $minSupp$ seuil du support minimum
Sortie : F ensemble de tous les motifs fréquents de $\mathcal{DB}$

1: $F_1 \leftarrow \{1\text{-motif fréquent}\}$
2: $k = 2$
3: **while** $F_{k-1} \neq \emptyset$ **do**
4: *//génération des candidats, voir algorithme 2*
5: $C_k \leftarrow Apriori - Gen(F_{k-1})$
6: *//calcul du support des candidats*
7: **for** $t \in \mathcal{DB}$ **do**
8: **for all** $c \in C_k$ **do**
9: **if** c est contenu dans t **then**
10: $count(c) + +$
11: *//incrémentation du nombre d'occurrence de c avec le compteur count*
12: **end if**
13: **end for**
14: *//sélection des motifs vérifiant la contrainte de support minSupp*
15: $F_k \leftarrow \{c \in C_k / supp(c) = \frac{count(c)}{|\mathcal{DB}|} \geq minSupp\}$
16: **end for**
17: $k = k + 1$
18: **end while**
19: $F \leftarrow \bigcup_k F_k$

Algorithm 2 Apriori-Gen $(F_k - 1)$

Entrée : $F_k - 1$
Sortie : C_k

1: $C_k = \emptyset$
2: **for all** $p \in F_{k-1}$ **do**
3: **for all** $q \in F_{k-1}$ **do**
4: **if** $p(1) = q(1), p(2) = q(2), ..., p(k-2) = q(k-2), p(k-1) < q(k-1)$ **then**
5: $c \leftarrow p \cup q(k-1)$
6: $C_k \leftarrow C_k \cup \{c\}$
7: **end if**
8: **for all** $s \subseteq c$ (avec s un $(k-1)$-motif) **do**
9: **if** $s \notin F_k - 1$ **then**
10: remove c from C_k
11: **end if**
12: **end for**
13: **end for**
14: **end for**
15: Retourner C_k

C_k, ceux qui sont contenus dans la transaction t voient leur nombre d'occurrences incrémenté dans la ligne 10. Par la suite, seuls ceux qui ont un support supérieur à minSupp sont retenus.

Pour générer les règles d'association, on considère l'ensemble F des motifs fréquents trouvés par l'algorithme 1. Pour chaque motif fréquent M, on considère tous ses sous-ensembles (tous fréquents d'après la propriété d'anti-monotonie du support). À partir de ces sous-ensembles fréquents, on génère toutes les règles $M_1 \rightarrow M \backslash M_1$ pour tout $M_1 \subset M$ telles que leurs confiances respectives dépassent le seuil minimum de confiance.

Variantes On trouve dans la littérature de nombreux algorithmes basés sur cette technique, permettant de générer tous les motifs fréquents d'une base transactionnelle. Ces algorithmes peuvent être classés en trois approches principales. La première consiste à parcourir itérativement par niveau l'ensemble des motifs. Cette approche inclut donc l'algorithme Apriori ainsi que, par exemple, AprioriTid (Agrawal & Srikant, 1994), Partition (Savasere et al., 1995), Sampling (Toivonen, 1994) ou l'algorithme DHP (Direct Hashing and Pruning) proposé par Park et al. (1995). La seconde est basée sur l'extraction des motifs fréquents maximaux. Parmi les algorithmes les plus efficaces basés sur cette approche, on peut citer MaxMiner (Bayardo, 1998), Pincer-Search (Lin & Kedem, 1998), MaxClique et MaxEclat (Zaki et al., 1997). La dernière est basée sur l'extraction de motifs fréquents fermés où un motif fermé est un ensemble maximal d'items communs à un ensemble d'objets. En ce qui concerne cette dernière approche, on peut citer Close citePasquier99a et A-Close (Pasquier et al., 1999b) qui sont aussi des algorithmes par niveau, Titanic (Stumme et al., 2002) et Charm (Zaki & Hsiao, 2002).

Extraction par la technique diviser-pour-régner

Les algorithmes reposant sur cette technique parcourent en profondeur l'espace de recherche et divisent la base de données en sous-ensembles de données, puis appliquent le processus d'extraction des motifs fermés (Pasquier et al., 1999c) récursivement sur ces sous-ensembles. Ce processus d'extraction repose sur un élagage de la base de données basé essentiellement sur une métrique statistique et des heuristiques.

Le principe de cette technique est d'éviter l'inconvénient de la technique « générer-et-élaguer », à savoir la génération d'un nombre excessif de candidats. L'exemple principal dans cette catégorie est l'algorithme FP-Growth (Frequent-Pattern Growth) (Han et al., 2000) qui construit les motifs fréquents sans génération de candidats. Cet algorithme compresse tout d'abord les motifs fréquents représentés dans la base de données à l'aide d'une structure compacte appelée FP-Tree (Frequent-Pattern tree) dont les branches contiennent les associations possibles des items. Il fouille ensuite le FP-tree, ce qui permet de générer tous les motifs fréquents possibles.

La technique « diviser-pour-régner » a été également implémentée dans d'autres algorithmes, comme par exemple Closet (Pei et al., 2000) inspiré de l'algorithme FP-Growth en utilisant la même structure de données. En revanche, cet algorithme est basé sur l'approche

d'extraction de motifs fermés fréquents. Plusieurs variantes de cet algorithme ont été proposées en gardant le même principe et en apportant des améliorations (Wang et al., 2003; Grahne & Zhu, 2003). Ces variantes adoptent la même structure de données FP-tree (Han et al., 2000) qui permet de compresser la base de données et de fusionner plusieurs transactions, lorsqu'elles partagent un même item.

Méthodes d'optimisation

Les algorithmes basés sur le principe d'Apriori souffrent de la gestion du nombre de candidats qu'ils peuvent générer, surtout pour des valeurs de support relativement faibles. Des travaux récents ont proposé une série d'algorithmes qui introduisent plusieurs optimisations et structures de données pour améliorer les performances du processus d'extraction de motifs fréquents. Ces algorithmes sont centrés essentiellement sur la réduction de la taille de l'espace de recherche dans le but de le stocker en mémoire et de réaliser moins d'entrées/sorties. Ils sont aussi focalisés sur la minimisation du coût de l'étape de calcul du support. Parmi ceux-ci, nous pouvons citer les méthodes proposées par Park et al. (1995); Brin et al. (1997b); Zaki (1998); Gardarin et al. (1998); Bastide et al. (2000); Han et al. (2000); Bykowski et Rigotti (2001); Bastide et al. (2002); Calders et Goethals (2002); Boulicaut et al. (2003); Geerts et al. (2005). Dans l'objectif de limiter le nombre de candidats générés, d'autres travaux introduisent des représentations condensées pour l'extraction d'ensembles de motifs condensés dont la cardinalité est plus réduite, mais avec le même niveau de pertinence que l'ensemble de tous les motifs fréquents (Mannila & Toivonen, 1996). Parmi ces représentations condensées, on peut citer les représentations closes (Pasquier et al., 1999c; Stumme et al., 2000; Pei et al., 2000; Zaki & Hsiao, 2002 ; Uno et al. 2003; 2005), les représentations par motifs maximaux (Zaki et al., 1997; Lin & Kedem, 1998 ; Lin et Kedem 1998; 1998 ; Burdick et al., 2005), les représentations par motifs non-dérivables (Calders et Goethals 2002; 2007) et les représentations par ensembles libres (Boulicaut et al., 2003).

Les approches citées ci-dessus sont marquées par leur effort algorithmique pour la réduction du temps de calcul de l'étape d'extraction des motifs intéressants. Ils existe d'autres approches, permettant une réduction sans perte d'information, reposent sur un ensemble de résultats issus de la théorie de l'analyse formelle de concepts (AFC) introduite par Wille (2009). Le principe de ces approches est tout d'abord de déterminer l'ensemble minimal de règles d'association présentées à l'utilisateur, tout en maximisant la quantité d'informations utiles véhiculée ; puis de disposer d'un mécanisme d'inférence qui, suite à la demande de l'utilisateur, permet de retrouver le reste des règles d'association tout en déterminant avec exactitude leur support et leur confiance sans accéder à la base de données (Pasquier, 2000; Gasmi et al., 2006). Dans ce contexte, de nombreux algorithmes de fouille de règles d'association basées sur les treillis des concepts ont été proposés, comme par exemple Touch et Talky-G (Szathmary et al., 2009). Il faut noter que l'AFC est également utilisée dans les représentations condensées closes des motifs citées ci-dessus.

Nous pouvons noter également une autre gamme d'algorithmes s'appuyant sur l'architecture des processeurs multi-cœurs, qui proposent des optimisations basées à la fois sur la

réduction de la base de données et sur le parallélisme multi-threads. A titre d'exemple, nous pouvons citer les travaux des thèses de Negrevergne (2011) et de Quintero Flores (2013).

Dans cette thèse, nous ne nous intéressons pas à l'optimisation d'algorithmes d'extraction de motifs fréquents basée d'une part sur la réduction du nombre de motifs extraits et d'autre part sur la réduction du temps de leur extraction. Nous n'avons pour cette raison pas détaillé la palette d'algorithmes cités ci-dessus.

1.1.2 Cas numérique : règles d'association quantitatives

Le problème originel de la recherche de règles d'association consiste à extraire des corrélations à partir de données binaires. Or les bases de données réelles contiennent non seulement des variables catégorielles mais aussi des variables numériques, discrètes ou pseudo-continues. Les règles d'association classiques ne peuvent donc pas leur être appliquées directement et ont été étendues aux *règles d'association quantitatives* (Agrawal & Srikant, 1994; Miller & Yang, 1997) qui visent à exprimer des corrélations pour de telles données.

Pour les variables catégorielles, chaque valeur possible est considérée individuellement comme un item dont on note la présence ou l'absence. Pour les variables numériques, on ne peut considérer chaque valeur individuelle, car son nombre d'occurrences serait trop faible. Il est alors nécessaire d'effectuer une discrétisation afin d'identifier autant d'intervalles de valeurs qui définissent un attribut catégoriel que d'items. Un item est défini comme un couple constitué d'un attribut avec un intervalle, par exemple (âge, $[30, 45]$). Il est alors possible de calculer la proportion de données possédant un item pour évaluer son support, et donc d'appliquer des algorithmes classiques d'extraction de motifs.

Les approches proposées pour cette extraction dans la littérature peuvent être classées en trois catégories principales que nous détaillons ci-dessous et comparons ensuite dans le tableau 1.2, page 28. Nous commençons par les approches basées sur une discrétisation préalable, puis les approches guidées par des schémas de règles, et enfin les approches fondées sur un algorithme génétique.

Approches fondées sur une discrétisation préalable

Parmi les méthodes les plus standard de discrétisation, on peut citer la discrétisation en k intervalles de même largeur ou de même fréquence ou encore en intervalles non réguliers qui s'appuient sur des connaissances du domaine (Agrawal & Srikant, 1994; Lent et al., 1997; Miller & Yang, 1997).

La difficulté de ces méthodes de découpage réside dans le choix du nombre d'intervalles et dans la disponibilité des connaissances a priori pour les intervalles non réguliers : en se basant sur des intervalles trop petits, on risque d'omettre des règles pour insuffisance de support, alors que si les intervalles sont trop grands, c'est par défaut de confiance qu'on est susceptible de les manquer.

Approche guidée par des schémas de règles

Cette approche proposée par Fukuda et al. (1996a; 1996b) n'identifie pas une discrétisation des attributs numériques mais extrait des intervalles particuliers, dits intervalles d'intérêt, qui satisfont des contraintes de pertinence : elle est basée sur une optimisation de critères de qualité mesurant cette pertinence. L'évaluation des intervalles d'intérêt candidats dépend de la qualité des règles qu'ils induisent selon, par exemple, le support, la confiance ou le gain (Fukuda et al. 1996a; 1996b).

Afin de limiter le coût de calcul, certaines approches sont fondées sur des schémas de règles restreints : un schéma de règle est une règle présentant dans chacun de ses membres gauche et droit des items catégoriels aux valeurs fixées et des items numériques dont les intervalles correspondants ne sont pas encore instanciés (Fukuda et al. 1996a; 1996b), limitant par exemple le nombre d'attributs numériques dans la prémisse et la conclusion. Un tel schéma de règle peut être illustré par l'exemple « âge $\in [v_1, v_2]$ et (région = sud) $\rightarrow$ (salaire = moyen) » où les attributs région et salaire sont instanciés et les valeurs v_1 et v_2 de l'intervalle correspondant à l'attribut âge ne sont pas encore instanciées.

Aumann et Lindell (2003) et Webb (2001) proposent une autre vision du problème : des statistiques (moyenne, variance, écart-type, minimum etc.) sur des distributions des attributs numériques sont autorisées dans la partie droite d'une règle. Deux sortes de règles sont considérées : la première représente le cas où la prémisse de la règle est un ensemble d'attributs catégoriels et son conséquent un ensemble de statistiques sur les distributions de plusieurs attributs numériques ; la deuxième représente le cas où la prémisse de la règle contient un seul attribut numérique et son conséquent une statistique sur la distribution d'un seul attribut numérique. Ces règles peuvent être illustrées par l'exemple : « région = centre $\rightarrow$ salaire : moyenne = 1200 euros » par mois. Cette approche est intéressante, mais elle contraint la forme des règles.

Approche reposant sur un algorithme génétique

L'optimisation est également la voie choisie dans les travaux de Mata et al. (2002), qui propose d'utiliser des algorithmes génétiques : un individu est représenté par une liste de couples (attribut numérique, disjonction d'intervalles). La qualité des individus est évaluée par une mesure permettant d'optimiser le support des motifs, tout en veillant à ne pas retenir les domaines entiers des attributs numériques et à favoriser les motifs les plus spécifiques. Le seul critère optimisé dans cet algorithme est le support, ce qui limite l'applicabilité d'une telle approche.

Une autre approche basée sur l'algorithme génétique suivant une organisation classique a été proposée par Nortet et al. (2006; 2013). Contrairement à l'approche précédente qui optimise un seul critère qui pourrait être insuffisant, celle-ci cherche le meilleur intervalle pour chaque attribut optimisant le support, la confiance, ainsi que la mesure de gain. Cette approche est basée également sur les schémas de règles, et la discrétisation obtenue varie donc pour chaque schéma tout en dépendant des attributs catégoriels et numériques qui le composent.

Approche	Discrétisation	Optimisation	Limites
Discrétisation préalable	discrétisation complète	deux étapes	• perte d'information
Schémas de règles	intervalle d'intérêt	une seule étape	• schémas de règles • format très limité
Algorithme génétique	intervalle d'intérêt	une seule étape	• schémas de règles limité à un intervalle • nombreux paramètres

Tableau 1.2 – Comparaison des trois principales catégories d'extraction de règles d'association quantitatives

Cet algorithme contient plusieurs paramètres à fixer : la taille de la population, le nombre de générations, les taux de mutation et de croisement. En outre, il n'est pas capable d'identifier plusieurs intervalles pertinents.

Synthèse

Ce paragraphe synthétise les travaux précédemment cités, classés selon différents critères listés dans le tableau 1.2 et présentés ci-dessous.

— Les méthodes effectuent-elles une discrétisation complète de l'univers pour identifier les intervalles souhaités ou extraient-elles seulement des intervalles d'intérêt ?

— Les méthodes utilisent-elles une optimisation en une seule étape ?

— La troisième ligne indique les limites de ces méthodes.

Les méthodes reposant sur une discrétisation a priori des attributs quantitatifs optimisent les intervalles d'intérêt en deux étapes, dont une étape de pré-discrétisation préalable induisant une perte d'information. Les approches basées sur les schémas de règles permettent quant à elles d'optimiser les intervalles d'intérêt en une seule étape, pendant la phase de génération des motifs fréquents, mais dans ce cas, le format des règles est souvent très limité. Les approches basées sur les algorithmes génétiques optimisent également les intervalles d'intérêt en une seule étape. Cependant, cette étape d'optimisation n'est pas effectuée pendant la phase de génération des motifs fréquents, mais pendant la phase de génération de règles. L'ensemble de ces approches sont limitées à l'identification d'un seul intervalle d'intérêt : elles n'utilisent pas la disjonction d'intervalles, comme cela est le cas dans les approches basées sur les schémas de règles. Elles reposent de plus sur plusieurs paramètres, pour lesquels il n'est pas aisé de trouver les valeurs optimales.

1.1.3 Cas flou

Alors que l'extension des règles d'association précédente prend en compte des données numériques, une autre extension vise à traiter des données floues, c'est-à-dire des données

dont les attributs sont des variables linguistiques associées à des modalités floues. Considérons par exemple un attribut correspondant au salaire d'un employé. Dans le cas classique, cet attribut est décrit par des valeurs numériques. Dans le cas flou, il peut être associé à trois modalités floues « faible », « moyen », et « élevé ». L'attribut est ensuite décrit par ses degrés d'appartenance à ces modalités. Un exemple d'une telle règle étendue est : « les employés jeunes et de faible niveau d'études ont des salaires faibles » où « employés », « niveau d'études » et « salaires » représentent les variables linguistiques et « jeune » et « faible » représentent leurs modalités floues respectives.

Définition 1.6 (Règle d'association floue)**.** Une règle d'association floue est de la forme générale $M_1 \rightarrow M_2$ avec $M_1 = (X, A)$ et $M_2 = (Y, B)$ où X, Y sont des attributs flous et A, B sont leurs modalités floues respectives.

Les règles d'association floues sont interprétées comme une généralisation des règles d'association appliquées à des données floues, indiquant que la présence floue de M_1 implique, au sens de la logique, la présence floue de M_2 (Hüllermeier, 2001). Ainsi, la règle « plus on est proche du mur, plus on freine fort » peut être considérée comme l'extension floue d'une règle d'association concernant la présence binaire des attributs distance au mur et freinage.

Le support de la règle est alors calculé comme la somme des contributions de chaque objet à l'implication : une règle est valide si les degrés d'appartenance aux modalités floues impliquées dans la règle satisfont l'implication floue, pour chaque objet de la base de données individuellement.

Définition 1.7 (Support d'une règle d'association floue)**.** Formellement,

$$Supp(M_1 \rightarrow M_2) = \sum_{o \in \mathcal{D}} i(M_1(o), M_2(o)) \tag{1.4}$$

où i est un opérateur d'implication résiduel, par exemple l'implication de Goguen définie par : $i(a, b) = \min(1, b/a)$ si $a \neq 0$, 1 sinon.

Bosc et al. (2001) et Hüllermeier (2001) proposent aussi d'étendre le concept de découverte de règles d'association, de façon à prendre en compte des propriétés graduelles et d'exprimer une contrainte sur les valeurs des attributs apparaissant dans la règle.

On peut noter que ce support ne s'applique pas à un motif, comme dans le cas des règles d'association classiques, mais à une règle, d'une façon asymétrique qui permet de distinguer $M_1 \rightarrow M_2$ de $M_2 \rightarrow M_1$.

Dans la littérature, différentes formes de règles graduelles sont distinguées (Dubois & Prade, 1992; Hüllermeier, 2001) suivant le type d'opérateur d'implication utilisé :

— les r-implications modélisant les règles graduelles floues de la forme « plus X est A, alors plus Y est B » ;

— les s-implications modélisant les règles floues de certitude de la forme « plus X est A, alors plus il est certain que Y est B », par exemple, « plus on se réveille tard, plus on est sûr d'être en retard ».

Un autre type de règle, représentant un croisement de règles d'association floues et de résumés linguistiques a été proposé par Bosc et al. (2001). Il consiste à faire reposer l'interprétation d'une règle d'association floue sur un calcul de cardinalités floues. Le principe est le suivant : comme dans le cas usuel, la validité de la règle $(X, A) \rightarrow (Y, B)$ dépend du nombre de données qui sont A d'une part et du nombre de données qui sont A et B d'autre part. La différence avec le cas usuel est qu'ici, il faut utiliser une cardinalité étendue puisque les ensembles de données considérés sont décrits par des modalités floues. Dans cette approche, la validité est définie comme le degré de nécessité de l'événement « Q données vérifient la règle », où Q désigne un quantificateur flou tel que « la plupart » ou « très peu ».

1.1.4 Autre extension : motifs séquentiels

Les extensions présentées ci-dessus considèrent des données différentes de celles considérées dans le cas classique. Il est important de noter qu'il existe une autre extension qui considère le même type de données que celles traitées dans le cas classique, mais enrichies par un attribut temporel. Il s'agit des motifs séquentiels : l'idée est de fouiller non plus les corrélations entre sous-ensembles de motifs, mais de fouiller les ordres répétitifs entre motifs.

Un motif séquentiel est défini comme une liste ordonnée et non vide de motifs (Agrawal & Srikant, 1995). De tels motifs sont par exemple de la forme : « les clients achètent du pain et du beurre, puis plus tard ils achètent du chocolat ».

De nombreux algorithmes efficaces ont été proposés pour extraire de tels motifs graduels tels que ceux proposés dans les travaux (Agrawal & Srikant, 1995; Masseglia et al., 1998; Zaki, 2001; Ayres et al., 2002; Pei et al., 2004; Chiu et al., 2004; Zaki & Hsiao, 2005)

1.2 Motifs graduels

Les motifs graduels constituent une variante des règles d'association qui permet également de traiter des données numériques, mais recherche un type de corrélation différent : il ne s'agit pas de co-occurence des items mais de co-variation des valeurs des attributs.

Dans cette section, nous commençons par définir les notions de base que nous utilisons dans la suite de cette thèse, dans le cas des données numériques puis floues, comme données par Berzal et al. (2007), Hüllermeier (2002), Di Jorio et al. (2009) et Bouchon-Meunier et al. (2010). Nous rappelons ensuite les différentes interprétations associées aux motifs et règles graduels, ainsi que leurs méthodes d'extraction et critères de qualité proposés pour leur évaluation.

1.2.1 Définitions et notations

Dans toute la section, on considère un ensemble de données, noté $\mathcal{D}$, constitué de n objets décrits par m attributs.

Id.	Vitesse V	Distance D	Freinage F
1	91	3400	1
2	95	2200	0
3	112	2000	2
4	104	1850	3
5	82	5000	2
6	95	1200	1
7	88	1850	5
8	98	1200	4

Tableau 1.3 – Exemple d'une base de données numériques

Cas de données numériques

Définition 1.8 (Item graduel)**.** Un *item graduel* est défini comme un couple constitué d'un attribut et d'une variation, notée $* \in \{\leq, \geq\}$, qui représente un opérateur de comparaison : un item graduel A^* représente le fait que les valeurs de l'attribut augmentent si $* = \geq$ ou diminuent si $* = \leq$.

Deux types de variations pour un item I sont distingués :

— une variation croissante de valeurs d'attributs, c'est-à-dire que la valeur augmente d'un objet à l'autre. Dans ce cas, l'item graduel est sémantiquement identifié comme « plus I est élevé » ou encore « plus I augmente ».

— une variation décroissante de valeurs d'attributs, c'est-à-dire que la valeur diminue d'un objet à l'autre. Dans ce cas, l'item graduel est sémantiquement identifié comme « moins I est élevé », ou « plus I est faible », ou « moins I augmente », ou encore « plus I diminue »

Les items graduels « plus I est élevé » et « moins I est élevé » peuvent être notés de différentes manières. Dans certains travaux comme par exemple les travaux de Dubois et al. (1995); Berzal et al. (2007) et Fiot et al. (2008), ils sont notés $I^{>}$ et $I^{<}$ en utilisant les opérateurs de comparaison stricts $>$ et $<$, alors qu'ils sont notés $I^{\geq}$ et $I^{\leq}$ en utilisant les opérateurs de comparaison larges $\geq$ et $\leq$ dans d'autres travaux, comme par exemple les travaux de Di Jorio et al. (2008; 2009) et de Laurent et al. (2009). Dans cette thèse, nous utilisons les opérateurs de comparaison larges $\{\geq, \leq\}$.

Ces principes peuvent être illustrés par l'exemple de la base de données numériques présentée dans le tableau 5.4 qui décrit $n = 8$ camions en mouvement selon $m = 3$ attributs : leur vitesse, leur distance à un mur et la force de freinage.

Six items graduels peuvent être considérés : $V^{\leq}$, $V^{\geq}$, $D^{\leq}$, $D^{\geq}$, $F^{\leq}$ et $F^{\geq}$, représentant respectivement (plus la vitesse est élevée), (moins la vitesse est élevée), (plus la distance est élevée), (moins la distance est élevée), (plus le freinage est fort) et (moins le freinage est fort).

Définition 1.9 (Motif graduel)**.** Un *motif graduel*, ou *itemset graduel*, noté $\{(A_i, *_i), i = 1...k\}$ ou $\{A_i^{*i}, i = 1...k\}$, est défini comme une combinaison de plusieurs items graduels et interprété sémantiquement comme leur conjonction.

Par exemple $M = V^{\geq}D^{\leq}$ est interprété comme « plus la vitesse est élevée et moins la distance est élevée ». Ceci impose une contrainte de variation de plusieurs attributs simultanément.

La longueur d'un motif graduel, notée k, est le nombre d'attributs qui y sont impliqués.

Définition 1.10 (Règle graduelle)**.** Une *règle graduelle*, notée $M_1 \rightarrow M_2$, est définie comme une paire de motifs graduels (M_1, M_2) sur laquelle est imposée une relation de causalité ; M_1 est appelé l'*antécédent* ou *prémisse*, M_2 le *conséquent*.

Une règle graduelle établit des relations de causalité entre les attributs et résume les tendances observées dans l'ensemble des données. Notons que c'est cette causalité qui fait la différence entre une règle et un motif.

A partir du tableau 5.4, nous pouvons extraire la règle $D^{\leq} \rightarrow F^{\geq}$ qui est lue comme « plus on est proche du mur alors plus on freine fort ».

Cas flou

La plupart des travaux existants sur les motifs graduels (Hüllermeier 2001; 2002 ; Berzal et al., 2007 ; Di Jorio et al. 2008; 2009 ; Laurent et al., 2009) s'appliquent à des données floues. Pour de telles données, les attributs sont associés à des modalités floues et les données sont décrites par leurs degrés d'appartenance à ces modalités. Ainsi, dans le cas de l'exemple des camions décrit dans le tableau 5.4, on peut obtenir une base de données floues en considérant que la vitesse, la distance et le freinage sont associés à des variables linguistiques ;la vitesse est associée à 3 modalités : lente, normale et élevée, la distance au mur à 2 modalités : proche et loin, et le freinage à 3 modalités : faible, normal et fort.

Les données sont ensuite décrites par leur degré d'appartenance aux modalités comme : par exemple, la vitesse de l'objet 1 appartient avec un degré $0,2$ à la modalité lente de l'attribut vitesse, avec un degré $0,3$ à la modalité normale et avec un degré $0,5$ à la modalité élevée.

Il faut noter qu'on peut avoir plus de deux modalités de degrés d'appartenance supérieurs à 0.

En notant $V(o)$ la valeur numérique de l'attribut « vitesse » décrivant un objet o, m et M représentent respectivement la valeur minimale et maximale décrivant l'attribut « vitesse », on définit formellement les deux fonctions d'appartenance aux modalités floues « élevée » et « lente » par les fonctions $\mu_{\text{élevée}}$ et μ_{lente} de la manière suivante :

$$\mu_{\text{élevée}}(o) = \begin{cases} 0 & \text{si } V(o) = m \\ 1 & \text{si } V(o) = M \\ \dfrac{A(o) - m}{M - m} & \text{si } m < V(o) < M \end{cases}$$

Id.	Vitesse			Distance		Freinage		
	lente	normale	élevée	proche	loin	faible	normal	fort
1	0.2	0.3	0.5	0.4	0.6	0.6	0.4	0.2
2	0.2	0.2	0.6	0.5	0.5	0.2	0.7	0.1
3	0	0.1	0.9	0.7	0.3	0	0.6	0.4
4	0	0.2	0.8	0.8	0.2	0.1	0.3	0.6
5	0.1	0.7	0.3	0.3	0.7	0.3	0.3	0.4
6	0.2	0.3	0.5	0.9	0.1	0.5	0.3	0.2
7	0	0.6	0.4	0.8	0.2	0.1	0.1	0.8
8	0.1	0.2	0.7	0.9	0.1	0	0.3	0.7

Tableau 1.4 – Exemple de base de données floues

$$\mu_{\text{faible}}(o) = \begin{cases} 0 & \text{si } V(o) = M \\ 1 & \text{si } V(o) = m \\ 1 - \dfrac{V(o) - m}{M - m} & \text{si } m < V(o) < M \end{cases}$$

Nous rappelons ci-dessous les définitions d'item graduel, motif graduel et règle graduelle dans le cas de telles données floues (Di Jorio et al., 2009; Laurent et al., 2009; Bouchon-Meunier et al., 2010).

Définition 1.11 (Item graduel flou)**.** Un *item graduel flou* est un *triplet* $(X, A, *)$ constitué d'un attribut X, une de ses modalités A et une variation, notée $* \in \{\geq, \leq\}$.

Un item graduel flou peut être illustré par l'exemple (vitesse, lente, $\geq$). Il est interprété comme « plus la vitesse est lente », ou plus précisément « plus le degré d'appartenance de la vitesse à lente est élevée ».

Il faut noter que les items graduels flous peuvent être représentés dans le même formalisme que les items graduels : il faut pour cela introduire un attribut pour chaque modalité floue, dont les valeurs sont les degrés d'appartenance. On peut ainsi créer, dans l'exemple précédent, trois attributs vitesseLente, vitesseNormale et vitesseElevée dont les valeurs sont les degrés d'appartenance. L'item graduel flou précédent peut alors être écrit $vitesseLente^{\geq}$.

Il est cependant important de souligner une différence sémantique entre les deux types d'items, que nous illustrons en considérant l'exemple de l'item graduel « plus la vitesse est élevée ». Dans le cas non flou il exprime une contrainte d'ordre sur tout l'univers des vitesses, défini par exemple sur $[0, 120]$. Dans le cas flou, « élevée » est associé à une modalité floue, par exemple de noyau $[110, 120]$ et de support $[100, 120]$. Le motif graduel flou s'applique aux degrés d'appartenance, dans l'univers $]0, 1]$. Il ne fait donc intervenir que les vitesses supérieures à 100, restreignant les données qui supportent le motif, et lui donne une interprétation plus locale.

Dans la suite du document, nous utilisons la notation $A^{\leq}$ et $A^{\geq}$ pour les deux cas de données numériques et de données floues. Pour tout o appartenant à $\mathcal{D}$, $A(o)$ désigne la valeur

de l'attribut A pour l'objet o ou le degré d'appartenance de o à l'attribut A qui représente l'attribut flou et la modalité considérée dans le cas de données floues.

Les notations de motifs et règles graduels flous sont alors identiques à celles des données numériques, bien que basées sur les items graduels flous.

Il existe également une autre différence théorique entre données numériques et données floues : pour un motif graduel flou $M = A_j^{*j}, j = 1...k$, si tous les sens de variation sont identiques, un degré d'appartenance au motif peut être défini comme $M(o) = \top_{j=1...k}(A_j(o))$ où $\top$ désigne une t-norme, puisque le motif est interprété comme une conjonction des items qu'il contient. Dans le cas de données numériques, l'agrégation des valeurs de plusieurs attributs impliqués dans le motif graduel est potentiellement plus problématique, et sa sémantique doit être examinée en fonction des données considérées.

Il est important de noter que la gradualité indiquée des règles d'association dans le cas flou (voir section 1.1.3) est particulière et différente de la gradualité indiquée dans cette section. En effet, ici la gradualité exprime une tendance globale à travers l'ensemble de données : elle s'applique à un sous-ensemble d'objets de manière transversale. Au contraire, l'interprétation de la gradualité décrite dans la section 1.1.3 s'applique à chaque objet individuellement : elle considère qu'une présence (floue) implique au sens flou une présence (floue), et que chaque objet a une contribution individuelle avec son propre degré d'appartenance. Une telle gradualité est interprétée par les différentes lignes de l'ensemble de données.

Extensions séquentielles

Certains travaux étendus des motifs graduels visent à prendre en compte la notion de temporalité et l'ordre des événements dans le but d'extraire des motifs graduels flous séquentiels (Fiot et al. 2008; 2009).

Rappelons qu'un motif séquentiel, contrairement aux règles d'association, décrit la fréquence de certains comportements successifs. Il est représenté par une liste de motifs ordonnés. L'ordre est associé à une mesure du temps et la gradualité peut être appliquée à deux niveaux :

- au niveau des items, ce qui traduit une co-variation dans le même sens entre plusieurs items. Par exemple, considérons l'attribut « salaire », un tel motif peut être illustré par « le salaire augmente au cours du temps ».
- au niveau de la relation entre les motifs, ce qui introduit les notions de « puis rapidement », « après une période de temps très courte », etc. Ainsi, la connaissance de la forme « Quand la vitesse d'un moteur augmente fortement, après une période de temps très courte, la vitesse du camion augmente légèrement pour une courte période » représente un tel motif.

1.2.2 Interprétation comme covariation de valeurs d'attributs

Dans cette section et les suivantes, nous nous intéressons aux différentes interprétations des motifs graduels ainsi qu'aux approches permettant leur extraction, synthétisées dans le tableau 1.5.

Approches	Principe d'extraction	Type de données	Critères d'évaluation	Référence
Hüllermeier (2002)	Régression linéaire	Floues	Qualité et coefficients de régression	section 1.2.2, page 34
Berzal et al. (2007)	Dépendance graduelle avec considération des variations des degrés entre deux objets	Floues	Pourcentage de couples d'objets vérifiant une variation	section 1.2.3, page 36
Laurent et al. (2009)	Extraction de couples concordants d'objets respectant la gradualité	Numériques	τ de Kendall	section 1.2.3, page 36
Di Jorio et al. (2008)	Heuristique basée sur les ensembles de conflits	Numériques	Cardinalité de la liste maximale d'objets respectant la gradualité	section 1.3.2, page 40
Di Jorio et al. (2009)	Recherche exhaustive basée sur les graphes de précédence			section 1.3.3, page 41

Tableau 1.5 – Approches traitant la gradualité : extraction de motifs/règles graduels.

Une première approche des règles graduelles floues comme contrainte de covariation de valeurs interprète une règle $A \rightarrow B$ comme une contrainte imposant qu'une augmentation des valeurs de l'attribut A s'accompagne d'une augmentation des valeurs de l'attribut B (Hüllermeier, 2002).

Afin d'identifier de telles règles, Hüllermeier (2002) propose d'effectuer une régression linéaire des valeurs des attributs, en appliquant la méthode des moindres carrés aux couples $(A(o), B(o))$ pour tous les objets o.

Cette définition et cette méthode d'extraction s'appliquent aux paires d'attributs. L'extension proposée de cette définition à des motifs de longueur supérieure à 2 considère le cas de données floues : l'extension exploite ce cadre de logique floue ainsi que le fait que des motifs soient interprétés comme conjonction des items qu'ils contiennent. Comme souligner précédemment, un degré d'appartenance à un motif peut être calculé en utilisant une t-norme, appliquée aux degrés d'appartenance des items du motif considéré. La tendance graduelle est alors comprise comme une contrainte de covariation entre les degrés d'appartenance agrégés. Ainsi des motifs de longueur supérieure à 2 peuvent être traités comme des motifs de longueur 2. Il faut noter que cette approche ne s'applique que pour les données floues : en effet, dans le cas des données non floues, la valeur associée à un motif graduel ne peut pas être calculée par agrégation.

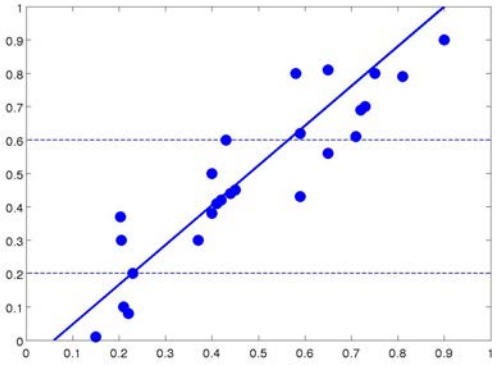


Figure 1.1 – Exemple de règle graduelle : l'abscisse représente les degrés d'appartenance à A et l'ordonnée les degrés d'appartenance à B

La validité de la règle est évaluée à partir de la qualité de la régression, mesurée par le coefficient de corrélation linéaire R^2, conjointement avec la pente de la droite de régression. Ainsi les attributs qui ne sont pas suffisamment corrélés sont rejetés, de même que ceux pour lesquels le degré d'appartenance de A reste constant alors que celui de B varie, ou inversement.

Considérons ainsi l'exemple illustré sur la figure 1.1. Comme on peut le voir sur la figure de gauche, il existe une forte corrélation entre A et B. La pente positive de la droite de régression propose en fait la tendance suivante : « plus A est élevé, plus B est élevé » avec un support $R^2 = 0.77$.

1.2.3 Interprétation comme contrainte d'ordres induits

Principe

Alors que l'approche précédente repose sur les valeurs numériques des attributs, d'autres approches considèrent la contrainte de covariation en terme de corrélation d'ordres (Berzal et al., 2007) : elles imposent que les classements des données induits par chacun des attributs intervenant dans le motif soient identiques. Ainsi, dans le cas d'un motif de taille 2, la règle $A^{\geq} \rightarrow B^{\geq}$ est valide si $\forall o, o' \in \mathcal{D}, A(o) < A(o')$ implique $B(o) < B(o')$; dans le cas de règles telles que la règle $A^{\leq} \rightarrow B^{\geq}$, la contrainte impose que les ordres induits soient inversés.

Définition 1.12 (Ordre induit par un motif)**.** L'ordre $\preceq_M$ induit par un motif $M = \{(A_j, *_j), j = 1..k\}$ est défini comme :

$$o \preceq_M o' \text{ ssi } \forall\ j \in [1, k]\ A_j(o) *_j A_j(o') \tag{1.5}$$

Formalisation comme transposition de règles d'association

Berzal et al. (2007) proposent une approche d'extraction automatique de telles règles graduelles, formulée comme la découverte de règles d'association dans un ensemble de tran-

sactions $\mathcal{D}'$ dérivé de l'ensemble de données initiales $\mathcal{D}$: chaque paire d'objets dans $\mathcal{D}$ est associée à une transaction t dans $\mathcal{D}'$, qui possède alors un item graduel A^* si la paire correspondante (o, o') de $\mathcal{D}$ satisfait la contrainte imposée par A^*, c'est-à-dire si $A(o) * A(o')$. Une règle graduelle dans $\mathcal{D}$ est donc formulée comme une règle d'association classique extraite de $\mathcal{D}'$ et son support est défini comme le support de la règle d'association correspondante. Avec la définition précédente de la base de transactions, pour un motif graduel $M = \{(A_j, *_j), j = 1..k\}$, ce support s'écrit

$$supp(M) = \frac{|\{(o, o')/o \preceq_M o'\}|}{|\mathcal{D}|(|\mathcal{D}| - 1)} \tag{1.6}$$

Dans le cas des motifs classiques, le support est le quotient entre le nombre d'objets vérifiant le motif par le nombre total d'objets de la base de données ; dans le cas de la gradualité, le support est la proportion de couples d'objets de la base de données vérifiant la contrainte d'ordre imposée par le motif graduel.

La construction de $\mathcal{D}'$ est coûteuse. Berzal et al. (2007) proposent une méthode d'approximation, basée sur la discrétisation des valeurs d'attributs, qui nécessite de garder en mémoire un tableau de taille p^k où p représente le niveau de discrétisation et k la longueur des motifs considérés. Ainsi, la complexité du calcul du support d'un motif de longueur p passe de $\mathcal{O}(n^2)$ dans le cas de la méthode de base (construction de $\mathcal{D}'$) à $\mathcal{O}(n + k^p)$ dans le cas de la méthode d'approximation. Cependant, lorsque des motifs de taille importante sont recherchés, le coût de calcul reste élevé, comme le montrent Berzal et al. (2007) dans des expériences effectuées sur un ensemble de données contenant seulement 6 attributs.

Berzal et al. (2007) formalisent la propriété de complémentarité du support, ce qui permet de diviser par deux l'espace de recherche, car une moitié des règles graduelles peut être automatiquement déduite de l'autre moitié. En effet, elle peut être obtenue en remplaçant simplement le couple d'objets (o, o') par (o', o) dans l'équation (1.6). Cette propriété réduit le temps de calcul de l'approche proposée.

Propriété (complémentarité)**.** *Soit* $c \in \{\leq, \geq\}$ *tel que* $c(\geq) = \leq$ *et* $c(\leq) = \geq$*, alors* $supp(X, A, *) = supp(X, A, c(*))$.

Extension pour la prise en compte d'amplitude

Molina et al. (2007) étendent l'approche de Berzal et al. (2007) afin de mesurer la force de variation d'une règle. Cette extension propose de prendre en compte la variation d'amplitude entre les couples d'objets (voir discussion à la section 1.3.4, page 43) : lors de la construction de $\mathcal{D}'$, au lieu de considérer de manière binaire que la valeur entre deux objets augmente ou diminue, les auteurs quantifient cette variation par la différence des valeurs. Celle-ci fournit une information indiquant dans quelle mesure les contraintes sont satisfaites. Les règles d'association sont ensuite appliquées afin d'extraire des informations de cet ensemble de données. Il faut noter que les données sont numériques et non binaires,

ce qui nécessite une adaptation de l'algorithme d'extraction des règles d'association (Molina et al., 2007). Une mesure adaptée des dépendances floues proposée par Berzal et al. (2005) est alors utilisée à la place du support utilisé par emciteberzal rappelé dans l'équation (1.6).

Formalisation comme corrélation d'ordres

Laurent et al. (2009) considèrent la même définition du support que Berzal et al. (2007) mais en donnent une interprétation différente. En effet, la définition du support de l'équation (1.6) est présentée dans un cadre de règles d'association, comme une généralisation du support classique à une base de transactions dérivée des données initiales. Laurent et al. (2009) l'interprètent dans un cadre de comparaison d'ordres multiples : en considérant que chaque attribut induit un ordre sur les données, ce support peut en effet être considéré comme quantifiant la ressemblance, ou le degré d'accord, entre ces ordres.

La notion de corrélation d'ordres a été étudiée dans le domaine des statistiques et plusieurs mesures ont été proposées : Laurent et al. (2009) proposent d'utiliser le τ de Kendall (Kendall & Babington, 1939), dont la définition est directement liée à la définition de support donnée dans l'équation (1.6). Ils proposent aussi d'utiliser une représentation binaire efficace des paires de données suivant le principe proposé par Di Jorio et al. (2009) : une *matrice de concordance binaire* est définie pour chaque motif $M = \{A_j^{*j}, j = 1..k\}$ telle que la valeur associée au couple (o, o') est 1 si $\forall j \in [1, k] A_j(o) *_j A_j(o'), 0$ sinon.

Cette représentation fournit d'une part une méthode efficace pour passer de motifs de longueur $k-1$ à des motifs de longueur k, puisque la liste des paires d'objets ordonnables pour être en accord avec le motif graduel de longueur k est équivalente à une conjonction logique appliquée aux matrices de concordance correspondant aux deux motifs de longueur $k-1$ considérés. D'autre part, le support d'un motif est obtenu très simplement, comme la somme des éléments de la matrice, divisée par le nombre total de paires d'objets. Les auteurs utilisent donc la mesure de support donnée dans l'équation (1.6) pour évaluer la qualité des motifs graduels extraits.

Cette approche ne nécessite pas d'effectuer une approximation comme celle proposée par Berzal et al. (2007), car elle repose sur une approche par niveau qui permet d'extraire les motifs graduels pertinents de longueur $k+1$ à partir de ceux obtenus au niveau k. L'algorithme proposé par Laurent et al. (2009) est très efficace, car le calcul de support ne nécessite aucune opération de comptage sur l'ensemble de données et il peut être déduit des informations du niveau précédent. En outre, cette information peut être traitée d'une manière efficace aussi, grâce à la représentation binaire des matrices de concordance.

1.3 Motifs graduels par identification de sous-ensembles de données compatibles

Di Jorio et al. (2008; 2009) proposent également une interprétation en termes de corrélation d'ordres. Cependant, plutôt que de rechercher le nombre de paires d'objets respectant

un motif, comme proposé par Berzal et al. (2007) ou Laurent et al. (2009), les auteurs cherchent à identifier des sous-ensembles de données qui satisfont la contrainte d'ordre exprimée par un motif donné et définissent le support de ce dernier comme le nombre maximal de données pouvant être extraites des données initiales et qui satisfont sa contrainte. Un tel sous-ensemble de données compatible avec un motif M est appelé *chemin* (voir définition 1.13 ci-dessous).

Dans ce contexte, deux approches d'extraction automatique ont été proposées par Di Jorio et al. (2008; 2009). La première est une heuristique que nous présentons brièvement dans la section 1.3.2. La seconde, nommée GRITE (GRadual ITemset Extraction), est une approche exacte que nous détaillons dans la section 1.3.3. Notre thèse s'appuie sur cette interprétation des motifs graduels car l'algorithme GRITE identifie de plus l'ensemble des chemins complets maximaux pour tout motif graduel valide dont nous avons besoin (voir section 1.3.3, page 41). Aussi, une section complète dédiée à cette approche est présentée ci-dessous.

Dans cette section, nous présentons tout d'abord dans la section 1.3.1 les notations ainsi que le vocabulaire utilisé dans la suite de ce chapitre, dans la section 1.3.2 la méthode approchée et dans la section 1.3.3 la méthode exacte et l'algorithme GRITE. Nous terminons, dans la section 1.3.4, par une comparaison de toutes les approches d'extraction de motifs graduels présentées dans ce chapitre. Cette comparaison est basée sur la prise en compte de l'amplitude de déviation des objets qui ne satisfont pas la contrainte d'ordre du motif considéré.

1.3.1 Formalisation et définition de chemin

Dans cette section, nous rappelons la notion de chemin ainsi que les définitions nécessaires à la suite du chapitre et plus généralement à l'ensemble de la thèse.

Définition 1.13 (Chemin)**.** Un sous-ensemble $D = \{o_1, ..., o_m\} \subseteq \mathcal{D}$ est un *chemin* supportant le motif $M = \{(A_j, *_j), j = 1..k\}$ s'il existe une permutation π telle que $\forall j \in [1, k], \forall l \in [1, m-1]$, $A_j(o_{\pi_l}) *_j A_j(o_{\pi_{l+1}})$.

Définition 1.14 (Chemin complet)**.** Un chemin est dit *complet* si aucun objet de $\mathcal{D}$ ne peut lui être ajouté sans violer la contrainte d'ordre imposée par M.

On note $\mathcal{L}(M)$ l'ensemble de chemins complets associés à un motif M.

Définition 1.15 (Chemin maximal)**.** Un *chemin maximal* est un chemin complet de longueur maximale.

On note $\mathcal{L}^*(M)$ l'ensemble des chemins maximaux.

Définition 1.16 (Objets compatibles)**.** o et o' sont *compatibles* selon l'ordre induit par le motif M si $o \preceq_M o'$ ou $o' \preceq_M o$ où $\preceq_M$ est défini comme dans la définition 1.12.

Afin d'illustrer ces notions, considérons la base de données contenant $n = 7$ objets décrits par 2 attributs représentés graphiquement sur la figure 1.2.

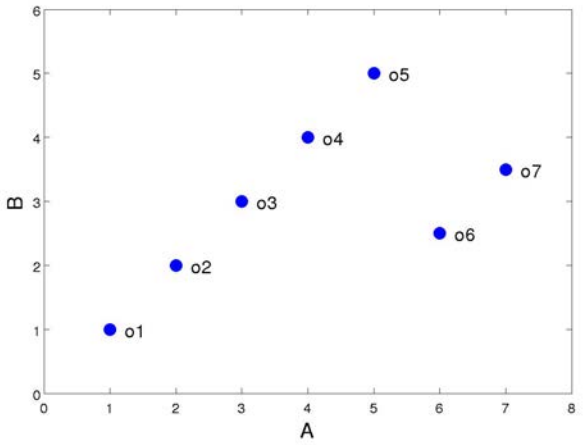


Figure 1.2 – Ensemble de données illustrant les chemins complets et maximaux

Pour le motif $M = A^{\geq}B^{\geq}$ on a $\mathcal{L}(M) = \{\{o_1, o_2, o_3, o_4, o_5\}, \{o_1, o_2, o_3, o_7\}, \{o_1, o_2, o_6, o_7\}\}$. En effet,$\{o_1 \preceq_M o_2 \preceq_M o_3 \preceq_M o_4 \preceq_M o_5\}$, $\{o_1 \preceq_M o_2 \preceq_M o_3 \preceq_M o_7\}$ et $\{o_1 \preceq_M o_2 \preceq_M o_6 \preceq_M o_7\}$. Comme de plus aucun objet ne peut être ajouté à aucune de ces listes sans violer la contrainte d'ordre imposée par M, tous les chemins de $\mathcal{L}(M)$ sont complets. Seul le premier est un chemin maximal de longueur maximale, égale à 5.

Définition 1.17 (Support graduel)**.** Le support d'un motif est défini comme la longueur de ses chemins maximaux rapportée au nombre total d'objets. Formellement :

$$SG(M) = \frac{1}{|\mathcal{D}|} \max_{D \in \mathcal{L}(M)} |D| \tag{1.7}$$

Le support représente donc la proportion maximale d'objets qu'on peut ordonner selon la contrainte d'ordre imposée par M. Pour l'exemple illustré par la figure 1.2, on a $SG(M) = 5/7$.

Définition 1.18 (Validité d'un motif graduel)**.** M est un motif *valide* si $SG(M) \geq s$ où s est un seuil fixé par l'utilisateur. Pour un motif graduel M, on note l'ensemble des chemins *complets valides* $\mathcal{L}_s(M) = \{D \in \mathcal{L}(M)/|D|/|\mathcal{D}| \geq s\}$.

Dans les sections suivantes, nous présentons les deux méthodes d'extraction automatique de motifs graduels qui exploitent la définition du support donnée ci-dessus : une méthode heuristique (Di Jorio et al., 2008), puis une méthode exacte qui présente l'avantage de la complétude (Di Jorio et al., 2009).

1.3.2 Méthode d'extraction approchée

Di Jorio et al. (2008) proposent, pour calculer le support d'un motif graduel, une méthode par niveau heuristique, basée sur des ensembles de *conflits*. Un ensemble de conflits est défini informellement comme les objets qui empêchent un nombre maximal d'autres objets d'être ordonnés.

Il est défini formellement pour un motif de longueur 2 comme suit :

Définition 1.19 (Ensemble de conflits)**.** Soit $M = A_1^{*1}, A_2^{*2}$ un motif graduel et E un sous-ensemble d'objets de $\mathcal{BD}$ ordonnés sur A_1^{*1} selon l'opérateur de comparaison $*_1$ puis sur A_2^{*2} selon l'opérateur de comparaison $*_2$. À chaque objet $o \in E$, un ensemble de conflit $\mathcal{C}_\neg$ est associé tel que $\forall\, o' \in \mathcal{C}_\neg,\ A_2(o) \neg *_2 A_2(o')$.

La méthode consiste à éliminer à chaque niveau les données dont l'ensemble de conflits est maximal, c'est-à-dire qui empêchent le plus grand nombre de données de satisfaire les contraintes d'ordre considérées et de garder une seule liste maximale d'objets ordonnés (un seul chemin maximal). Il s'agit d'une heuristique, car le choix d'une autre donnée à un niveau peut conduire à de meilleurs résultats au niveau suivant. Ce choix n'est pas une tâche triviale, puisqu'il influe sur les supports des motifs de longueur supérieure. Un objet qui satisfait le plus de contraintes lors des futures jointures doit être choisi parmi plusieurs, car lors de la jointure de deux listes candidates, des objets en conflit peuvent apparaître.

L'heuristique proposée est efficace en temps, elle calcule le support dans un processus par niveau et considère les motifs de taille croissante. Cette heuristique adopte une technique « générer-et-élaguer » et utilise un algorithme de génération basé sur Apriori.

1.3.3 Méthode d'extraction exacte : algorithme GRITE

L'inconvénient majeur de la méthode exposée ci-dessus est qu'elle n'est pas complète. En effet, l'utilisation d'une heuristique implique que, dans certains cas, le support d'un motif graduel peut être sous-évalué. Ainsi, si l'utilisateur fixe un seuil minimal légèrement supérieur au support obtenu, et si ce support n'est pas le support réel, ce motif graduel ne sera pas extrait. Ce phénomène s'explique par le fait qu'un choix est fait à chaque fois que plusieurs ensembles de conflits maximaux sont rencontrés. Ainsi, quel que soit le choix, le support calculé sera toujours inférieur ou égal au support réel. Ce phénomène s'applique pour les sur-motifs, et non au niveau local.

Afin de pallier le problème du choix des objets à éliminer dans la méthode approchée décrite ci-dessus, une méthode exacte très efficace, appelée GRITE (GRadual ITemset Extraction), a été proposée par Di Jorio et al. (2009). Elle consiste à conserver tous les choix possibles, c'est-à-dire tous les chemins possibles pour chaque motif graduel. Cette méthode est basée sur les graphes de précédence : les données sont représentées dans un graphe dont les arcs expriment les relations de précédence induites par les attributs impliqués dans les motifs considérés. Un exemple illustratif pour les données du tableau 5.4 est donné dans la figure 1.3.

Ce graphe est représenté par sa matrice d'adjacence, sous la forme d'une matrice binaire $n \times n$: s'il existe une relation d'ordre entre un objet o et un objet o', alors le bit correspondant à la ligne o et à la colonne o' vaut 1, et 0 sinon. Aussi, pour un motif $M = A_1^{*1}, \ldots, A_p^{*p}$ où $*j = \{\geq, \leq\}$ et A_j^{*j} est un item graduel avec $j \in [1, p]$. o précède o' si $\forall j \in [1, p], A_j(o) \leq_j A_j(o')$, le coefficient correspondant à la paire d'objets (o, o') est 1 si $\forall j \in [1, p]$, on a $A_j(o) *_j A_j(o')$, il vaut 0 sinon.

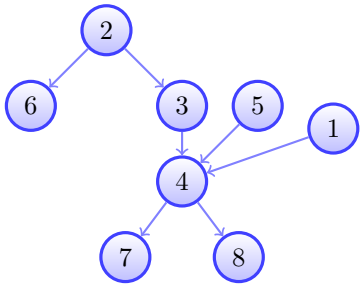


Figure 1.3 – Graphe de précédence correspondant au motif graduel $M = D^{\leq}F^{\geq}$ pour les données du tableau 5.4, page 127

	1	2	3	4	5	6	7	8
1	0	0	0	1	0	0	1	1
2	0	0	1	1	0	1	1	1
3	0	0	0	1	0	0	1	1
4	0	0	0	0	0	0	1	1
5	0	0	0	1	0	0	1	1
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0

Tableau 1.6 – Matrice de concordance du motif $D^{\leq}F^{\geq}$

Le support d'un motif peut ensuite être obtenu à partir de la longueur maximale de ses chemins.

Le principal intérêt de la structure binaire réside dans le fait qu'elle facilite l'opération de jointure. En effet, l'ensemble de toutes les solutions étant mémorisé, il est possible d'effectuer une conjonction logique entre deux matrices binaires. Cela garantit la préservation des ordres communs. Ainsi, cette opération binaire rend l'approche proposée très efficace pour générer des motifs graduels de longueur $k+1$ à partir de motifs de taille k : si M_3 est un motif généré à partir des motifs M_1 et M_2, sa matrice d'adjacence $AdjM_3 = AdjM_1$ & $AdjM_2$ où & est l'opération de ET logique.

La figure 1.3 montre le graphe associé au motif graduel $M = D^{\leq}F^{\geq}$ pour les données du tableau 5.4. Dans ce graphe, deux chemins maximaux vérifient la contrainte d'ordre imposée par M : $D_1 = \{2, 3, 4, 7\}$ et $D_2 = \{2, 3, 4, 8\}$.

Le tableau 1.6 montre la matrice binaire correspondant au motif graduel M.

Les nœuds isolés, qui n'ont ni père, ni fils, sont éliminés du graphe de précédence. Dans la matrice binaire, les nœuds isolés sont les objets pour qui leur ligne et leur colonne sont nulles. Ces objets sont donc supprimés de la matrice afin de réduire la complexité spatiale de la mémoire. Ainsi, pour le motif $D^{\leq}F^{\geq}$, le graphe n'a aucun nœud isolé : sa matrice n'est donc pas réduite.

L'algorithme GRITE (Di Jorio et al., 2009) constitue une méthode efficace d'extraction

de motifs valides, qui associe à chacun l'ensemble des chemins complets maximaux sur lesquels il s'appuie.

1.3.4 Discussion sur le rôle de l'amplitude de déviation

Les méthodes d'extraction présentées dans les sections 1.2 et 5.4.1 diffèrent par leur traitement des amplitudes de déviation par rapport aux motifs graduels considérés. Ceci est illustré sur la figure 1.4 qui représente deux ensembles de données décrits par deux attributs, pour lesquels l'objet o_2 est en contradiction avec la contrainte d'ordre imposée par le motif $A^{\geq}B^{\geq}$. Cependant, la déviation induite par o_2 est moins importante pour l'ensemble de données de gauche, noté $\mathcal{D}_1$, que pour l'ensemble de droite, noté $\mathcal{D}_2$.

La définition de support proposée par Di Jorio et al. (2008; 2009) constitue une différence majeure avec les méthodes d'extraction de motifs graduels présentées dans la section 1.2, concernant les données qui ne satisfont pas la contrainte d'ordre du motif considéré. En effet, les approches basées sur la régression (Hüllermeier, 2002), sur les règles d'association classiques (Berzal et al., 2007) et sur les comparaisons d'ordres (Laurent et al., 2009), prennent en compte l'amplitude de déviation par rapport aux motifs graduels considérés : dans chacun des deux ensembles de données $\mathcal{D}_1$ et $\mathcal{D}_2$, l'objet o_2 empêche le motif graduel « plus A, plus B » d'être complètement vrai, mais dans le deuxième cas, la déviation qu'il amène est beaucoup plus haute. En d'autres termes, il représente plutôt une exception dans le cas de $\mathcal{D}_1$.

Pour la définition par régression (Hüllermeier, 2002), cette différence est reflétée par une pente plus forte pour $\mathcal{D}_1$ que pour $\mathcal{D}_2$. Dans les approches basées sur les règles d'association (Berzal et al., 2007) et sur les comparaisons d'ordres (Laurent et al., 2009), le support est également plus faible pour $\mathcal{D}_2$ que pour $\mathcal{D}_1$: o_2 conduit à un nombre plus important de couples de données qui ne vérifient pas le motif graduel, à savoir (o_2, o_3), (o_2, o_4), (o_2, o_5) et (o_2, o_6), alors que pour $\mathcal{D}_1$, seul le couple (o_2, o_3) contredit le motif.

L'amplitude de déviation est également prise en compte dans les travaux de Molina et al. (2007) et cette différence est reflétée, non pas par les couples contredisant la contrainte d'ordre du motif, mais par la quantification de la variation entre deux objets. Le support est également plus faible pour $\mathcal{D}_2$ que pour $\mathcal{D}_1$.

Au contraire, dans la définition de support proposée par (Di Jorio et al. 2008; 2009), $\mathcal{D}_1$ et $\mathcal{D}_2$ conduisent au même support, puisqu'il suffit dans les deux cas de supprimer l'objet o_2 pour obtenir un ordre parfait suivant la contrainte d'ordre imposée par le motif « plus A, plus B ».

Conclusion

Dans la première partie de ce chapitre, nous avons rappelé les définitions préliminaires à la fouille de données, liées aux règles d'association. Nous avons ensuite illustré diverses variantes

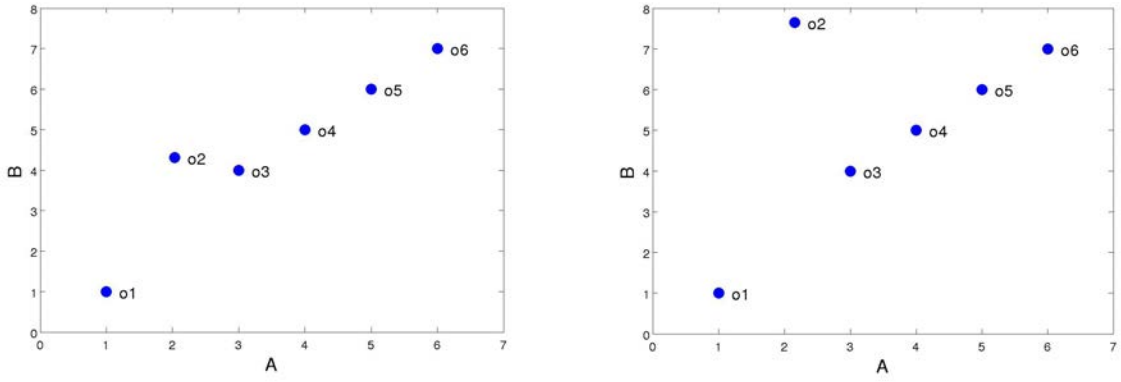


Figure 1.4 – Rôle de l'amplitude de déviation

pour différents types de données, avec des interprétations variées, en considérant tout d'abord le cas binaire qui représente le cas classique, puis leurs extensions aux cas numérique et flou.

La deuxième partie est centrée sur les motifs graduels qui constituent le cadre dans lequel cette thèse s'inscrit. Nous avons d'abord rappelé les notions et définitions liées aux motifs graduels et aux règles graduelles. Nous avons ensuite examiné les interprétations et approches proposées pour leur extraction.

Dans les chapitres suivants, nous proposons d'enrichir ces motifs graduels, en tenant compte de différentes formes de contextualisation : soit par rapport aux autres motifs pour gérer le problème éventuel de motifs graduels contradictoires, soit par complément d'information associant aux motifs extraits une meilleure interprétation et un nouveau contexte qui les rend plus vrais.

Dans la fouille de données, la contrainte d'interprétabilité des connaissances quelle que soit leur forme est primordiale. Nous proposons dans cette thèse des approches permettant de travailler sur cette contrainte ainsi que sur la sémantique des motifs graduels, proposant une approche permettant de traiter le problème de motifs graduels contradictoires et différents modes d'enrichissement des motifs graduels.

2

Renforcement par un nouvel attribut : nouveaux critères et extension

Sommaire

Introduction

Dans ce chapitre, nous nous intéressons à l'enrichissement de motifs graduels selon le principe du renforcement proposé par Bouchon-Meunier et al. (2010) pour des données floues, qui consiste à ajouter une clause linguistiquement introduite par « *d'autant plus que* ». Ce renforcement est interprété comme une validité accrue, ce qui signifie que, quand les données sont restreintes à celles satisfaisant la clause de renforcement, la validité du motif

doit augmenter. Un tel motif graduel enrichi peut être illustré par l'exemple « plus on est proche du mur, plus on freine fort, d'autant plus que la vitesse est élevée » : les informations sur la vitesse permettent d'enrichir par une précision supplémentaire la relation établie entre la distance par rapport au mur et le freinage. Cela fournit une interprétation plus riche à la connaissance extraite.

Ce chapitre est organisé comme suit : dans la section 2.1, nous rappelons la définition du renforcement des motifs graduels, les différentes interprétations envisagées par les auteurs et les critères de qualité des motifs graduels flous renforcés. Dans la section 2.2, nous présentons l'étude complémentaire que nous avons réalisée, portant sur l'extraction de ces motifs par filtrage, ainsi que la définition de nouveaux critères de qualité et leur analyse. Enfin, dans la section 2.3, nous étendons notre étude complémentaire en étudiant la transposition de la notion de renforcement aux règles d'association classiques : nous introduisons les règles d'association renforcées et nous discutons de leurs interprétations possibles et de l'apport. Nous définissons ensuite les critères de qualité qui peuvent être utilisés pour mesurer leur pertinence. Enfin, nous montrons que leur apport est limité, en étudiant leur validité par rapport aux règles d'association classiques.

2.1 État de l'art : enrichissement proposé pour des données floues

Dans cette section, nous rappelons le formalisme ainsi que les différentes interprétations, proposées par Bouchon-Meunier et al. (2010), pouvant être associées aux motifs graduels renforcés, puis les critères de qualité qui mesurent leur qualité.

2.1.1 Définition et exemple

Définition 2.1 (Motif graduel renforcé)**.** Les motifs graduels flous renforcés sont des motifs graduels enrichis par une clause de renforcement introduite par l'expression linguistique « *d'autant plus que* ». Ils sont formellement représentés sous la forme $M_1; M_2$ où M_1 est un motif graduel flou et M_2 est un motif flou classique non graduel qui représente la clause de renforcement.

Cette dernière, composée d'attributs flous, permet d'enrichir la relation existante entre les attributs du motif M_1.

Il est important de noter que cet enrichissement est appliqué aux données floues qui sont considérées de nature numérique par le motif graduel à enrichir et de nature présentielle floue par la clause de renforcement, dans la mesure où la présence de cette clause supplémentaire conduit à une restriction des données définie par la présence de valeurs spécifiques possédant la clause de renforcement.

Considérons par exemple les données artificielles présentées dans le tableau A.3 donné en annexe en page 147, illustrées sur la figure 2.4, page 59. Les abscisses représentent la modalité

floue « proche » associée à l'attribut flou « distance de mur », les ordonnées la modalité floue « fort » associée à l'attribut flou « freinage » et la taille des objets la modalité floue « élevée » associée à l'attribut flou « vitesse ». On peut remarquer que 6 objets parmi 8, représentés en bleu, vérifient le motif graduel flou « plus on est proche du mur, plus on freine fort ». On peut observer aussi que les 2 objets qui ne vérifient pas le motif, représentés en rouge, n'appartiennent pas suffisamment à la modalité floue « élevée » qui décrit la « vitesse » et leur taille est très petite par rapport aux objets vérifiant le motif graduel. Celui-ci peut donc être enrichi par l'information concernant la taille importante des objets qui le vérifient. Cela motive l'extraction du motif graduel renforcé « *plus on est proche du mur, plus on freine fort, d'autant plus que la vitesse est élevée* » : les informations sur la vitesse permettent d'enrichir la relation établie entre la distance du mur et le freinage par une précision supplémentaire.

2.1.2 Interprétations de motifs graduels renforcés

Bouchon-Meunier et al. (2010) ont proposé plusieurs interprétations qui peuvent être envisagées pour le renforcement, selon l'interprétation choisie pour les motifs graduels, en considérant principalement le cas où les motifs graduels sont interprétés comme des contraintes de co-variation.

Tout d'abord, la différence entre les motifs graduels renforcés et les règles conjonctives de la forme $(M_1 \wedge M_3) \rightarrow M_2$ a été soulignée. Avec l'exemple illustratif considéré, cette règle serait « *plus on est proche du mur et plus la vitesse est élevée, alors plus on freine fort* ». La sémantique d'une telle règle est différente du motif graduel renforcé considéré. En effet, dans les règles graduelles conjonctives, M_3 joue un rôle causal sur M_2 et dans l'interprétation de co-variation des motifs graduels flous, il impose une contrainte forte : il faut que les ordres induits par les trois motifs M_1, M_2 et M_3 soient identiques ou très corrélés. Au contraire, dans le cas du renforcement, la contrainte d'ordre ne concerne que les deux motifs graduels M_1 et M_2. L'effet de renforcement s'applique au motif composé de M_1 et M_2 et non pas sur M_2 seul.

Interprétation comme contrainte sur l'intensité des variations

Les auteurs proposent ensuite que le renforcement puisse être compris comme une contrainte sur l'intensité des variations de M_2 : lorsque la relation entre M_1 et M_2 est établie, la relation « d'autant plus que » peut être interprétée comme une intensification des variations des attributs impliqués dans M_2 en fonction des valeurs de M_3. En considérant le même exemple précédent « *plus on est proche du mur, plus on freine fort ; d'autant plus que la vitesse est élevée* », cette interprétation signifie tout d'abord qu'une augmentation du rapprochement au mur implique une augmentation de la force de freinage, et que, de plus, l'augmentation est en corrélation avec la vitesse élevée.

Une formalisation de cette interprétation a été fournie par les auteurs comme une conjonction des deux règles : $M_1 \rightarrow M_2$ et $M_3 \rightarrow \Delta M_2$ où ΔM_2 représente les variations de M_2. Il faut noter qu'ici, cette interprétation est associée aux règles graduelles floues et non pas aux

motifs graduels. Cette interprétation nous a conduite à définir la notion d'accélération des valeurs d'attributs par rapport aux autres attributs, en exploitant la notion d'intensification, afin d'extraire *les motifs graduels accélérés* présentés dans le chapitre 5.

Interprétation comme validité accrue

Une autre interprétation comme *validité accrue* a été proposée et retenue par Bouchon-Meunier et al. (2010) : elle consiste à évaluer l'influence du renforcement M_3 sur le motif composé des motifs M_1 et M_2, par le fait que le motif est plus satisfait lorsque les objets considérés sont ceux qui possèdent M_3 à un degré élevé, plutôt que lorsque tous les objets sont considérés. Pour cette interprétation, le renforcement est dit présentiel. Aussi, le motif M_1M_2 est combiné avec la présence floue de M_3, de telle sorte que sa validité, quand on se restreint aux objets qui possèdent M_3, doit être alors supérieure à sa validité sur la base de données totale.

Dans la suite de ce chapitre, nous adoptons cette interprétation comme *validité accrue.*

Pour cette interprétation, l'approche d'extraction des motifs graduels par l'algorithme GRITE (Di Jorio et al., 2009), rappelé dans la section 1.3.3, page 41, est particulièrement pertinente. Il est en effet facile de mesurer une présence floue renforcée de M_3 quand le motif M_1M_2 est vérifié, puisque la méthode extrait explicitement les sous-ensembles de données pour lesquels le motif est vérifié. Elle est utilisée par Bouchon-Meunier et al. (2010), pour définir les critères de qualité d'un motif graduel flou renforcé.

Il est important de noter que le renforcement des motifs graduels flous est appliqué à des données floues. Ceci est imposé par l'interprétation en termes de présence floue qui lui est associée. Par conséquent, il n'est clairement pas possible d'appliquer le renforcement des motifs graduels flous aux données quantitatives. Cependant, ce dernier peut être pertinent pour traiter des attributs binaires, comme par exemple « *plus la population est élevée, plus la pollution est élevée, d'autant plus que la ville est Paris* ». Ces motifs peuvent facilement être traités dans ce contexte : les degrés d'appartenance valent 1 pour tous les objets possédant l'attribut.

2.1.3 Critères de qualité des motifs graduels flous renforcés

Dans la suite du chapitre, on utilise les notations M_1 pour le motif graduel à enrichir et M_2 pour le motif flou de la clause de renforcement.

Principe

Un motif graduel flou renforcé $M_1; M_2$ doit être évalué en fonction de ses deux composantes : le motif graduel flou M_1 et la clause de renforcement M_2. La qualité de la première est mesurée par le support du motif M_1 et celle de la seconde par les critères proposés par Bouchon-Meunier et al. (2010). Ces critères sont introduits par analogie avec les critères des règles d'association classiques $A \rightarrow B$ où A et B sont des motifs classiques, en utilisant

la mise en correspondance suivante : A représente le fait de vérifier le motif graduel et B de vérifier le renforcement. Ainsi $n(AB)$ correspond au nombre d'objets qui possèdent M_2 et qui, de plus, peuvent être ordonnés selon la contrainte d'ordre imposée par M_1. Ces objets vérifiant la contrainte d'ordre imposée par M_1 sont extraits par l'algorithme GRITE qui fournit l'ensemble des chemins complets maximaux $\mathcal{L}^*(M)$ pour tout motif graduel M valide (voir section 1.3.3, page 41).

Pour un chemin maximal $D \in \mathcal{L}^*(M)$, on note $M_2(D) = \sum_{o \in D} M_2(o)$ le cardinal flou de D selon M_2.

Il est primordial de noter que les critères de qualité proposés pour évaluer la qualité des motifs graduels renforcés sont appelés *support* et *confiance*, mais ils sont différents de ceux proposés dans le cas classique. En effet, il n'y a pas d'effet de causalité dans les motifs graduels renforcés : ce sont des motifs graduels classiques enrichis par une nouvelle information, il ne s'agit pas de règles.

Le critère de confiance est donc un critère supplémentaire évaluant la même chose que le support, c'est-à-dire évaluant la qualité d'un motif et non pas d'une règle.

Support renforcé

Comme dans le cas classique, les auteurs ont défini le support d'un motif renforcé, SR, comme une évaluation de sa fréquence. Ils ont utilisé pour cela une approche sigma-comptage, pour mesurer le degré de présence de M_2 parmi les objets qui vérifient le motif graduel M_1.

Définition 2.2 (Support renforcé)**.** Pour un motif graduel M_1 et un motif graduel flou M_2, le support renforcé de $M_1; M_2$ est défini comme :

$$SR\ (M_1; M_2) = \max_{D \in \mathcal{L}^*} \sum_{o \in D} M_2(o) = \max_{D \in \mathcal{L}^*} M_2(D) \tag{2.1}$$

Il faut noter que la définition proposée n'est pas symétrique, en raison du rôle spécifique de M_2. En outre, cette définition ne tient pas compte d'une covariation des degrés d'appartenance à la clause de renforcement M_2 avec le motif graduel M_1 : elle utilise un sigma-comptage sur la présence de M_2.

Il faut également noter que cette mesure est anti-monotone par rapport à la taille du motif flou M_2. Ainsi, si on considère deux motifs graduels renforcés : $M_{r_1} = M_1; M_2$ et $M_{r_2} = M_1; M_3$ avec $M_2 \subset M_3$ alors $SR(M_{r_1}) \geq SR(M_{r_2})$. En effet, M_{r_1} et M_{r_2} sont des motifs graduels renforcés dont le motif graduel est le même, M_1, ils ont donc le même ensemble de chemins maximaux $\mathcal{L}^*$. En outre, $\forall o \in D$ tel que $D \in \mathcal{L}^*$, $M_2(o) \geq M_3(o)$. En effet, Les motifs sont interprétés comme une conjonction des items qu'ils contiennent. Par conséquent, en notant $M = M_3 \setminus M_2$, $M_3 = M_2 \cup M$, et $\forall o, M_3(o) = \top(M_2(o), M(o)) \leq M_2(o)$. Ainsi, $\forall D \in \mathcal{L}^*, M_3(D) \leq M_2(D)$.

Cette propriété est notamment utilisée dans l'algorithme d'extraction de motifs graduels renforcés décrit à la fin de cette section, page 51.

Pour éviter de prendre en compte des objets qui ne représentent pas assez M_2, les auteurs ont proposé d'utiliser un sigma-comptage à seuil, le seuil étant défini par l'utilisateur.

La mesure de support proposée n'est pas normalisée. Habituellement cette mesure varie dans l'intervalle $[0, 1]$, comme par exemple le support graduel. Cette absence de normalisation pose problème, notamment pour montrer la validité accrue des motifs, qui est l'interprétation associée au renforcement. En effet, la comparaison du support renforcé et du support graduel n'est pas réalisable, puisque ces deux mesures varient sur des intervalles de valeurs différents. Afin d'augmenter une certaine lisibilité des motifs renforcés, nous proposons deux variantes de normalisation de cette mesure d'intérêt dans la section 2.2.2.

Confiance renforcée

Le second critère proposé évalue la force de l'effet de renforcement par rapport au motif non renforcé M_1 : la confiance renforcée consiste à calculer le rapport entre le cardinal flou selon M_2 et le cardinal de chaque chemin maximal vérifiant la contrainte de classement individuellement, et de garder le rapport ayant la valeur maximale.

Définition 2.3 (Confiance renforcée)**.** Elle est définie formellement comme :

$$CR\ (M_1; M_2) = \max_{D \in \mathcal{L}^*} \frac{M_2(D)}{|D|} \tag{2.2}$$

Il faut noter que, comme le support renforcé, et contrairement à la mesure de confiance classique, la confiance renforcée est anti-monotone par rapport à la taille du motif M_2 (Bouchon-Meunier et al., 2010). Cette propriété permet d'extraire directement les motifs graduels flous renforcés avec une confiance élevée, supérieure à un seuil défini par l'utilisateur, au lieu de filtrer a posteriori les motifs de faible confiance après leur extraction : elle permet d'extraire efficacement les motifs graduels renforcés d'intérêt.

La confiance renforcée, tout comme le support renforcé, possède la propriété de dissymétrie en raison du rôle spécifique de M_2.

Lift renforcé

Les mêmes auteurs notent que d'autres critères de qualité classiques peuvent également être étendus, comme la mesure du lift (Brin et al., 1997a) : la confiance renforcée et le support renforcé sont sensibles à la fréquence de M_2 (de la même manière que le support classique et la confiance sont sensibles à la fréquence du conséquent du motif), alors que le lift filtré ne l'est pas.

Définition 2.4 (Lift renforcé)**.** Il est défini formellement comme :

$$LR\ (M_1; M_2) = \max_{D \in \mathcal{L}^*} \frac{\frac{M_2(D)}{|D|}}{\frac{M_2(\mathcal{D})}{|\mathcal{D}|}} = \max_{D \in \mathcal{L}^*} \frac{M_2(D)}{|D|} \times \frac{|\mathcal{D}|}{M_2(\mathcal{D})} \tag{2.3}$$

Dans le cas des règles d'association, le critère de lift permet de rejeter des règles d'association candidates telles que les attributs du conséquent sont observés dans l'ensemble des données. De même, le but du lift filtré est de rejeter des clauses de renforcement basées sur des modalités telles que $\forall o, M_2(o) = 1$. Il compare le degré moyen d'appartenance à M_2 des données satisfaisant la contrainte d'ordre au degré moyen d'appartenance à M_2 dans l'ensemble de données.

2.1.4 Algorithme d'extraction des motifs graduels flous renforcés

L'algorithme proposé par Bouchon-Meunier et al. (2010) pour extraire les motifs graduels flous renforcés opère en deux étapes, en exploitant la propriété d'anti-monotonie de support renforcé :

1. Extraction de motifs graduels flous et de leurs chemins maximaux, en utilisant GRITE (Di Jorio et al., 2009, voir section 1.3.3, page 41).
2. Identification des clauses de renforcement, en utilisant le même principe qu'Apriori pour générer des motifs de longueur croissante. Cette étape est scindée en deux sous-étapes :
   (a) génération des motifs renforcés de longueur $k+1$ à partir des motifs renforcés de longueur k valides,
   (b) validation des clauses de renforcement en utilisant le critère de qualité proposé du support renforcé.

2.2 Étude complémentaire des motifs graduels flous renforcés

Dans cette section, nous approfondissons l'étude du renforcement réalisée par Bouchon-Meunier et al. (2010). Nous discutons tout d'abord l'extraction des motifs graduels renforcés par filtrage et en étudions deux méthodes différentes. Nous proposons ensuite de nouveaux critères de qualité en étudiant leurs propriétés et en les comparant aux critères existants. Nous illustrons enfin leur pertinence en réalisant des expérimentations sur des données jouets, puis sur des données réelles.

2.2.1 Discussion sur l'extraction par filtrage

Nous détaillons ici l'équivalence indiquée par Bouchon-Meunier et al. (2010) entre deux explications de l'interprétation retenue, pour évaluer l'influence du renforcement de la présence de M_2 sur le motif graduel M_1. La première interprète que l'influence du renforcement M_2 sur le motif M_1 comme le fait que le motif est mieux satisfait lorsque les objets considérés sont ceux qui possèdent M_2 plutôt que lorsque l'ensemble des données est pris en compte. La deuxième interprète l'influence du renforcement de M_2 sur le motif M_1 par le degré avec lequel M_2 est possédé par les objets vérifiant le motif M_1. Aussi, la première considère une restriction appliquée à la base de données entière alors que la deuxième l'applique au chemin maximal. Deux méthodes d'extraction par filtrage peuvent être distinguées selon ces deux explications :

Id.	proche du mur	freine fort	vitesse élevée
1	1	0	1
2	1	1	1
3	1	1	1
4	1	1	0
5	0	0	0
6	0	1	1
7	0	1	0
8	1	1	1

Tableau 2.1 – Exemple de base de données binaires

1. la première consiste à filtrer la base de données selon le motif de renforcement M_2, puis à appliquer l'algorithme d'extraction de motifs graduels flous à la base filtrée, pour chaque candidat M_1.
2. la seconde consiste à extraire les motifs graduels flous de la base entière et filtrer directement les chemins maximaux par M_2. Cette méthode évite de passer deux fois par la base de données et applique le processus d'extraction une seule fois, c'est-à-dire qu'elle combine les deux étapes de la première en une seule étape. Cela permet de réduire la complexité en termes de temps de calcul de l'extraction.

Nous illustrons ces méthodes d'extraction en considérant tout d'abord le cas de données binaires, puis le cas général pour le motif graduel renforcé utilisé comme exemple tout au long du chapitre « *plus on est proche du mur, plus on freine fort ; d'autant plus que la vitesse est élevée* ». On note M_1 le motif graduel « *plus on est proche du mur, plus on freine fort* » et M_2 la clause de renforcement « *d'autant plus que la vitesse est élevée* ».

Cas binaire

Les données binaires considérées sont représentées dans le tableau 2.1 : elles conduisent à quatre chemins maximaux de support $\{1, 3, 4, 7\}$, $\{2, 3, 4, 7\}$, $\{1, 3, 4, 8\}$ et $\{2, 3, 4, 8\}$, respectivement associés à un cardinal selon M_2 de 2, 2, 3 et 3. Le support renforcé est donc 3, atteint pour les chemins $\{1, 3, 4, 8\}$ et $\{2, 3, 4, 8\}$.

La figure 2.1 illustre les deux méthodes d'extraction précédentes : $\mathcal{D}_f$ représente la base de données filtrée selon M_2 et D_{if} représente à la fois le résultat du filtrage selon le chemin D et selon M_2. (A) et (B) correspondent aux étapes de la deuxième méthode : (A) est l'étape d'extraction des motifs graduels flous (application de l'algorithme GRITE) et (B) est l'étape de filtrage des chemins maximaux ; (C) et (D) correspondent aux étapes de la première méthode : (C) est le filtrage de la base de données selon la présence de M_2 et (D) est l'étape d'extraction des motifs graduels flous de cette base filtrée (utilisation de l'algorithme GRITE).

On constate qu'au lieu de passer deux fois par la base de données, ce qui augmente le temps de calcul de l'algorithme, on peut ne passer qu'une seule fois, en filtrant directement

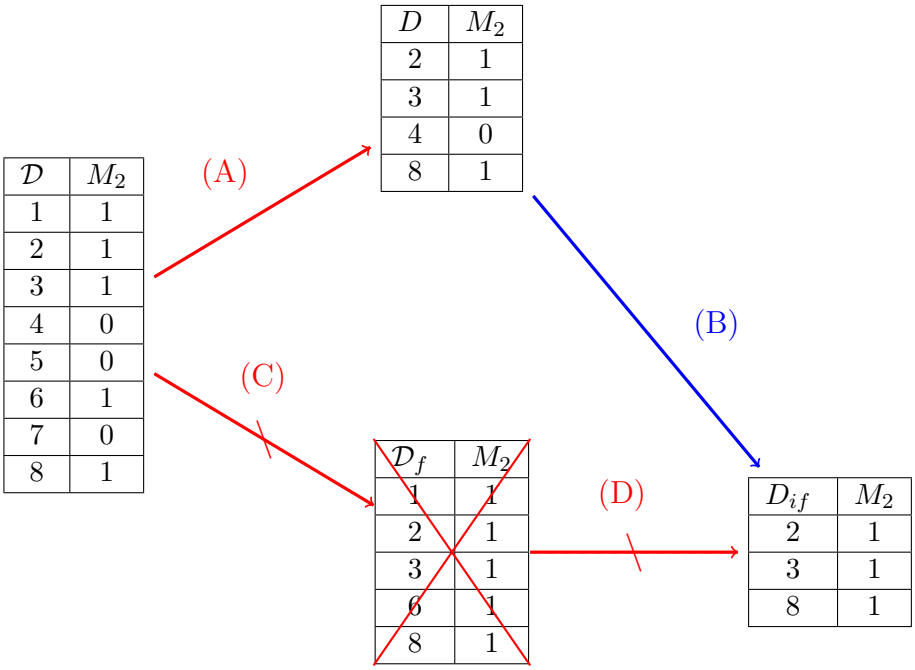


Figure 2.1 – Extraction et filtrage des chemins maximaux.

les chemins maximaux obtenus à partir de la première phase d'extraction, comme illustré par l'étape (B).

Cas flou

La méthode de filtrage peut également être appliquée aux données floues, comme illustré dans la figure 2.2 qui représente le même exemple pour les données représentées dans le tableau 1.4, page 33 du chapitre précédent. Elles ont été choisies de façon à donner les mêmes chemins maximaux que le cas binaires, respectivement associés aux cardinaux selon M_2 valant 2.6, 2.7, 2.9 et 3. Le support renforcé est donc de 3, à nouveau obtenu pour le chemin maximal $\{2, 3, 4, 8\}$.

Contrairement au cas binaire où les objets pour lesquels $M_2 = 0$ sont supprimés et ceux pour lesquels $M_2 = 1$ sont conservés, dans le cas flou, les objets contribuent avec des poids qui constituent leur degré d'appartenance aux modalités floues.

Par analogie avec le cas binaire, on peut manipuler les degrés d'appartenance à la clause de renforcement pour les objets vérifiant le motif graduel à la place des 1 et des 0, correspondant à la flèche en haut à droite (B) de la figure 2.1. Ceci est illustré sur la figure 2.2.

2.2.2 Extension des critères de qualité

Dans cette section, nous présentons les extensions que nous avons proposées concernant les critères de qualité des motifs graduels renforcés. Nous discutons d'abord la normalisation

$\mathcal{D}$	M_2
1	0.5
2	0.6
3	0.9
4	0.8
5	0.3
6	0.5
7	0.4
8	0.7

D	M_2
2	0.6
3	0.9
4	0.8
8	0.7

Figure 2.2 – Filtrage dans le cas flou.

du support renforcé évoquée dans la section 2.1.3, qui garantit des valeurs dans l'intervalle $[0, 1]$: nous proposons deux normalisations, qui conduisent à deux nouveaux critères de sémantiques différentes, que nous appelons respectivement *support graduel renforcé*, SGR, et *support graduel filtré*, SGF. Nous discutons ensuite d'une extension du risque relatif, RR.

Nos motivations concernant ces nouvelles mesures sont doubles. D'une part, nous souhaitons enrichir l'évaluation de la qualité des motifs graduels renforcés, puisque le nombre de ces motifs extraits est souvent plus élevé que celui des motifs graduels classiques (voir section 3.4.1, page 59). D'autre part, nous garantissons une lisibilité des motifs graduels renforcés en considérant l'unique intervalle $[0, 1]$ où les mesures de qualité varient.

Support graduel renforcé

La normalisation classique du support, dans le cas des règles d'association, rapporte le nombre de co-occurrences au nombre total de données dans la base de données. Ceci permet d'obtenir une mesure de qualité qui varie dans l'intervalle $[0, 1]$. Aussi, par analogie, on peut normaliser le support renforcé défini dans l'équation (2.1) par le nombre total d'objets $|\mathcal{D}|$.

Définition 2.5 (Support graduel renforcé)**.** Pour un motif graduel renforcé $M = M_1; M_2$, le support graduel renforcé est défini comme :

$$\text{SGR (M)} = \frac{1}{|\mathcal{D}|} \max_{D \in \mathcal{L}^*} M_2(D) \tag{2.4}$$

En plus de garantir une valeur dans l'intervalle $[0, 1]$, cette normalisation permet de vérifier que les objets considérés sont suffisamment représentatifs de la base de données. Elle évalue le degré moyen d'appartenance à M_2 sur la base de données totale, $\mathcal{D}$, qui est calculé pour chacun des chemins maximaux, et leur maximum représente le support graduel renforcé du motif considéré.

A titre d'exemple, pour le motif graduel renforcé « *plus on est proche du mur, plus on freine fort, d'autant plus que la vitesse est élevée* » extrait à partir de la base de données floues du tableau 1.4, page 33, on a $SR = 3$ et $SGR = \frac{3}{8} = 0.375$.

La valeur de SR ne permet pas de préciser à quel degré les objets vérifiant le motif sont représentatifs de la base de données selon la modalité floue « vitesse élevée ». Au contraire, la valeur du SGR est facilement interprétable. Elle signifie que les objets vérifiant le motif appartiennent à la modalité floue « vitesse élevée » avec un degré d'appartenance moyen de 0.375 sur la base de données totale $\mathcal{D}$, qui n'est donc ici pas très représentatif.

Support graduel filtré

On peut également considérer le support renforcé comme un filtrage du support graduel par le biais de M_2 : au lieu de compter $\max_{D \in \mathcal{L}^*} |D|$ comme dans le support graduel classique défini dans le chapitre 1 (équation (1.7)), on considère $\max_{D \in \mathcal{L}^*} M_2(D)$ dans l'équation (2.1).

On peut donc proposer de normaliser en utilisant le même filtre, c'est-à-dire en remplaçant le cardinal $|\mathcal{D}|$ par son cardinal flou selon M_2.

Définition 2.6 (Support graduel filtré)**.** Pour un motif graduel renforcé $M = M_1; M_2$, le support graduel filtré est défini comme :

$$\text{SGF (M)} = \frac{1}{M_2(\mathcal{D})} \max_{D \in \mathcal{L}^*} M_2(D) \tag{2.5}$$

Le support graduel filtré permet de mesurer à quel degré les objets vérifiant le motif graduel flou appartiennent à M_2. Il évalue le degré d'appartenance à M_2 des objets vérifiant le motif M_1 par rapport aux degrés d'appartenance à M_2 des objets de la base de données totale. La valeur de cette mesure varie également dans l'intervalle $[0, 1]$. Elle indique à quel point les objets considérés sont suffisamment présents dans la base de données.

Considérons le même exemple que précédemment, où M_2 est la modalité floue « *vitesse élevée* » ; pour la base de données illustrée dans le tableau 1.4, page 33, on a $M_2(\mathcal{D}) = 4.7$ et $SGF = \dfrac{3}{4.7} = 0.64$.

On peut souligner que, si l'on considère le quotient SGF/SG afin de quantifier la validité accrue des motifs graduels, on retrouve le critère du lift filtré LF (voir équation (2.3)) : celui-ci compare en effet le support graduel avant et après filtrage, et valide un motif graduel renforcé seulement s'il a une valeur significativement supérieure à 1. Il compare donc la validité du motif M_1 sur la base de données filtrée selon M_2 à la validité de la base de données entière.

Autres critères de qualité

D'autres critères de qualité classiques d'évaluation des règles d'association que nous avons décrits dans le chapitre 1, page 20 (voir Lenca et al., 2007) peuvent également être étendus. On peut noter qu'il existe des mesures de qualité qui s'intéressent à A, c'est-à-dire à la présence de A, telles que les mesures de support, et d'autres à $\overline{A}$, c'est-à-dire à l'absence de A, telles que le risque relatif. Or dans le cas classique, correspondant à des données binaires, A représente la présence de l'item A et $\overline{A}$ son absence.

Dans les sous-sections ci-dessus, seules des mesures de qualité reposant sur la présence de A ont été présentées. Dans cette section, nous montrons que les mesures reposant sur $\overline{A}$, dans le cas des motifs graduels renforcés, ne conviennent pas, en raison de la complexité de la transposition de $\overline{A}$ pour le cas de motifs graduels. La démonstration est essentiellement concentrée sur la mesure du risque relatif.

Le risque relatif, noté RR, est utilisé dans le domaine de l'épidémiologie et représente le rapport entre le risque de survenue d'une maladie chez les personnes exposées au facteur de risque et le risque chez les personnes non exposées.

De façon générale, pour une règle d'association $A \rightarrow B$, le RR est défini comme le rapport entre la proportion de transactions qui vérifient B parmi celles qui vérifient A et la proportion de celles qui ne vérifient pas A.

Définition 2.7 (Risque relatif)**.** Le risque relatif est défini comme :

$$\mathrm{RR(A \rightarrow B)} = \frac{\frac{n(AB)}{n(A)}}{\frac{n(\overline{A}B)}{n(\overline{A})}} = \frac{n(AB)}{n(A)} \times \frac{n(\overline{A})}{n(\overline{A}B)} \tag{2.6}$$

Il faut noter que RR est intéressant s'il est soit significativement supérieur soit significativement inférieur à 1. En effet, si le RR est supérieur à 1, cela signifie que la proportion de transactions qui vérifient B parmi celles qui vérifient A est bien supérieure à celle qui ne vérifient pas A. Au contraire, si le RR est inférieur à 1, cela signifie que la proportion de transactions qui vérifient B parmi celles qui vérifient A est bien inférieure à celle qui ne vérifient pas A.

Nous définissons le risque relatif renforcé, RRR, pour un motif graduel flou renforcé, $M_1; M_2$, par analogie avec la définition classique de l'équation (2.6), en utilisant la mise en correspondance suivante : A représente le fait de vérifier le motif graduel M_1 et $\overline{A}$ de ne le pas vérifier, c'est-à-dire $\overline{M_1}$. Ainsi $n(\overline{M_1})$ correspond au nombre d'objets qui ne peuvent pas être ordonnés pour vérifier la contrainte d'ordre imposée par le motif M_1. Cela signifie que $n(\overline{M_1}) = |\mathcal{D} - \bigcup_{D \in \mathcal{L}^*(M_1)} D|$. RRR est donc le rapport entre le degré moyen d'appartenance à M_2 des objets qui appartiennent à l'un des chemins D et le degré moyen d'appartenance à M_2 des objets qui n'appartiennent à aucun D.

Définition 2.8 (Risque relatif renforcé)**.** Pour un motif graduel renforcé $M = M_1; M_2$, le risque relatif renforcé est défini comme :

$$\mathrm{RRR} = \max_{D \in \mathcal{L}^*} \frac{M_2(D)}{|D|} \times \frac{|\mathcal{D} - \bigcup_{D \in \mathcal{L}^*(M_1)} D|}{M_2(\mathcal{D} - \bigcup_{D \in \mathcal{L}^*(M_1)} D)} \tag{2.7}$$

Le RRR est important uniquement s'il est supérieur à 1 car on souhaite que la présence de M_2 des objets vérifiant le motif M_1, c'est-à-dire $M_2(D)$ soit supérieure à celle des objets qui ne vérifient pas le motif M_1, ce qui impose que $M_2(D)/|D|$ soit supérieur à $\frac{M_2(\mathcal{D} - \cup D)}{|\mathcal{D} - \bigcup_{D \in \mathcal{L}^*(M_1)} D|}$.

Autrement dit, $M_2(D)$ doit être élevé et $M_2(\mathcal{D} - \underset{D\in\mathcal{L}^*(M_1)}{\cup} D)$ faible, ce qui conduit à une contradiction : si le motif est vérifié par plusieurs chemins maximaux et qu'il existe un objet o, tel que o appartient à un chemin et pas à un autre, alors, d'une part, la condition selon laquelle $M_2(D)$ est suffisamment grand impose que o appartienne considérablement à M_2 ; d'autre part, la condition avec laquelle $M_2(\mathcal{D} - \cup D)$ est assez faible impose que o n'appartienne pas considérablement à M_2. La satisfaction des deux conditions revient donc à calculer la moyenne, c'est-à-dire à chercher un $RRR = 1$. Toutefois, nous nous intéressons uniquement aux cas où RR est strictement supérieur à 1.

Ceci nous amène à exclure le critère RRR des critères de mesure de qualité des motifs graduels flous renforcés. Il en serait de même pour toutes les mesures de qualité reposant sur $\overline{A}$. En effet, on peut facilement généraliser cette constatation à toutes les mesures reposant sur $\overline{A}$: dans le cas des motifs graduels renforcés, $n(\overline{A})$ correspond au nombre d'objets qui ne peuvent pas être ordonnés pour vérifier la contrainte d'ordre imposée par le motif M_1, ce qui peut conduire à une contradiction. Cette contradiction est due au fait que, sur un ensemble de chemins maximaux, quelques objets peuvent vérifier A uniquement sur certains chemins, ce qui signifie qu'ils peuvent vérifier $\overline{A}$ sur d'autres chemins. Nous illustrons ce principe sur l'exemple ci-dessous en considérant la mesure du RRR.

Exemple Soit M_1 un motif graduel obtenu à partir d'une base de données contenant 8 objets numérotés de 1 à 8 avec M_1 vérifié par deux chemins maximaux D_1 et D_2 tels que : $D_1 = \{1, 3, 6, 7, 8\}$ et $D_2 = \{4, 3, 1, 6, 8\}$.

On a

$$M_2(\mathcal{D} - D_1) = M_2(2) + M_2(4) + M_2(5) \tag{2.8}$$

$$M_2(D_2) = M_2(4) + M_2(3) + M_2(1) + M_2(6) + M_2(8) \tag{2.9}$$

Du fait de la contrainte selon laquelle $M_2(D)$ est élevé, on souhaite donc que $M_2(2)$, $M_2(4)$ et $M_2(5)$ soient faibles dans l'équation (2.8). Simultanément, la contrainte selon laquelle $M_2(\mathcal{D} - \cup D)$ est assez faible impose que $M_2(4)$, $M_2(3)$, $M_2(1)$, $M_2(6)$ et $M_2(8)$ soient élevés dans l'équation (2.9).

Ainsi $M_2(4)$ doit être faible dans l'équation (2.8) et élevé dans l'équation (2.9), ce qui mène à un problème de contradiction. Afin de pallier ce problème, il faut calculer la moyenne, c'est-à-dire chercher $RR = 1$. Mais ceci ne peut pas confirmer la pertinence d'un motif (ou d'une règle) puisque uniquement les cas où $RR > 1$ sont intéressants.

Cet exemple montre que les critères reposant sur $\overline{A}$ ne conviennent pas dans le cas de motifs graduels flous renforcés.

2.2.3 Étude et illustration de la complémentarité de critères

Nous discutons l'apport des critères complémentaires proposés dans la section précédente en mettant l'accent sur leur intérêt individuel apporté dans l'évaluation des motifs graduels flous renforcés. Tout d'abord, nous illustrons cela sur les exemples de données artificielles

présentées dans les tableaux A.2 et A.3 donnés en annexe en page 147, illustrées sur les figures 2.3 et 2.4, en confirmant nos propos liés à l'exemple parfait de la figure 2.4. Ensuite, pour une discussion plus élaborée, nous reprenons la base de données jouet fournie dans le tableau 1.4, page 33 qui contient 8 objets décrits par 8 modalités floues.

Les critères définis dans les sections 2.1.3 et 2.2.2 ont chacun un intérêt individuel tout en se complétant entre eux. En effet, le support graduel est donné pour l'évaluation des motifs graduels flous. Le support graduel filtré doit être plus élevé que le support graduel pour qu'un motif graduel flou renforcé soit valide. Le lift filtré compare ces deux derniers supports et permet de confirmer la validité accrue du motif, qui est l'interprétation associée au renforcement.

Définition 2.9 (Validité d'un motif graduel flou renforcé)**.** Nous appelons *un motif graduel flou renforcé valide* un motif pour lequel les quatre critères SG, SGF, SGR et le LF dépassent les seuils minimaux fixés par l'utilisateur (pour LF, le seuil doit être supérieur à 1).

Intérêt du support graduel renforcé : SGR complète les trois autres critères

Les figures 2.3 et 2.4 représentent deux ensembles de données décrits par trois attributs A, B et C, respectivement indiqués par l'abscisse, l'ordonnée et la taille des points. Le motif graduel $M_1 = A^{\geq}B^{\geq}$ est supporté par le chemin représenté par les points bleus. Les valeurs des critères de qualité du motif $A^{\geq}B^{\geq}; C$ sont données dans le tableau 2.2.

Si un seuil de support graduel est fixé à une valeur inférieure ou égale à 3/4 alors dans les deux cas le motif graduel M_1 est valable puisqu'il suffit de supprimer 2 objets pour qu'une co-variation parfaite soit obtenue.

La différence provient d'une part du fait que, dans l'exemple de la figure 2.3, les degrés d'appartenance à C sont globalement beaucoup plus faibles que pour l'exemple de la figure 2.4, et surtout, du fait que les 2 points rouges qui ne vérifient pas le motif M_1 sont précisément ceux pour lesquels le degré d'appartenance à C est faible dans l'exemple de la figure 2.3, alors que ce n'est pas le cas dans l'exemple de la figure 2.4. Le support graduel filtré ne permet pas de capturer cette différence et donne des valeurs proches pour les deux bases de données. Il en est de même pour la mesure du lift filtré. On est donc face à une situation où les trois mesures de qualité ne suffisent pas pour déterminer dans quel cas le motif graduel renforcé est réellement pertinent. Cependant, comme on peut le voir sur le tableau 2.2, le support graduel renforcé est discriminant et permet de distinguer les deux motifs graduels renforcés.

Illustration à l'aide de la base de données jouet données dans le tableau 1.4, page 33

Nous illustrons maintenant de manière détaillée l'apport des critères de qualité proposés avec des exemples obtenus à l'aide de la base de données floues présentée dans le tableau 1.4, page 33 du chapitre 1.

Le tableau 2.3 montre quelques motifs graduels renforcés que l'on peut extraire de la base de données jouet en fixant le seuil minimum du support graduel à 50%, les autres critères,

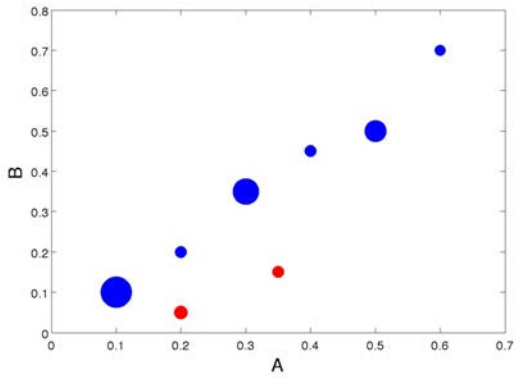


Figure 2.3 – $M = A^{\geq}B^{\geq}; C$.

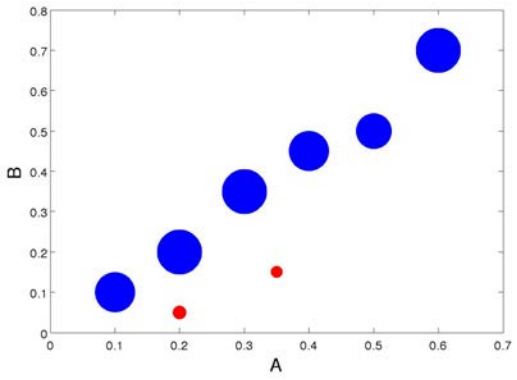


Figure 2.4 – $M = A^{\geq}B^{\geq}; C$.

Critères de qualité	Exemple de la figure 2.3	Exemple de la figure 2.4
Support graduel	$\frac{6}{8} = 75\%$	$\frac{6}{8} = 75\%$
Support graduel filtré	$\frac{0.52}{0.78} = 67\%$	$\frac{5.84}{6.1} = 95\%$
Lift renforcé	1.1	1.5
Support graduel renforcé	$\frac{0.52}{8} = 6.5\%$	$\frac{5.84}{8} = 73\%$

Tableau 2.2 – Comparaison des critères de qualité du motif graduel renforcé $A^{\geq}B^{\geq}; C$ pour les deux exemples illustrés dans les figures 2.3 et 2.4.

c'est-à-dire le support graduel filtré et le support graduel renforcé à 10%. Le lift filtré est mesuré pour tous les motifs satisfaisant ces contraintes. À chaque motif graduel renforcé, nous avons associé dans le tableau 2.4 les valeurs obtenues pour les critères de qualité SG, SGR, LF et SGF. Ces motifs sont ordonnés par ordre décroissant de SG.

De manière surprenante, avec les conditions de seuils fixées, le nombre de motifs graduels renforcés extraits est de 40, pour une base de données contenant seulement 8 objets. Le filtrage avec de telles conditions n'est donc pas pertinent. Pourtant le seuil de SG est fixé à 50%, ce qui illustre un problème classique, se posant également pour les motifs graduels, de quantité de règles extraites. Il faudrait alors être plus exigeant avec les seuils des autres critères. En effet, l'extraction de motifs graduels renforcés risque d'augmenter énormément : pour des données décrites par m attributs, le nombre de motifs graduels fréquents potentiels est, de l'ordre de 2^m ; chacun pourrait être renforcé par un nombre de clauses différent de l'ordre de 2^{m-k} où k est le nombre de modalités floues composant le motif à renforcer. Au total, le nombre de motifs graduels renforcés fréquents potentiels est de $2^m \times 2^{m-k}$. Ceci explique l'intérêt de s'appuyer sur les critères proposés pour évaluer la qualité des motifs graduels renforcés et de filtrer les motifs extraits : les seuils choisis ci-dessus peuvent être fixés à des valeurs élevées ou inclure le lift filtré.

Id.	Motifs graduels		Clause de renforcement
M_1	Freinage.faible$\geq$,	Vitesse.normale$\geq$	Distance.proche
M_2	Freinage.faible$\geq$,	Vitesse.normale$\geq$	Distance.loin
M_3	Freinage.faible$\leq$,	Distance.loin$\geq$	Vitesse.lente
M_4	Freinage.faible$\leq$,	Distance.loin$\geq$	Vitesse.normale
M_5	Freinage.faible$\leq$,	Distance.loin$\geq$	Vitesse.élevée
M_6	Vitesse.normale$\geq$,	Distance.loin$\leq$	Freinage.faible
M_7	Freinage.normal$\geq$,	Vitesse.lente$\geq$	Distance.loin
M_8	Freinage.normal$\geq$,	Vitesse.lente$\geq$	Distance.proche
M_9	Freinage.normal$\geq$,	Vitesse.élevée$\geq$	Distance.proche
M_{10}	Freinage.faible$\geq$,	Vitesse.lente$\geq$	Distance.proche

Tableau 2.3 – Motifs graduels renforcés obtenus pour la base de données du tableau 1.4, page 33.

Id.	SG	SGR %	LR	SGF %
M_1	75	23	0.89	67
M_2	75	53	1.06	79
M_3	75	51	1.04	78
M_4	62.5	24	1.17	73
M_5	62.5	38	1.02	64
M_6	62.5	15	1.07	67
M_7	62.5	6	1.0	63
M_8	62.5	51	1.25	78
M_9	50	57	0.99	87
M_{10}	50	30	1.02	89

Tableau 2.4 – Valeurs des critères de qualité des motifs du tableau 2.3.

En tenant compte de LF (parmi les motifs satisfaisant les contraintes des seuils minimaux des autres critères, extraire ceux pour qui la valeur de LF est supérieure à 1), nous nous apercevons que le nombre de motifs graduels renforcés extraits est environ réduit de moitié : 22 motifs graduels renforcés sont obtenus. Il est donc intéressant de tenir compte de LF dans l'évaluation de la qualité des motifs graduels renforcés.

Comme indiqué ci-dessus, 2^m est un nombre approché du nombre de motifs pouvant être extraits pour des données décrites par m attributs numériques. Dans le cas des données floues, le nombre de motifs graduels flous fréquents potentiels est inférieur à 2^m, puisqu'un motif ne peut contenir des modalités floues associées à un même attribut flou.

Il faut noter que 2^{m-k} est le nombre le plus élevé de clauses de renforcement qui peuvent être identifiées. La valeur de k considérée est la valeur qui induit le plus petit nombre de motifs, composés de 1 ou plusieurs items apparaissant dans le motif à enrichir, ce qui correspond à 2^k clauses à écarter du comptage du nombre de clauses de renforcement, c'est-à-dire que 2^k clauses ne peuvent donc pas renforcer le motif. Or, les k items peuvent apparaître dans d'autres motifs qui ne sont pas forcément composés de ces items. 2^{m-k} est donc une valeur

Motif graduel renforcé	$SG\%$	$SGR\%$	$SGF\%$	LF
Freinage.faible$\geq$, Vitesse.normale$\geq$; Distance.loin	75	53	79	1.06
Freinage.faible$\leq$, Distance.loin$\geq$; Vitesse.lente	75	51	78	1.04
Freinage.normal$\geq$, Vitesse.lente$\geq$; Distance.proche	62.5	51	78	1.25

Tableau 2.5 – Motifs graduels flous renforcés validés pour la base de données du tableau 1.4, page 33 en fixant les seuils minimaux des critères de qualité à 50% et $LF > 1$.

approchée, car toutes les clauses où un des k items apparaît ne peuvent pas être des clauses de renforcement du motif composé des k items.

En se basant sur cette observation, on peut se demander s'il ne serait pas suffisant d'évaluer la qualité des motifs graduels renforcés en s'appuyant uniquement sur LF. Dans les exemples fournis dans le tableau 2.3, le motif M_4 illustre le cas où le lift filtré seul ne suffit pas : bien qu'il vaille 1.17, soit supérieur au seuil de 1, et que SGF et SG soient élevés, SGR est très faible. Le motif M_4 ne peut donc pas être considéré comme valide. Il en est de même pour les motifs M_7 et M_6. Le lift filtré seul ne suffit donc pas pour évaluer les motifs graduels flous renforcés.

La comparaison des motifs M_7 et M_8 montre l'intérêt de l'utilisation du support graduel renforcé : ils correspondent en effet à 2 renforcements différents d'un même motif graduel. On constate qu'ils ont des valeurs de support graduel égales et celles du support graduel filtré et du lift filtré presque égales. Si l'on tient compte uniquement de ces valeurs, alors les deux motifs ont la même validité et on aboutit à deux motifs graduels renforcés exprimant deux connaissances contradictoires. Toutefois, la valeur du support graduel renforcé permet une nouvelle fois de distinguer les deux motifs, comme pour les exemples illustrés sur les figures 2.3 et 2.4 : les objets vérifiant le motif « Freinage.normal$\geq$, Vitesse.lente$\geq$ » appartiennent suffisamment à la clause de renforcement « Distance.proche » mais pas suffisamment à la clause « Distance.loin ».

Le motif ayant un support graduel renforcé élevé est validé et celui ayant un support graduel renforcé faible est rejeté. On peut en conclure que le motif graduel « *Freinage.normal$\geq$, Vitesse.lente$\geq$* » est mieux renforcé avec la clause « *Distance.proche* ». Nous avons illustré, avec les deux exemples précédents, la nécessité et l'intérêt d'introduire le support graduel renforcé pour évaluer les motifs graduels flous renforcés.

Si l'on souhaite être exigeant sur les contraintes des autres critères et que l'on fixe cette fois tous les critères à 50% et le lift filtré supérieur à 1, alors ceci aide à réduire considérablement le nombre de motifs graduels extraits et ne laisse que les motifs graduels renforcés intéressants. Dans le cas de la base de données utilisée dans cet exemple, seuls les 3 motifs M_2, M_3 et M_8 sont conservés. Le tableau 2.5 les récapitule.

Synthèse

Les nouveaux critères de qualité proposés améliorent considérablement l'évaluation de la qualité des motifs graduels renforcés. De plus, ces critères filtrent davantage les motifs extraits et réduisent donc leur nombre, en validant uniquement ceux qui sont pertinents. Il faut noter que la réduction du nombre de motifs extraits peut se réaliser simplement par le durcissement de la contrainte du support renforcé, sans prendre en compte les critères proposés. Cela n'est cependant pas l'objectif premier. Le but des critères que nous avons proposés est de valider les motifs extraits qui sont réellement intéressants, et le critère de support renforcé seul ne pourrait pas répondre à cet objectif, même en lui fixant une valeur élevée.

2.2.4 Étude expérimentale du temps de calcul et de l'occupation mémoire

Dans cette section, nous nous intéressons à la fois au nombre de motifs extraits en fonction des seuils minimaux utilisés pour les critères de qualité et au coût de l'algorithme d'extraction des motifs graduels flous renforcés.

Données considérées

Pour illustrer le comportement de notre méthode, nous avons effectué des tests sur plusieurs bases de données jouet.

Comme la base de données illustrative utilisée ci-dessus ne contient que 8 objets, les résultats obtenus en termes de temps et de mémoire sur cette base ne permettent pas d'affirmer que l'approche proposée est capable de traiter des seuils faibles des critères de qualité. Pour cette raison, nous avons effectué des expérimentations sur la base de données d'UCI (Bache & Lichman, 2013) *Wine quality red* que nous avons fuzzifiée.

Celle-ci contient 1599 objets correspondant à des vins rouges décrits par 12 attributs numériques et un degré de qualité entre 0 et 10. Chaque attribut a été remplacé par deux modalités floues *faible* et *élevée*, avec des degrés d'appartenance obtenus grâce à une fonction d'appartenance trapézoïdale illustrée par la figure 2.5 : m et M représentent respectivement la valeur minimale et maximale décrivant l'attribut à fuzzifier. Les degrés d'appartenance à la modalité floue « faible » sont calculés avec une interpolation linéaire entre les points $(m, 1)$ et $(M, 0)$ et les degrés d'appartenance à la modalité floue « élevée » sont calculés avec une interpolation linéaire entre les points $(M, 1)$ et $(m, 0)$.

En notant $A(o)$ la valeur numérique de l'attribut A décrivant un objet o, on définit formellement les fonctions d'appartenance aux modalités floues « élevée » et « faible » par les fonctions $\mu_{\text{élevée}}$ et μ_{faible} de la manière suivante :

$$\mu_{\text{élevée}}(o) = \begin{cases} 0 & \text{si } A(o) = m \\ 1 & \text{si } A(o) = M \\ \dfrac{A(o) - m}{M - m} & \text{si } m < A(o) < M \end{cases}$$

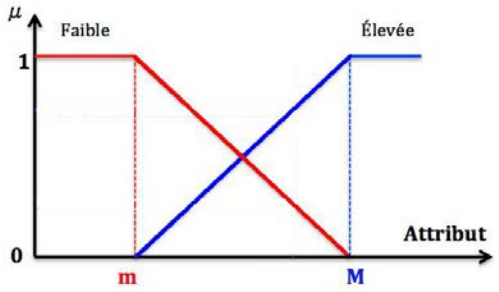


Figure 2.5 – Représentation des modalités floues « faible » et « élevée » associées à chaque attribut numérique.

Configuration	minSG %	minSGR %	minSGF %	nb motifs extraits	temps (min)	mémoire (mo)
C_1	10	10	10	12000	80	190
C_2	10	20	20	710	11	110
C_3	60	30	30	511	5	70
C_3	65	40	40	201	2	30

Tableau 2.6 – Résultats obtenus pour l'expérimentation avec la base de données *Wine quality red*

$$\mu_{\text{faible}}(o) = \begin{cases} 0 & \text{si } A(o) = M \\ 1 & \text{si } A(o) = m \\ 1 - \dfrac{A(o) - m}{M - m} & \text{si } m < A(o) < M \end{cases}$$

Évaluation des performances de la méthode

Les expérimentations ont été menées afin de mesurer les consommations en termes de temps et de mémoire, en fonction des seuils minimaux des critères de qualité. Ces expérimentations ont été réalisées sur un ordinateur Intel(R) Core(TM)2 Duo CPU E8500 3.16GHz doté de 4Go de RAM.

Les résultats obtenus sont résumés dans le tableau 2.6. Ce tableau montre, pour chaque configuration correspondant aux seuils des critères minimaux utilisés, le nombre de motifs graduels renforcés extraits et le temps nécessaire pour leur extraction et la mémoire utilisée.

La première configuration montre qu'un nombre important de motifs est extrait, bien que la base de données ne contienne pas un grand nombre d'attributs. Ce nombre de motifs générés justifie les ressources consommées. En effet, la génération d'un grand nombre de motifs implique le stockage d'un grand nombre de matrices binaires, ce qui explique l'importance de la mémoire consommée et le temps nécessaire pour le calcul.

La deuxième configuration a la même valeur de $minSG$ que la première configuration et les autres seuils sont fixés à des valeurs légèrement plus élevées que les précédentes : on constate que le nombre de motifs graduels renforcés extraits chute considérablement. Cela confirme qu'il est important d'être exigeant avec les contraintes imposées aux critères évaluant le renforcement. Comme détaillé ci-dessus, le nombre de motifs graduels renforcés est généralement beaucoup plus élevé que celui des motifs graduels classiques. Ceci est justifié par le fait qu'un motif graduel peut être renforcé par un nombre important de clauses de renforcement, ce nombre étant relatif au nombre d'attributs de la base de données utilisée.

Le temps d'extraction pour un support graduel minimal de 60% ou de 65%, et pour $minSGR$ et $minSGF$ allant de 30 à 40 est très acceptable, puisqu'il varie de deux minutes à environ cinq minutes, pour extraire environ 500 motifs graduels renforcés. En revanche, en dessous de 60%, le temps d'extraction et le nombre de motifs extraits croissent considérablement. Au-delà des performances de calcul, on peut ainsi noter que baisser les seuils des critères en dessous de 60% pour SG et en dessous de 30 pour les autres critères génère trop de motifs ; il n'est donc pas facile de les interpréter. De plus, baisser ces seuils ne signifie pas pour autant que très peu de motifs graduels renforcés sont générés. En effet, la base de données contient un très grand nombre de motifs graduels dont le support graduel est proche de 60%, ce qui permet de générer suffisamment de motifs graduels renforcés dont le SGF et le SGR sont proches de 40%.

2.3 Renforcement des règles d'association

Dans cette section, nous nous intéressons à la transposition de la notion de renforcement au cas des règles d'association classiques. Pour cela, nous définissons les règles d'association renforcées ainsi que les critères de qualité qui peuvent être utilisés pour mesurer leur pertinence. Nous discutons ensuite de leurs interprétations possibles et de l'apport de ce renforcement par rapport à une règle d'association classique. Nous montrons qu'il est limité, en établissant qu'une règle d'association renforcée est équivalente à deux règles d'association classiques.

Il est primordial de souligner que le renforcement de règles d'association correspond en fait à un renforcement de motifs : le même renforcement s'applique indifféremment pour $A \rightarrow B$ et pour $B \rightarrow A$.

2.3.1 Définition et interprétation des règles d'association renforcées

Une règle d'association renforcée est représentée sous la forme $A \rightarrow B; C$. Elle se compose d'une règle d'association $A \rightarrow B$ qui exprime le lien entre les motifs A et B, et d'une clause de renforcement C introduite par l'expression « *d'autant plus que* ». Ces règles peuvent être illustrées par l'exemple « *si on achète du lait, alors on achète du pain, d'autant plus qu'on achète du beurre* ». Dans cet exemple, les informations sur l'achat du beurre permettent d'enrichir la relation établie entre les items lait et pain par une précision supplémentaire.

Critères de qualité	$A \rightarrow B; C$
SAR = Support d'association renforcé	$\frac{n(ABC)}{n}$
CAR = Confiance d'association renforcée	$\frac{n(ABC)}{n(AB)}$
SAF = Support d'association filtré	$\frac{n(ABC)}{n(C)}$
LAR = Lift d'association renforcé	$\frac{n(ABC)}{n(AB)} \times \frac{n}{n(C)}$

Tableau 2.7 – Critères de qualité d'une règle d'association renforcée

Nous proposons d'interpréter l'influence du renforcement du motif C sur la règle $A \rightarrow B$ de la même façon que pour les motifs graduels renforcés, par principe de validité accrue : la règle doit être plus satisfaite lorsque les transactions considérées sont les transactions possédant C plutôt que toutes les transactions. Ainsi, dans le cas des motifs graduels (voir section 2.2.1, page 51), C doit servir de filtrage.

2.3.2 Critères de qualité d'une règle d'association renforcée

Une règle d'association renforcée $A \rightarrow B; C$ est évaluée en fonction de ses deux composantes, à savoir la règle d'association $A \rightarrow B$ et son renforcement C.

En ce qui concerne les règles d'association, on peut utiliser les critères classiques rappelés dans la section 1.1.1, page 19.

Pour mesurer l'intérêt du renforcement des règles d'association, nous nous basons sur les critères de qualité définis pour les motifs graduels flous renforcés dans les sections 2.1.3 et 2.2.2 et nous les transposons en utilisant la mise en correspondance suivante : étant donné $M_1; M_2$ un motif graduel renforcé et D un chemin maximal vérifiant M_1. M_1 représente la règle $A \rightarrow B$ et M_2 le renforcement C. Ainsi, une donnée appartient à D si et seulement si elle contient à la fois A et B, ce qui est représenté traditionnellement par $n(AB)$, $M_2(D)$ correspond donc au cardinal flou de l'ensemble des transactions qui contiennent à la fois A, B et C, soit $n(ABC)$. $M_2(\mathcal{D})$ correspond au cardinal flou de l'ensemble des transactions qui contiennent C dans toute la base de données, soit $n(C)$.

En appliquant cette correspondance aux équations (2.2), (2.3), (2.4) et (2.5), on obtient alors les critères de qualité indiqués dans le tableau 2.7. Ils ont le même rôle que ceux des motifs graduels renforcés. Ainsi, le support filtré permet de mesurer à quel point la règle d'association $A \rightarrow B$ est valide quand on se restreint aux transactions qui contiennent C. Il évalue le nombre de transactions qui contiennent A, B et C à la fois par rapport à la présence de C dans la base de transactions totale.

Une règle d'association renforcée est valide si tous les critères donnés dans le tableau 2.7 et le support de la règle classique sont supérieurs aux seuils fixés par l'utilisateur.

Règles d'association	Support	Confiance	Lift
$R_1 = A \rightarrow B \wedge C$	$\frac{n(ABC)}{n}$	$\frac{n(ABC)}{n(A)}$	$\frac{n(ABC)}{n(A)} \times \frac{n}{n(BC)}$
$R_2 = A \wedge B \rightarrow C$	$\frac{n(ABC)}{n}$	$\frac{n(ABC)}{n(AB)}$	$\frac{n(ABC)}{n(AB)} \times \frac{n}{n(C)}$
$R_3 = C \rightarrow A \wedge B$	$\frac{n(ABC)}{n}$	$\frac{n(ABC)}{n(C)}$	$\frac{n(ABC)}{n(C)} \times \frac{n}{n(AB)}$

Tableau 2.8 – Critères de qualité des trois règles d'association issus du motif ABC

2.3.3 Comparaison entre règles d'association classiques et renforcées

Pour étudier l'intérêt apporté par le renforcement des règles d'association, nous nous concentrons sur les trois règles d'association : $R_1 = A \rightarrow (B \wedge C)$, $R_2 = (A \wedge B) \rightarrow C$, $R_3 = C \rightarrow (A \wedge B)$ obtenues à partir du même motif ABC. Le but est alors d'étudier leurs relations et interprétations, par rapport à la règle d'association renforcée $A \rightarrow B; C$.

Le tableau 2.8 donne les critères de qualité des règles R_1, R_2, R_3.

Analyse

En comparant les critères de qualité associés à la règle d'association renforcée $A \rightarrow B; C$ et aux trois règles d'association R_1, R_2 et R_3, on constate que

1. SAR = support (R_1) = support(R_2) = support(R_3)
2. CAR = confiance(R_2)
3. SAF = confiance(R_3)
4. LR = lift(R_2) = lift(R_3)

Ainsi les critères de qualité de la règle d'association renforcée sont en relation avec ceux des deux règles R_2 et R_3. Toutefois, cela ne signifie pas qu'elles ont la même validité : celle-ci dépend aussi des seuils de leurs critères de qualité.

En effet, si les seuils sont égaux (seuils des critères de qualité des règles d'association renforcées et ceux des règles classiques), on peut conclure que lorsqu'une règle d'association renforcée $A \rightarrow B; C$ est valide, alors les deux règles d'association R_2 et R_3 sont aussi valides. La validité de la règle $A \rightarrow B; C$ implique alors la validité de (R_2 et R_3).

Si les seuils des critères des règles d'association classiques sont inférieurs à ceux des règles d'association renforcées, alors on peut obtenir les deux règles R_2 et R_3 et non la règle $A \rightarrow B; C$.

Si au contraire les seuils des critères des règles d'association renforcées sont inférieurs à ceux des règles classiques, alors on peut avoir la règle $A \rightarrow B; C$ mais pas les deux règles R_2 et R_3.

Bilan

D'après l'étude que nous avons effectuée, on constate que, sous certaines conditions de seuils de critères de qualité, deux règles d'association peuvent être exprimées par une seule règle d'association renforcée. Ainsi, nous pouvons conclure que, contrairement au cas des motifs graduels flous où le renforcement est intéressant, il ne présente pas d'apport dans le cas des règles d'association. En effet, les règles d'association ont une sémantique différente de celle des motifs graduels. Le renforcement des règles d'association est présentiel, comme le renforcement des motifs graduels flous : l'effet de renforcement est mesuré par la présence de C parmi les transactions où A et B sont présents, ce qui revient à chercher toutes les transactions où A, B et C sont présents, soit $n(ABC)$. Ce dernier correspond au support de n'importe quelle règle d'association composée de ces trois items A, B et C.

Un motif renforcé $AB; C$ revient en fait à une notion de causalité telle qu'elle est définie pour les règles d'association classiques.

2.4 Conclusion

Nous avons présenté dans ce chapitre l'enrichissement par renforcement, basé sur la méthode d'extraction des motifs graduels renforcés proposée par Bouchon-Meunier et al. (2010). Nous avons montré que les résultats des deux approches par filtrage obtiennent les mêmes résultats. Puis, nous avons proposé de compléter les propositions présentées par Bouchon-Meunier et al. (2010), en particulier les critères de qualité. Nous avons étudié leur complémentarité sur différents exemples : nous avons montré que les résultats sont meilleurs, notamment en utilisant les critères proposés dans ce chapitre, au travers de la réduction du nombre de motifs graduels renforcés extraits et de leur pertinence. Nous avons également implémenté la méthode et mis en œuvre un protocole expérimental où nous nous sommes concentrées sur les coûts en temps de calcul et occupation mémoire de la méthode.

À la fin du chapitre, nous avons posé la question d'une possible transposition de l'enrichissement par renforcement aux règles d'association, et avons ainsi proposé des critères de qualité associés ainsi qu'une nouvelle forme sémantique de règles d'association, dans le but d'apporter une nouvelle précision à ces règles, permettant de les interpréter aisément. Puis, nous avons effectué une étude par comparaison entre le renforcement de ces règles et des règles classiques. D'après cette étude, nous avons montré l'existence d'une équivalence entre une règle d'association renforcée et deux règles d'association classiques, sous certaines conditions sur les seuils minimaux des critères de qualité.

Pour finir, nous avons évoqué dans la section 3.4.1 l'apparition des motifs graduels renforcés contradictoires. Ce fait n'est pas nouveau en fouille de données : il arrive régulièrement

que les algorithmes extraient des informations contradictoires. Ce phénomène est étudié et traité dans le chapitre suivant.

3

Motifs graduels contradictoires

Sommaire

Introduction

Les méthodes d'extraction de motifs graduels peuvent générer des motifs contradictoires, produisant par exemple simultanément les motifs « plus A, plus B » et « plus A, moins B ». Ceux-ci nuisent à la qualité et à la lisibilité de l'information ainsi extraite des données. Afin de résoudre l'ambiguïté résultante, nous proposons de raffiner le critère de qualité des motifs

graduels : le nouveau critère que nous introduisons, appelé « support graduel propre global », ne dépend pas uniquement du motif considéré, mais aussi de ses contradicteurs potentiels. Cette nouvelle mesure est contrainte et permet d'extraire des motifs graduels valides sans ambiguïté d'interprétation.

Le chapitre est organisé de la façon suivante : la section 3.1 présente la définition des motifs graduels contradictoires et le principe de l'approche proposée. Dans la section 3.2, nous illustrons et formalisons le principe décrit dans la section 3.1. La section 3.3 présente le critère de qualité proposé pour évaluer la qualité des motifs graduels extraits et illustre son calcul sur des exemples de différentes complexités. La mise en œuvre du critère de qualité proposé pour l'extraction des motifs graduels fréquents est détaillée dans la section 3.4. Enfin, la section 3.5 présente les résultats expérimentaux obtenus sur des données réelles météorologiques.

Ces travaux ont été publiés dans (Oudni et al. 2012; 2013c).

3.1 Motivation et principe

Dans les approches existantes pour l'extraction des motifs graduels, rappelées dans le chapitre 1, page 34, le support graduel est défini indépendamment pour chaque motif graduel : il peut donc conduire à l'identification simultanée de motifs contradictoires, qui vérifient tous la condition de support minimum et apparaissent comme valides. Dans cette section, nous formalisons la définition de la contradiction et nous décrivons ensuite le principe général de l'approche proposée.

3.1.1 Définition formelle des motifs contradictoires

Définition 3.1 (Motifs graduels contradictoires)**.** Deux motifs graduels I et J sont dits *contradictoires* si et seulement s'ils s'écrivent sous la forme $I = MA^*$ et $J = MA^{c(*)}$ où M est un motif graduel, A un item graduel, $*, c(*) \in \{\geq, \leq\}$ et $c(*) \neq *$.

Nous introduisons la notation $\mathcal{I}_c$ utilisée par la suite, qui à un motif I associe l'ensemble des contradicteurs.

Définition 3.2 (Ensemble des contradicteurs)**.** Soit $\mathcal{I}$ l'ensemble des motifs graduels, $\mathcal{P}(\mathcal{I})$ l'ensemble des parties de $\mathcal{I}$, la fonction $\mathcal{I}_c$ est définie comme

$$\mathcal{I}_c \;:\; \mathcal{I} \;\rightarrow\; \mathcal{P}(\mathcal{I})$$

$$I \;\mapsto\; \mathcal{I}_c(I) = \{J \in \mathcal{I}, J = MA^{c(*)}\}, \text{tel que } I = MA^*$$

On peut noter qu'un motif graduel de longueur 2 possède un seul motif contradicteur, c'est-à-dire pour un motif I de longueur 2, $|\mathcal{I}_c(I)| = 1$. Dans le cas général, il peut exister de multiples contradicteurs. Ainsi le motif $I_1 = A^{\geq}B^{\geq}C^{\geq}$ est en contradiction avec $I_2 = A^{\geq}B^{\leq}C^{\geq}$ à cause de l'attribut B, mais également avec $I_3 = A^{\geq}B^{\geq}C^{\leq}$ à cause de C.

Avec cette définition, et ces notations I et J sont contradictoires si et seulement si $J \in \mathcal{I}_c(I)$.

3.1.2 Principe de la méthode proposée

Les figures 3.1 et 3.2 constituent des exemples de motifs contradictoires de longueur 2, $I = A^{\geq}B^{\geq}$ et $J = A^{\geq}B^{\leq}$: les abscisses représentent les valeurs de l'attribut A et les ordonnées celles de B. Si le seuil de support graduel est fixé à 3/8, les deux motifs I et J sont valides dans les deux cas. Ces figures illustrent cependant deux types de contradictions que nous proposons de distinguer :

1. Les deux motifs graduels ont pour support deux ensembles d'objets dont les intervalles de l'item qui diffère, c'est-à-dire l'item dont la variation est opposée, ici A, sont disjoints, comme illustré sur la figure 3.1.
2. Les deux motifs graduels ont pour support deux ensembles d'objets dont les intervalles de l'item qui diffère, ici A, sont joints, comme illustré sur la figure 3.2.

Dans le cas de la figure 3.1, la contradiction peut être gérée facilement : I couvre l'intervalle $[15; 30]$ pour les valeurs de l'attribut A et J couvre l'intervalle $[30; 40]$. Il est donc possible de distinguer deux intervalles disjoints de l'item qui diffère sur lesquels chaque motif est vérifié, ce qui permet de résoudre le problème d'ambiguïté entre les deux motifs contradictoires. En effet, chacun possède un intervalle propre.

En revanche, dans le deuxième cas illustré sur la figure 3.2, la contradiction des motifs graduels paraît difficile à gérer. En effet, le motif graduel I couvre l'intervalle $[15; 38]$ et le motif graduel J couvre l'intervalle $[18; 40]$. Ces deux intervalles ne sont pas disjoints : les deux motifs graduels contradictoires étant vérifiés sur la même plage de valeurs, il n'est pas possible de les différencier par leurs intervalles et le problème de contradiction persiste.

Il n'est donc pas souhaitable de conserver les deux motifs contradictoires dans le deuxième cas. En effet, leur présence simultanée dans l'ensemble de connaissances extraites induit des ambiguïtés dans leur interprétabilité. Ce chapitre propose une nouvelle définition de support qui permet de résoudre de manière efficace la problématique des motifs graduels contradictoires.

L'extraction conjointe de tels motifs signifie que chacun possède au moins un chemin valide qui le supporte. Nous proposons de contraindre la notion de chemin de support, afin de réduire la longueur des motifs et de limiter l'apparition de tels cas, en introduisant la notion de *chemin propre* : ce dernier est défini comme un sous-chemin d'un chemin maximal, qui induit une zone de l'espace ou d'une plage de valeurs sur laquelle un seul motif graduel est vérifié. Un tel motif est alors défini comme valide, uniquement si un tel chemin propre existe avec une taille suffisante, où la notion de taille suffisante est définie par le seuil de support. Ce chemin propre garantit l'existence d'une plage de valeurs, qui est à la fois de taille suffisante et propre au motif, c'est-à-dire sur laquelle il est seul valide.

Deux cas peuvent être distingués. D'une part, il est possible qu'au moins l'un des supports des deux motifs passe en dessous du seuil de support, c'est-à-dire qu'au moins l'un des motifs

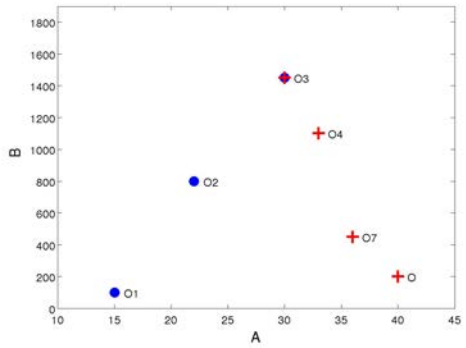


Figure 3.1 – Motifs $I = A^{\geq}B^{\geq}$ et $J = A^{\geq}B^{\leq}$ dont les uniques chemins maximaux, représentés respectivement par • et +, couvrent deux intervalles disjoints.

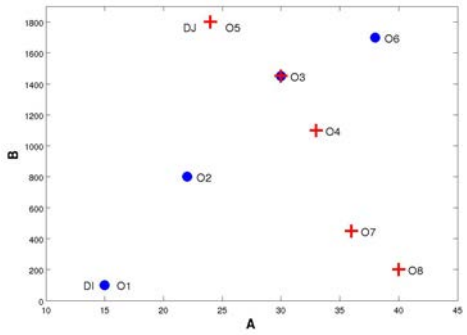


Figure 3.2 – Motifs $I = A^{\geq}B^{\geq}$ et $J = A^{\geq}B^{\leq}$ dont les uniques chemins maximaux, représentés respectivement par • et +, couvrent deux intervalles joints.

ne soit plus valide, en raison de la définition restreinte de chemin propre. Il reste donc au plus un motif valide et, par conséquent, il n'y a plus de contradiction. D'autre part, il est possible que les deux motifs possèdent un chemin propre d'une taille suffisante et qu'ils aient tous deux un support qui reste plus élevé que le seuil. Ce cas peut être décrit comme une contradiction justifiable : chaque motif est vérifié par la plage de valeurs qui lui est associée exclusivement et la caractérisation fournie par cette information résout le problème de la contradiction. Il faut souligner que la caractérisation ici est déterminée par l'attribut lui-même et non pas par un autre attribut, comme dans l'approche par renforcement discutée dans le chapitre précédent.

Pour rendre un chemin maximal propre, nous proposons de supprimer les objets qui sont responsables des chevauchements des intervalles évoqués ci-dessus. La question est alors de savoir quels objets sont à supprimer, afin d'éliminer l'ambiguïté entre les deux motifs, c'est-à-dire de définir des chemins qui couvrent des intervalles disjoints. La section 3.2 présente la définition de chemin propre respectant cette contrainte. Il reste ensuite à déterminer quels motifs graduels sont à conserver, c'est-à-dire lesquels sont réellement importants. Il est donc nécessaire de comparer leur qualité. Nous répondons à cette question dans la section 3.3, en proposant un nouveau critère de qualité appelé *support graduel propre global*, noté *SGPG*.

3.2 Définition du chemin propre

Cette section présente la notion de chemin propre, en illustrant puis en formalisant le principe de suppression d'objets introduit ci-dessus.

3.2.1 Exemples illustratifs

Nous proposons de contraindre la définition du chemin en excluant les objets ambigus qui introduisent des contradictions avec d'autres motifs, c'est-à-dire les objets qui sont responsables de l'absence d'une plage de valeurs propres. Pour l'illustrer, considérons l'exemple de

la figure 3.2 qui contient $n = 8$ objets et deux motifs graduels, $I = A^{\geq}B^{\geq}$ and $J = A^{\geq}B^{\leq}$. Chacun est supporté par un unique chemin maximal : $\mathcal{L}(I) = \{D_I\} = \{\{1, 2, 3, 6\}\}$, représenté par les $\bullet$ bleus et $\mathcal{L}(J) = \{D_J\} = \{\{5, 3, 4, 7, 8\}\}$, représenté par les $+$ rouges.

Dans cet exemple, la suppression d'objets consiste à éliminer les objets ambigus 5 et 6, responsables du chevauchement des intervalles : le chemin propre de I est restreint à $\{1, 2, 3\}$ et celui de J à $\{3, 4, 7, 8\}$, conduisant respectivement aux intervalles disjoints $[15; 30]$ et $[30; 40]$.

Le chemin propre contient donc moins d'objets que le chemin complet valide, ce qui signifie que le support du motif diminue quand les chemins propres sont considérés. Deux cas peuvent être distingués : si le seuil de support est supérieur à $3/8$, I n'est pas considéré comme valide puisque son chemin propre est de longueur 3, et le problème de contradiction est donc résolu. Si le seuil de support est inférieur à $3/8$, les deux motifs I et J sont considérés comme valides, mais ils peuvent être justifiés par leurs intervalles respectifs : la contradiction est donc résolue par cette contextualisation.

3.2.2 Formalisation du chemin propre

Dans cette section, nous formalisons la définition du chemin propre. On note $\mathcal{L}_s(I)$ l'ensemble des *chemins complets valides* associés à un motif donné I, $\mathcal{L}$ l'ensemble des chemins, I et J deux motifs graduels contradictoires, D_I et D_J leurs chemins respectifs.

Pour rendre un chemin D_I propre, on le décompose en deux sous-chemins, constitués, respectivement, des objets qui précèdent et qui succèdent (au sens du pré-ordre induit par le motif considéré, voir définition 1.12, page 36) aux objets de chevauchement. Ceux-ci sont définis comme les objets appartenant à l'intersection des chemins considérés D_I et D_J. On ne conserve ensuite que le sous-chemin le plus long. Le chemin propre contient donc soit moins d'objets que le chemin maximal comme dans le cas de l'exemple de la figure 3.1, soit le même nombre d'objets que le chemin maximal comme dans l'exemple de la figure 3.2.

Formellement, l'extraction du chemin propre d'un motif I est modélisée par la fonction appelée *propre* qui prend en argument le chemin D_I à partir duquel le chemin propre doit être extrait, et le chemin D_J par opposition auquel le chemin propre est extrait. Elle dépend de deux fonctions auxiliaires *inférieur* et *supérieur* qui renvoient, respectivement, le chemin inférieur et le chemin supérieur de D_I par opposition à D_J.

Définition 3.3 (Chemin propre)**.** Pour un motif I donné, les fonctions $supérieur_I$, $inférieur_I$ et $propre_I$ sont définis comme :

$$
\begin{array}{rcll}
supérieur_I & : & \mathcal{L} \times \mathcal{L} & \rightarrow \quad \mathcal{L} \\
 & & (D_I, D_J) & \mapsto \quad D'_I = \left\{ \begin{array}{ll} \emptyset & \text{si } D_I \cap D_J = \emptyset \\ \{o \in D_I / \forall o' \in D_I \cap D_J, o \succeq_I o'\} & \text{sinon} \end{array} \right.
\end{array}
$$

$$\begin{array}{rcccl} inférieur_I & : & \mathcal{L}\times\mathcal{L} & \rightarrow & \mathcal{L} \\ & & (D_I, D_J) & \mapsto & D'_I = \begin{cases} \emptyset & \text{si } D_I \cap D_J = \emptyset \\ \{o \in D_I / \forall o' \in D_I \cap D_J, o \preceq_I o'\} & \text{sinon} \end{cases} \end{array}$$

Le *chemin propre* de D_I par opposition à D_J est alors le sous-ensemble maximal de D_I qui n'entre pas en contradiction avec D_J :

$$\begin{array}{rcccl} propre_I & : & \mathcal{L}\times\mathcal{L} & \rightarrow & \mathcal{L} \\ & & (D_I, D_J) & \mapsto & \begin{cases} supérieur_I(D_I, D_J) \text{ si} \\ \qquad |inférieur_I(D_I, D_J)| \leq |supérieur_I(D_I, D_J)| \\ inférieur_I(D_I, D_J) \text{ sinon} \end{cases} \end{array}$$

Il faut noter que l'on peut calculer simultanément les deux résultats $propre_I(D_I, D_J)$ et $propre_J(D_J, D_I)$.

3.2.3 Agrégation : chemin propre global

Le calcul du chemin propre défini dans la section précédente est appliqué sur un unique chemin. Or, le plus souvent, un motif graduel possède plusieurs chemins contradicteurs, et il faut donc généraliser la définition précédente à ce cas. Cette section est consacrée à cette généralisation, qui repose sur l'agrégation des chemins contradicteurs en un *chemin propre global* prenant en compte l'existence de multiples chemins contradicteurs.

Ces chemins sont liés à un ou plusieurs motifs contradictoires de I. Cette fonction est illustrée dans la section 3.3.2 sur divers cas.

Définition 3.4 (Chemin propre global)**.** Pour un seuil $s \in [0, 1]$, la fonction $propreGlobal_s$ est définie comme :

$$\begin{array}{rcccl} propreGlobal_s & : & \mathcal{L} & \rightarrow & \mathcal{L} \\ & & D_I & \mapsto & propreGlobal_s(D_I) = \bigcap_{\substack{J \in \mathcal{I}_c(I) \\ D_J \in \mathcal{L}_s(J)}} propre_I(D_I, D_J) \end{array} \tag{3.1}$$

La fonction $propreGlobal_s$ dépend du seuil s, car on s'intéresse uniquement aux chemins de longueur suffisante qui pourraient conduire à la validation des motifs contradictoires.

En effet, il est important de noter que seuls les chemins contradictoires de taille suffisante $D_J \in \mathcal{L}_s(J)$ sont pris en compte. Les autres chemins sont considérés comme peu importants et pourraient réduire D_I de façon drastique : si tous les chemins complets, y compris ceux qui ne sont pas valides, $\mathcal{L}(J)$, étaient considérés, alors un nombre trop important de chemins propres serait obtenu. L'intersection de ces derniers pourrait alors conduire à un chemin propre global de taille faible, pouvant ne pas vérifier le seuil du support minimum.

3.3 Critère de qualité : le support graduel propre global

Dans cette section, nous formalisons le critère de qualité proposé pour évaluer la qualité des motifs graduels extraits, appelé le support graduel propre global et noté $SGPG$. Nous illustrons ensuite son calcul sur des exemples de complexité croissante. Puis, nous présentons l'algorithme de calcul du $SGPG$.

3.3.1 Définition

Définition 3.5 (Support graduel propre global)**.** Pour un motif I et une valeur $s \in [0, 1]$ fixée, le support graduel propre global, $SGPG$, est défini comme :

$$SGPG_s(I) = \frac{1}{|\mathcal{D}|} \max_{D \in \mathcal{L}(I)} |propreGlobal_s(D)| \tag{3.2}$$

Il est équivalent au support graduel défini dans l'équation (1.7), page 40, mais s'appuie sur les chemins propres globaux au lieu des chemins complets. Ainsi, le support d'un motif ne dépend pas seulement de lui-même, mais aussi des motifs qui le contredisent.

Il est important de souligner qu'habituellement, le support graduel, SG, est calculable uniquement à partir des données. Au contraire, le critère de qualité que nous proposons, $SGPG$, dépend de la valeur choisie pour le paramètre s, car son calcul s'appuie sur la fonction *propreGlobal* qui elle-même dépend de cette valeur. Ceci implique qu'un changement d'exigence vis-à-vis du seuil amène à recalculer le support. De plus, le calcul de ce support ne s'appuie pas forcément sur les chemins maximaux, comme nous le commentons plus loin à la page 77.

Par exemple, le calcul du support $SGPG(I)$ du motif graduel $I = A^{\geq}B^{\geq}$, illustré sur la figure 3.2, dépend de son contradicteur $J = A^{\geq}B^{\leq}$, plus précisément, de son chemin contradicteur D_J. Il vaut $3/8 = 37\%$, alors que son support graduel vaut $4/8 = 50\%$.

Un motif I est ensuite dit *valide* si est seulement si son $SGPG$ est supérieur au seuil s qui est le même que pour le support graduel : il doit posséder un chemin *propre* de longueur suffisante par opposition à tous ses contradicteurs simultanément.

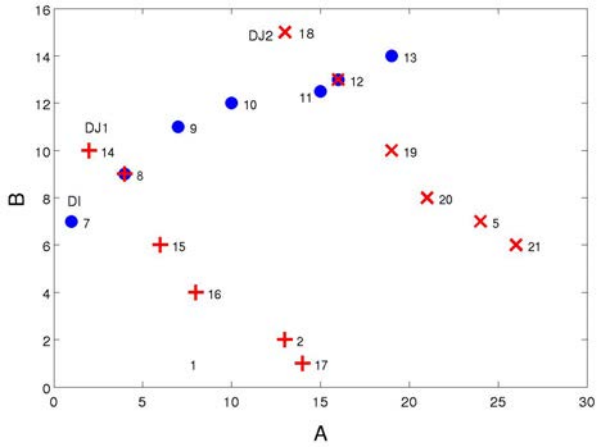


Figure 3.3 – Motif graduel, $I = A^{\geq}B^{\geq}$, supporté par le chemin D_I représenté par $\bullet$ et avec plusieurs chemins contradictoires, D_{J1} représenté par $+$ et D_{J2} représenté par $\times$.

3.3.2 Exemples illustratifs

Dans ce qui suit, nous illustrons et commentons les définitions précédentes avec des exemples de complexité croissante. Nous illustrons et examinons :

— le cas de l'utilisation de multiples chemins contradictoires, en particulier la nécessité d'utiliser l'intersection des chemins propres dans l'équation (3.1) ;

— le cas de la présence simultanée de multiples chemins contradictoires et de chemins complets valides vérifiant le motif considéré, justifiant la présence du maximum dans l'équation (3.2) ;

— la présence de $\mathcal{L}(I)$ dans l'équation (3.2) et non $\mathcal{L}^*(I)$: ceci est justifié par l'analyse du cas où un chemin complet valide permet d'obtenir un chemin propre global de longueur supérieure à celui obtenu avec un chemin maximal.

Le cas le plus simple correspondant à des motifs graduels de longueur 2 supportés par un seul chemin maximal a été déjà illustré dans les sections précédentes par l'exemple de la figure 3.2, qui représente le cas de référence dans toutes les sections précédentes.

Prise en compte de multiples chemins contradictoires

Dans le cas général, il peut y avoir plusieurs chemins complets valides générant des contradictions, comme la figure 3.3 l'illustre, pour $I = A^{\geq}B^{\geq}$ et $J = A^{\geq}B^{\leq}$: I est supporté par le seul chemin maximal D_I (représenté par les $\bullet$ bleus), alors que J, qui le contredit, est supporté par deux chemins maximaux D_{J1} (représenté par les $+$ rouges) et D_{J2} (représenté par les $\times$ rouges). Deux possibilités peuvent être considérées, conduisant à deux chemins propres : en calculant l'intersection de D_I avec D_{J1}, on obtient $D_{g1} = \ propre(D_I, D_{J1}) = \{8, 9, 10, 11, 12, 13\}$; en considérant D_{J2}, on obtient $D_{g2} = propre(D_I, D_{J2}) = \{7, 8, 9, 10, 11, 12\}$. Comme indiqué dans l'équation (3.1), nous les combinons en calculant leur intersection, ce qui conduit au chemin propre global $D_g = \{8, 9, 10, 11, 12\}$.

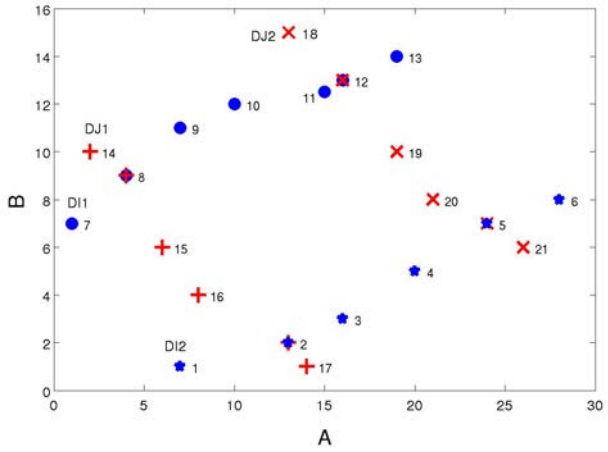


Figure 3.4 – Motif graduel, $I = A^{\geq}B^{\geq}$, supporté par plusieurs chemins D_{I1}, représenté par $\bullet$ et D_{I2} représenté par $*$, et avec plusieurs chemins contradictoires, D_{J1} représenté par des $+$ et D_{J2} représenté par des $\times$.

L'intersection proposée dans l'équation (3.1) représente l'opérateur d'agrégation des chemins propres. En effet, tous les chemins complets valides de tous les contradicteurs sont pris en compte et chacun induit un chemin propre. Le fait que D_{J1} et D_{J2} soient issus d'un même contradicteur, comme dans l'exemple précédent, ou de plusieurs contradicteurs, ne change pas le calcul du chemin propre global.

Prise en compte de multiples chemins

Dans le cas plus général, le motif à traiter est lui-même supporté par plusieurs chemins, comme illustré sur la figure 3.4 : les deux chemins maximaux D_{I1} (représenté par les $\bullet$ bleus) et D_{I2} (représenté par les $*$ bleus) supportent le motif I qui est en contradiction avec le motif J, supporté par D_{J1} et D_{J2}. Afin de calculer le chemin propre global de I, on rend propres tous ses chemins : par opposition à D_{J1} et D_{J2}, D_{I1} conduit à deux chemins propres dont l'intersection $D_{I1g} = \{8, 9, 10, 11, 12\}$ représente le chemin propre global qui lui est associé. On traite de même D_{I2}, ce qui conduit à $D_{I2g} = \{2, 3, 4, 5\}$. Enfin on conserve les chemins propres globaux de longueur maximale (cf. équation (3.2)), ici D_{I1g}. Le $SGPG(I)$ est $5/21 = 24\%$.

Prise en compte des chemins complets

Dans la définition de chemin propre global de l'équation (3.2), on prend en compte tous les chemins complets $\mathcal{L}(I)$ et non uniquement les chemins maximaux $\mathcal{L}^*(I)$. En effet, il peut exister des chemins non maximaux qui, à l'issue du traitement qui les rend propres, sont de longueur supérieure aux chemins issus des chemins maximaux.

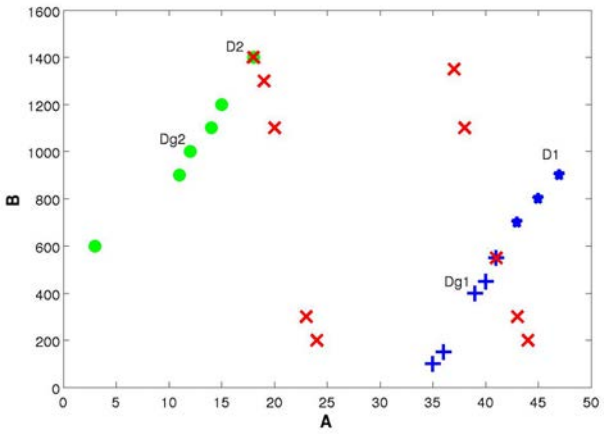


Figure 3.5 – Motif graduel de longueur 2, $I = A^{\geq}B^{\geq}$, supporté par un unique chemin maximal D_1, représenté par des $*$ et $+$, et un chemin complet valide D_2 représenté par des $\bullet$.

3.3.3 Algorithme de calcul des chemins propres globaux

Après avoir défini formellement le support contraint, nous considérons dans cette sous-section la question de son calcul efficace. La procédure que nous proposons est basée sur le choix de l'ordre dans lequel les chemins complets sont rendus propres, qui est défini par l'ordre décroissant de longueur. De la sorte, on peut en effet définir un critère d'arrêt efficace : on peut cesser le traitement dès que la longueur du chemin suivant à traiter est inférieure à la longueur maximale du chemin propre global issu des chemins traités précédemment. En effet, les chemins propres globaux que l'on pourrait obtenir par la suite auraient alors nécessairement des longueurs inférieures et ne seraient pas conservés.

Il faut noter que, si l'on cherche à identifier les motifs graduels valides et non à calculer le $SGPG$ pour tous les motifs, il n'est pas nécessaire de considérer la totalité des chemins complets. On peut en effet se restreindre aux chemins complets valides, c'est-à-dire, formellement, remplacer $\mathcal{L}(I)$ par $\mathcal{L}_s(I)$ dans l'équation (3.2). En effet, après application de la fonction qui les rend propres, les chemins de longueur inférieure à s sont plus petits encore. Ils ne peuvent donc pas constituer un support du motif par rapport au paramètre s.

La méthode décrite ci-dessus est implémentée par l'algorithme 3, qui prend en entrée un motif graduel I, l'ensemble de ses chemins complets $\mathcal{L}(I)$ et ses contradicteurs $\mathcal{I}_c(I)$, ainsi que leurs chemins $\mathcal{L}_s(\mathcal{I}_c(I))$. Le résultat de cet algorithme est le support graduel propre global du motif considéré, $SGPG$.

Il faut noter que cet algorithme peut renvoyer également les chemins propres globaux associés.

3.4 Mise en œuvre pour l'extraction de motifs

Dans la section précédente, nous avons proposé une nouvelle définition du support qui permet d'évaluer les motifs graduels contradictoires ainsi qu'un algorithme efficace pour le

Algorithm 3 Algorithme de calcul du $SGPG$

Input : un motif graduel I, l'ensemble de ses chemins complets $\mathcal{L}(I)$ et ses contradicteurs $\mathcal{I}_c(I)$, ainsi que leurs chemins $\mathcal{L}_s(\mathcal{I}_c(I))$.
Output : $SGPG(I)$
$l^* = 1$
$k = 1$
Tri des $D \in \mathcal{L}(I)$ par taille décroissante $l_1 > l_2 > ... > l_p$
while $k \leq p$ **and** $l_k > l^*$ **do**
  for all $D \in \mathcal{L}(I)$ tel que $|D| = l_k$ **do**
    calculer $D' = propreGlobal_s(D)$ selon l'équation (3.2)
    $l^* \leftarrow \max(l^*, |D'|)$
  end for
  $k \leftarrow k + 1$
end while
$SGPG = l^*$

calcul de ce support. Dans cette section, nous considérons le problème d'extraction de motifs graduels fréquents selon la nouvelle définition du support : nous expliquons comment le traitement des motifs contradictoires, tel que présenté dans la section 3.2, peut être appliqué à l'ensemble des contradictions et à quel niveau ce traitement est appliqué.

Nous proposons deux approches : la première, dite par filtrage a posteriori, consiste à appliquer le traitement après que tous les motifs ont été générés, afin de supprimer les contradictions a posteriori. La seconde, dite méthode intégrée, permet d'intégrer ce traitement dans la phase de génération et de l'appliquer comme un filtrage avant la génération des motifs du niveau supérieur. Les deux approches reposent sur l'algorithme GRITE (Di Jorio et al., 2009) qui associe aux motifs qu'il extrait leurs chemins complets valides (voir rappel dans son principe dans la section 1.3.3, page 41).

3.4.1 Filtrage a posteriori

La première approche consiste à considérer l'ensemble des motifs extraits par l'algorithme GRITE, puis à supprimer les contradictions a posteriori grâce aux informations extraites sur les chemins complets valides : la méthode consiste à calculer les $SGPG$ de tous les motifs contradictoires extraits, selon l'approche décrite dans la section 3.1. On effectue ensuite une élimination progressive des motifs contradictoires.

Le but de cette dernière est, d'une part, de garder uniquement les motifs graduels contradictoires ayant un $SGPG$ élevé, et d'autre part, de s'assurer que l'on n'élimine pas de motifs graduels de $SGPG$ élevé. Ce principe permet donc d'éliminer moins de motifs.

On pourrait en effet envisager de valider les motifs ayant un $SGPG$ supérieur à s et de supprimer tous les autres, mais cette approche est brutale : si l'on considère par exemple trois motifs contradictoires I, J et K, tels que tous les trois ont un $SGPG$ inférieur à s, la considération du seuil de $SGPG$ conduit à éliminer les trois motifs dès lors que la valeur est

Motif	SG	I_1	I_2	I_3	I_4	$SGPG$
I_1	$31,25\%$	-	21%	25%	-	15.63%
I_2	21%	12.5%	-	-	$18,75\%$	12.5%
I_3	21%	15.63%	-	-	21%	15.63%
I_4	23%	-	$18,75\%$	21%	-	$18,75\%$

Tableau 3.1 – Supports graduels propres des quatre motifs I_1, I_2, I_3 et I_4 pour les données du tableau A.1, page 146.

en dessous du seuil. Toutefois, il se peut que le support propre de I par opposition à J seul, soit supérieur à s : si K est éliminé, I peut être conservé.

Aussi nous proposons d'éliminer les motifs progressivement, en introduisant une priorité de traitement basée sur l'ordre croissant des $SGPG$, selon la démarche suivante en trois étapes :

— suppression du motif de $SGPG$ minimal

— mise à jour des $SGPG$ des motifs restants, en ignorant, dans l'équation (3.2), les contradicteurs qui ont été éliminés

— itération jusqu'à ce que tous les motifs restants vérifient $SGPG > s$

On peut noter que l'arrêt de ces itérations est garanti puisqu'à chaque itération, moins de motifs sont conservés et que leur $SGPG$ croît : on atteint nécessairement une étape où il n'y a plus de motifs dont le $SGPG$ est inférieur au seuil.

Exemple illustratif

Illustrons le déroulement de l'approche en utilisant la base de données jouet illustrée dans le tableau A.1 donné en annexe en page 145, qui contient 31 objets décrits par 3 attributs, A, B et C. Le seuil du support utilisé est $s = 18\%$.

Nous considérons quatre motifs graduels contradictoires : $I_1 = A^{\geq}B^{\geq}C^{\geq}$, $I_2 = A^{\geq}B^{\geq}C^{\leq}$, $I_3 = A^{\geq}B^{\leq}C^{\geq}$ et $I_4 = A^{\geq}B^{\leq}C^{\leq}$ générés par GRITE.

Les résultats sont représentés dans le tableau 3.1 : chaque ligne contient un motif. La colonne SG représente leurs supports graduels classiques respectifs. La case M_{ij} du tableau contient le $SGPG$ de I_i en considérant I_j comme unique contradicteur, c'est-à-dire en appliquant l'équation (3.2) et en remplaçant dans l'équation (3.1) $J \in \mathcal{I}_c(I_i)$ par $J = I_j$. Lorsque deux motifs ne sont pas en contradiction, on met le caractère $-$ dans la case correspondante.

Le motif I_4 est validé directement car son $SGPG$ est supérieur à s. Il est donc retiré des lignes et conservé dans les colonnes du tableau 3.1, puisqu'il est en contradiction avec I_3.

I_2 ayant le $SGPG$ minimum, inférieur à s est supprimé. Après mise à jour des $SGPG$, on obtient les résultats représentés dans le tableau 3.2.

Motif	SG	I_1	I_3	I_4	$SGPG$
I_1	$31,25\%$	-	25%	-	25%
I_3	21%	15.63%	-	$18,75\%$	15.63%

Tableau 3.2 – Mise à jour des $SGPG$

À l'issue de ce traitement, I_1 est validé et il n'est donc plus dans la liste des motifs à traiter. Pour I_3, le $SGPG$ reste inférieur à s, il est donc éliminé. Ainsi, I_1 et I_4 sont conservés.

On peut noter que si la priorité de traitement n'avait pas été intégrée dans le filtrage a posteriori, les trois motifs graduels I_1, I_2 et I_3 auraient été supprimés au lieu de deux.

Discussion sur la priorité de traitement

Une problématique importante concerne l'ordre dans lequel les motifs sont traités : si un ordre aléatoire est considéré, alors cela peut conduire à la suppression de tous les motifs. Pour cela, nous avons défini un critère de priorité d'ordre de traitement qui est l'ordre croissant des $SGPG$ présenté dans la section 3.4.1. Nous discutons ici la possibilité de considérer d'autres critères et expliquons en quoi ces derniers ne constituent pas un bon choix.

D'autres critères peuvent en effet être envisagés, comme la suppression par ordre croissant

— des minimums des supports graduels propres, c'est-à-dire des minimums des valeurs représentées dans les cases M_{ij} pour un i donné dans le tableau 3.1.

— des moyennes des supports graduels propres, c'est-à-dire des moyennes des valeurs représentées dans les cases M_{ij} pour un i donné dans le tableau 3.1.

— des maximums des supports graduels propres, c'est-à-dire des maximums des valeurs représentées dans les cases M_{ij} pour un i donné dans le tableau 3.1.

— des supports graduels, SG.

Nous avons choisi et privilégié le critère de priorité par ordre croissant du $SGPG$, car il est le seul à traduire l'information du chemin propre global qui détermine la validité d'un motif : le $SGPG$ est inférieur ou égal aux quatre critères cités ci-dessus, et il est le seul critère qui garantit la non contradiction des motifs graduels conservés. En effet, si un autre critère est considéré et qu'on supprime le motif ayant la valeur minimale de ce critère, cela ne signifie pas qu'on a supprimé le motif le moins pertinent et le plus ambigu, puisque tout autre critère ne traduit pas l'information de chemin propre global.

3.4.2 Approche intégrée

Le but de la seconde approche est d'éviter de générer des motifs de longueur $k+1$ qui s'appuient sur des motifs de longueur k contradictoires, donc susceptibles d'être éliminés, pour filtrer davantage en cours de génération. En d'autres termes, la génération des motifs

du niveau $k+1$ s'appuie uniquement sur les motifs du niveau k dont le $SGPG$ vérifie la condition de support minimum.

Cette information est intégrée dans le processus de génération, plus précisément dans la matrice de concordance manipulée par l'algorithme (Di Jorio et al., 2009) comme rappelé dans le chapitre 1, page 41 : celle-ci est modifiée pour représenter uniquement des chemins propres et non l'ensemble des chemins support. Aussi, les objets qui appartiennent à un chemin complet valide mais qui n'appartiennent pas à un chemin propre global sont supprimés, afin qu'ils ne soient pas considérés comme candidats de support dans la génération des motifs de niveau supérieur.

Il est important de noter qu'au niveau $k = 2$, les motifs valides sont extraits suivant l'approche a posteriori, et qu'ensuite le principe de l'approche intégrée est appliqué pour générer les motifs valides de niveau supérieur.

Formalisation de la mise à jour des matrices de concordance

Définition 3.6 (Matrice de concordance propre)**.** Soit I un motif et M sa matrice de concordance. La matrice de concordance propre M' est initialisée comme égale à M, puis pour chaque chemin complet valide D, en notant D_g le chemin propre global correspondant : si un objet o appartient à D et pas à D_g, c'est-à-dire si $(o_i \in D) \wedge (o_i \notin D_g)$ alors $\forall j, M_{ij} = M_{ji} = 0$.

Cette matrice propre est exploitée pour générer les motifs de niveau supérieur.

Il faut noter que les composantes de la matrice correspondant à des objets appartenant à un chemin complet non valide ne sont pas modifiées, c'est-à-dire qu'on n'apporte aucune modification sur les autres composantes de la matrice initiale : seules les composantes correspondant aux objets ambigus sont modifiées. Ceci permet de conserver une flexibilité pour la recherche de chemins lors du traitement des motifs de longueur $k+1$: un chemin vérifiant un motif de longueur supérieure qui ne s'appuie pas forcément sur les objets du chemin propre global du motif de longueur inférieure pourrait être identifié.

Exemple illustratif

La base de données utilisée ici est la même que celle utilisée pour illustrer l'approche a posteriori dans la section précédente.

Toutefois, on n'illustre pas les mêmes motifs que ceux de la section précédente, car la mise à jour des matrices de concordance se fait dès le niveau 2, pour éviter de générer des motifs contradictoires de longueur 3. Les exemples exploités pour cette illustration sont donc des motifs de longueur 2. En revanche, dans l'approche a posteriori, il est nécessaire d'exploiter des motifs de longueur supérieure à 2 pour illustrer le cas de multiples contradicteurs.

Parmi les motifs extraits à partir de cette base de données, nous considérons les deux motifs contradictoires $I = A^{\geq}B^{\geq}$ et $J = A^{\geq}B^{\leq}$. Déroulons maintenant la fonction de mise à jour des matrices de concordance sur celle qui correspond au motif $A^{\geq}B^{\geq}$, donnée dans le tableau 3.3. Dans le but de mieux visualiser la matrice, les 0 ne sont pas représentés.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0					1	1	1	1	1	1	1	1	1	1	1	1	1	1		**1**	1	1	1	1	1	1	1			1	1
1					1	1	1	1	1	1	1	1	1	1	1	1	1	1		**1**	1	1	1	1	1	1	1			1	1
2					1	1	1	1	1	1	1	1	1	1	1	1	1	1		**1**	1	1	1	1	1	1	1			1	1
3					1	1	1	1	1	1	1	1	1	1	1	1	1	1		**1**	1	1	1	1	1	1	1			1	1
4						1	1	1	1				1	1	1	1	1	1		1	1	1	1	1						**1**	
5							1		1				1	1	1	**1**	1	1			1	1		1							
6																						1									
7									1														1	1							
8																								1							
9							1	1	1													1	1	1	1						
10							1	1	1	1		1										1	1	1	1	1	1				
11									1															1							
12																															
13													1				1	1													
14													1				**1**	1													
15													1		**1**		1	1													
16													1					**1**													
17													**1**																		
18													1	1	1	1	1	1													
19					**1**	1	1	1	1				1	1	1	1	1	1			1	1	1	1						1	
20						**1**	1		1				1	1	1	1	1	1				1		1							
21							1																								
22								1	1															1							
23									1																						
24							1	1	1	1												1	1	1							
25							1	1	1	1	1	1										1	1	1	1		1				
26							1	1	1	1	1	1										1	1	1	1	1					
27													1	1	1	1	1	1													
28													1	1	1	1	1	1										1			
29						1	1	1	1				1	1	1	1	1	1			1	1	1	1						1	
30													1	1	1	1	1	1													

Tableau 3.3 – Matrice de concordance correspondant au motif graduel $A^{\geq}B^{\geq}$.

Pour chaque chemin complet valide supportant le motif I, nous calculons le chemin propre global qui lui correspond. Dans cet exemple, I est vérifié par quatre chemins maximaux : $D_{I_1} = \{0, 19, 4, 29, 20, 5, 15, 14, 16, 17, 12\}$, $D_{I_2} = \{1, 19, 4, 29, 20, 5, 15, 14, 16, 17, 12\}$, $D_{I_3} = \{2, 19, 4, 29, 20, 5, 15, 14, 16, 17, 12\}$ et $D_{I_4} = \{3, 19, 4, 29, 20, 5, 15, 14, 16, 17, 12\}$. Ces quatre chemins diffèrent uniquement par leur premier objet, représenté respectivement en bleu, en vert, en rouge et en en magenta dans la matrice de concordance. Les objets restants sont représentés en bleu.

— le chemin propre global de D_{I_1} par opposition à D_J est $D_{1g} = \{19, 4, 29, 20, 5, 15, 14, 16, 17, 12\}$, donc l'objet 0 est supprimé.

— le chemin propre global de D_{I_2} par opposition à D_J est $D_{2g} = \{19, 4, 29, 20, 5, 15, 14, 16, 17, 12\}$, donc l'objet 1 est supprimé.

— le chemin propre global de D_{I_3} par opposition à D_J est $D_{3g} = \{19, 4, 29, 20, 5, 15, 14, 16, 17, 12\}$, donc l'objet 2 est supprimé.

— le chemin propre global de D_{I_4} par opposition à D_J est $D_{4g} = \{19, 4, 29, 20, 5, 15, 14, 16, 17, 12\}$, donc l'objet 3 est supprimé.

Dans cet exemple, tous les chemins conduisent au même chemin propre global.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0					0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0			0	0
1					0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0			0	0
2					0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0			0	0
3					0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0			0	0
4						1	1	1	1				1	1	1	1	1	1		1	1	1	1	1						**1**	
5							1		1				1	1	1	**1**	1	1			1	1		1							
6																						1									
7									1														1	1							
8																								1							
9							1	1	1													1	1	1	1						
10							1	1	1	1		1										1	1	1	1	1	1				
11									1															1							
12																															
13													1				1	1													
14													1				**1**	1													
15													1		**1**		1	1													
16													1					**1**													
17													**1**																		
18													1	1	1	1	1	1													
19					**1**	1	1	1	1				1	1	1	1	1	1			1	1	1	1						1	
20						**1**	1		1				1	1	1	1	1	1				1		1							
21							1																								
22								1	1															1							
23									1																						
24							1	1	1	1												1	1	1							
25							1	1	1	1	1	1										1	1	1	1		1				
26							1	1	1	1	1	1										1	1	1	1	1					
27													1	1	1	1	1	1													
28													1	1	1	1	1	1										1			
29						1	1	1	1				1	1	1	1	1	1			1	1	1	1							
30													1	1	1	1	1	1													

Tableau 3.4 – Mise à jour de la matrice de concordance du motif graduel $A^{\geq}B^{\geq}$.

La mise à jour de la matrice de concordance du motif I entraîne donc la suppression des 4 objets ambigus $\{0, 1, 2, 3\}$, ce qui conduit à la matrice présentée dans le tableau 3.4.

De la même façon, ce traitement permet de mettre à jour la matrice de concordance du motif J.

La matrice propre possède un nombre de 1 inférieur à celui de la matrice de départ, car les objets problématiques ont été supprimés afin de les empêcher de reproduire le problème au niveau supérieur. On remarque que suffisamment d'objets appartenant aux chemins complets valides de I sont à 1, ce qui va permettre de trouver un support suffisant pour des motifs de longueur supérieure ou égale à 3, basés sur le motif $I = A^{\geq}B^{\geq}$.

Il est de plus important de noter que, pour obtenir, avec l'approche intégrée, le motif graduel de niveau 3, $I_1 = A^{\geq}B^{\geq}C^{\geq}$, il faut traiter également les deux motifs contradictoires de longueur 2, $A^{\geq}C^{\geq}$ et $A^{\geq}C^{\leq}$ au niveau 2 pour établir leur matrice de concordance propre. La génération du motif I_1 est donc basée sur des matrices de concordance propres, ce qui conduit à l'extraction du motif graduel I_1 sans ambiguïté.

Afin de vérifier que le motif I_1 n'entre pas en contradiction avec son contradicteur validé I_4, qui est également un motif validé avec l'approche a posteriori, et que les autres motifs, I_2 et

I_3 supprimés par l'approche a posteriori, n'ont pas de chemin propre global de taille suffisante permettant de les valider, le même principe appliqué aux motifs $A^{\geq}B^{\geq}$ et $A^{\geq}B^{\leq}$ doit être appliqué à tous les motifs de longueur 2 sur lesquels sont basés les motifs I_2, I_3 et I_4. Ce long processus n'est pas détaillé ici.

3.5 Résultats expérimentaux

Nous avons réalisé une étude expérimentale du support graduel propre global proposé, ainsi que des deux méthodes proposées pour l'extraction des motifs graduels fréquents non contradictoires, pour comparer à la fois les motifs extraits et les performances en termes de temps de calcul et d'occupation mémoire.

Pour ces expérimentations, nous avons considéré la base de données réelles météorologiques présentée en annexe A.1, page 145.

3.5.1 Exemples de motifs graduels extraits

Nous fixons dans ces expérimentations le support graduel minimum à $s = 20\%$ pour les deux approches proposées.

Les résultats de nos expérimentations montrent que la contradiction est bien éliminée et que seuls les motifs graduels non ambigus sont extraits. En effet comme attendu, on observe que, soit les contradicteurs sont supprimés, soit ils sont conservés mais ils sont vérifiés par des sous-ensembles d'objets qui leur sont propres et qui induisent des intervalles disjoints. Voici quelques motifs graduels non ambigus extraits ordonnés par ordre décroissant de leur $SGPG$.

— Plus la température est élevée, plus l'humidité est élevée
$SG = 70\%$, $SGPG = 33\%$

— Plus la vitesse du vent est élevée, plus la pression est élevée
$SG = 61.5\%$, $SGPG = 25.2\%$

— Plus la température est élevée, moins la quantité de la pluie est élevée
$SG = 51\%$, $SGPG = 23.2\%$

— Plus la température est élevée, plus la quantité de la pluie est élevée
$SG = 48\%$, $SGPG = 21\%$

Pour les deux premiers exemples, leur contradicteurs respectifs (« plus la température est élevée, moins l'humidité est élevée » : $SGPG = 12,5\%$, « plus la vitesse du vent est élevée, moins la pression est élevée » : $SGPG = 10.2\%$) ont été supprimés, faute d'avoir un $SGPG$ suffisant. Au contraire, les deux derniers motifs illustrés annoncent deux informations contradictoires, mais les deux sont validés car leur support graduel propre global est supérieur au seuil de support minimal. Chacun a un sous-ensemble d'objets propre qui induit un intervalle sur lequel il est seul validé : le troisième motif est vérifié pour une quantité de pluie $\in [0.13, 0.27]$ et le quatrième motif pour une quantité de pluie $\in [0.27, 0.5]$; les deux intervalles

propres sont disjoints. Ainsi, ils ne sont pas réellement contradictoires, mais s'appliquent à différents sous-ensembles de données.

On peut observer une très forte diminution des valeurs des supports graduels, illustrée par exemple sur ces quatre motifs.

3.5.2 Comparaison des deux approches d'extraction

Le tableau 3.5 quantifie les résultats obtenus, en indiquant le nombre de motifs générés, contradictoires et validés pour chacune des deux méthodes d'extraction proposées ; le tableau 3.6 détaille le nombre de motifs validés selon la longueur des motifs pour les deux approches. Il donne aussi des indications sur la longueur des motifs graduels écartés par la méthode intégrée.

Pour l'approche a posteriori, les 835 motifs générés sont ceux de GRITE. Ils correspondent à 717 motifs possédant au moins un contradicteur, ainsi la proportion de motifs contradictoires comparée au nombre de motifs générés est de 86%. Les 118 autres motifs n'ont pas de contradicteur et sont valides. À la fin du processus de traitement, 546 motifs graduels sont validés, c'est-à-dire que 428 motifs correspondent à des motifs graduels contradictoires validés. Nous observons que, même si la proportion des motifs contradictoires est élevée, la proportion des motifs contradictoires validés est aussi élevée : elle est de $428/717 = 60\%$. Ceci signifie beaucoup d'entre eux possèdent en fait une plage de valeurs propres où ils sont seuls vérifiés.

Par construction, il n'y pas de différence entre les approches pour les motifs de longueur 2. La proportion de suppression par rapport aux motifs contradictoires vaut 36%. En revanche, aucune suppression n'est produite à partir du niveau 3 avec l'approche intégrée, contrairement à l'approche a posteriori où 64% des motifs de longueur supérieure à 3 sont supprimés. La proportion de suppression totale par rapport au nombre de motifs générés dans l'approche intégrée est égale à 8%. Elle est donc, comme attendu, inférieure à celle de l'approche a posteriori qui est égale à 35%.

Globalement, on constate que la méthode intégrée génère légèrement moins de motifs, (11 de moins), que la méthode a posteriori, en identifiant un sous-ensemble seulement. Cette différence est due au fait que la méthode intégrée modifie les matrices de concordance dès le niveau 2 : lors de ces modifications, des objets qui appartiennent à un chemin complet valide au niveau inférieur sont supprimés. Aussi, un chemin qui aurait pu être support d'un motif de niveau supérieur ne l'est plus. Plus précisément, le problème vient du choix du sous-chemin, inférieur ou supérieur, dans le calcul des chemins propres : le choix optimal au niveau 2 ne l'est pas nécessairement pour les niveaux suivants. En effet, lors du couplage avec un item graduel supplémentaire, il est possible que le sous-chemin rejeté se révèle préférable au sous-chemin choisi.

À partir de la longueur 6, les motifs validés sont les mêmes dans les deux méthodes, ce qui s'explique par le fait que les motifs de longueur 5 obtenus uniquement dans l'approche a posteriori ne conduisent pas à des motifs de longueur 6.

	♯ de motifs générés	♯ de motifs contradictoires	♯ de motifs validés
A posteriori	835	717	546
Intégrée	583	132	535

Tableau 3.5 – Nombre de motifs générés, contradictoires et validés pour les deux approches.

	Longueur des motifs					
	2	3	4	5	6	7
A posteriori	84	148	118	118	66	12
Intégrée	84	144	113	116	66	12

Tableau 3.6 – Nombre de motifs validés par niveau pour les deux approches.

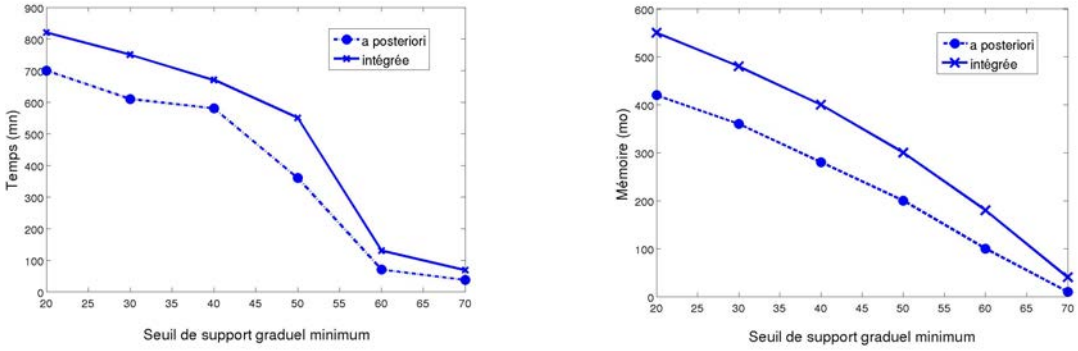


Figure 3.6 – Temps de calcul et consommation mémoire en fonction du seuil de support minimal pour les deux approches.

Le tableau 3.7 montre le nombre de motifs extraits par les deux méthodes pour différents seuils de support : on observe que, comme précédemment, les méthodes sont proches quelle que soit la valeur de s et que c'est toujours l'approche intégrée qui extrait légèrement moins de motifs. Néanmoins, quand on utilise un seuil de support élevé, les deux approches obtiennent exactement le même résultat.

3.5.3 Évaluation des performances

Nous avons également comparé les performances des deux approches en termes de temps et de mémoire utilisés. La figure 3.6 montre cette comparaison des performances en fonction du seuil de support minimal.

On observe que la consommation de la mémoire est plus élevée pour l'approche intégrée : environ 550 Mo sont utilisés pour extraire 535 motifs, tandis que 420 Mo seulement sont nécessaires pour en extraire presque le même nombre avec la méthode a posteriori. Le temps

Seuil s (%)	a posteriori	intégrée
20	546	535
30	437	432
40	361	359
50	137	137
60	63	63
70	3	3
80	0	0

Tableau 3.7 – Nombre de motifs validés pour différents seuils de support minimal.

de calcul est également plus long. Cela est dû à la manipulation et la modification des matrices de concordance ainsi qu'à leur sauvegarde à chaque niveau d'extraction.

La méthode intégrée apparaît donc comme plus coûteuse et légèrement plus sévère que la méthode a posteriori, bien qu'elle évite de générer des contradictions.

3.6 Conclusion

Dans ce chapitre, nous avons considéré le problème des motifs graduels contradictoires, qui pose des difficultés de lisibilité et d'interprétabilité pour l'utilisateur. Nous avons formalisé ces motifs contradictoires, puis nous avons proposé d'ajouter une information supplémentaire qui considère une vision globale et non individuelle sur les motifs graduels. Ce complément d'information conduit à une nouvelle définition de support plus stricte, le *support graduel propre global*, modifiant l'ensemble d'objets considérés comme compatibles, afin de pénaliser de tels motifs. Ce support ne dépend pas uniquement du motif lui-même, mais aussi des motifs contradictoires potentiels.

Nous avons proposé deux approches d'extraction de motifs fréquents selon cette définition de support : la première gère le traitement indépendamment de la génération et la seconde intègre le traitement dans la génération, afin d'éviter de générer des motifs qui s'appuient sur des motifs contradictoires.

Les deux approches proposées extraient des motifs graduels propres, c'est-à-dire des motifs sans ambiguïté : soit des motifs sans contradicteurs, soit des motifs contradictoires pour lesquels il y a suffisamment de données propres pour chacun. En effet, ces motifs s'appliquent à différents sous-ensembles de données qui induisent des intervalles sur lesquels ils sont seuls vérifiés. Ces motifs peuvent donc être considérés comme n'étant pas réellement contradictoires, et extraits en tant que tels.

Le chapitre suivant se focalise sur une telle caractérisation, par l'identification d'intervalles caractéristiques de chaque motif. Toutefois, celle-ci ne vise pas à résoudre l'ambiguïté éventuelle entre motifs contradictoires, mais à augmenter la validité des motifs graduels. Elle est appliquée indépendamment pour chaque motif, alors que le traitement de la contradiction proposé dans ce chapitre adopte un point de vue global, qui traite simultanément des ensembles de motifs contradictoires.

4

Caractérisation de motifs graduels

Sommaire

Introduction

L'objectif de ce chapitre est d'introduire un nouveau type d'enrichissement par une caractérisation des motifs graduels. La caractérisation est exprimée par des clauses introduites par l'expression linguistique « surtout si » : les motifs graduels caractérisés peuvent être illustrés par l'exemple « plus on est proche du mur, plus on freine fort, surtout si la distance au mur est dans $[0, 50]$m », ou plus généralement « M, *surtout si* $J \in R$ », où J est un ensemble d'attributs, appelés *attributs caractéristiques*, appartenant au motif graduel M et R est un ensemble d'intervalles, appelés *intervalles d'intérêt*, définis pour chaque attribut dans J. Nous proposons d'interpréter les motifs graduels caractérisés comme une *validité accrue*, quand les données sont restreintes aux objets satisfaisant la clause de caractérisation.

Cette caractérisation est donc basée sur la restriction d'un ensemble à un sous-ensemble de données, de même que pour le renforcement introduit dans le chapitre 2. Cependant, la restriction est définie ici par des contraintes d'intervalles extraits automatiquement, et non par des modalités existant dans les données. De plus, les motifs graduels caractérisés s'appliquent à des données numériques et non à des données floues.

Cette caractérisation présente également un lien avec le traitement de la contradiction décrit dans le chapitre précédent, mais présente des différences majeures : d'abord, le traitement de la contradiction est un traitement global qui s'applique simultanément à un ensemble de motifs contradictoires, alors que la caractérisation s'applique à un motif individuellement, sans connaissance a priori du reste des motifs. Ensuite, l'objectif ici est d'identifier un intervalle de valeurs sur lequel la validité du motif doit augmenter. Il diffère de l'objectif du traitement de la contradiction qui vise à garantir la non ambiguïté des motifs graduels extraits. Enfin, le traitement de la contradiction garantit un sous-ensemble d'objets propre au motif considéré. Ce sous-ensemble d'objets induit un intervalle de valeurs propre, mais celui-ci n'est pas identifié, en raison de la difficulté d'identification d'attributs caractéristiques dans le cas de motifs graduels de longueur supérieure à 2.

Dans un premier temps, nous présentons dans la section 4.1 les travaux liés à la problématique de la caractérisation, portant sur l'extraction d'intervalles d'intérêt, en particulier dans le cas d'identification de partition floues. Nous formalisons le problème et décrivons ensuite le principe de notre proposition dans la section 4.2. Nous détaillons ensuite les étapes de l'approche proposée dans les sections 4.3 à 4.5. Une étude et une analyse des paramètres de l'approche proposée sont fournies à la section 4.6. Enfin, la section 4.7 présente les expérimentations réalisées en soulignant les points forts et les points faibles de notre approche.

Ces travaux ont été publiés dans (Oudni et al. 2013b; 2013a).

4.1 Travaux liés à l'identification d'intervalles d'intérêt

D'après la formulation indiquée dans l'introduction, la caractérisation proposée peut être formulée comme un problème d'identification d'intervalles d'intérêt. Cette section présente les approches existantes, liées à cette problématique dans des contextes différents.

Ces intervalles d'intérêt représentent des valeurs numériques décrivant des attributs numériques : leur identification est donc liée à la discrétisation des attributs numériques continus, qui consiste à découper le domaine d'un attribut numérique en un nombre fini d'intervalles.

Dans cette section, nous présentons d'abord le principe général de discrétisation d'attributs numériques dans le cadre de l'apprentissage supervisé. Nous détaillons ensuite la méthode de discrétisation floue, proposée par Marsala et Bouchon-Meunier (1996), permettant l'identification de partitions floues. Cette méthode est basée sur les outils de morphologie mathématique dont les principes sont présentés dans la section 4.4.1 rappelant la définition de la morphologie mathématique.

4.1.1 Discrétisation en apprentissage supervisé

Une partie des méthodes d'apprentissage automatique s'applique à des données décrites par des attributs à valeurs discrètes, comme par exemple la construction d'arbres de décision. Leur mise en œuvre pour des données décrites par des attributs numériques nécessite donc de discrétiser ces données, c'est-à-dire de découper leur domaine en un nombre fini d'intervalles chacun identifié par un code. Son objectif est de trouver un compromis entre la qualité des intervalles et leur taille. Les critères de type χ^2 privilégient la garantie d'une taille suffisante des intervalles, tandis que ceux basés sur la mesure de l'entropie privilégient la qualité des intervalles. Parmi ces méthodes de discrétisation, on peut citer par exemple : C4.5 (Quinlan, 1993) qui utilise le critère d'entropie de Shannon, CART (Breiman et al., 1984) qui utilise l'indice de Gini, mesurant l'impureté des intervalles, et CHIMERGE (Kerber, 1992) qui utilise la mesure de χ^2 pour fusionner les intervalles de valeur entre eux, jusqu'à ce qu'un certain seuil de χ^2 soit atteint.

Les modèles à base d'arbres de décision flous utilisent des méthodes de discrétisation floues, qui construisent des partitions floues sur un ensemble de valeurs des attributs numériques.

Parmi celles-ci, on peut citer l'approche proposée par Marsala et Bouchon-Meunier (1996), basée sur l'utilisation d'opérateurs de morphologie mathématique formalisés à l'aide de la théorie des langages formels. L'approche que nous proposons dans ce chapitre est également basée sur ces outils de morphologie mathématique, c'est pourquoi nous détaillons cette approche ci-dessous.

Parmi les méthodes de discrétisation, on distingue classiquement des méthodes descendantes et ascendantes (voir Boullé (2006) pour plus de détails). Les méthodes descendantes partent du domaine numérique complet à discrétiser et le scindent en deux récursivement ; les méthodes ascendantes partent d'intervalles élémentaires mono-valeur et les fusionnent itérativement, jusqu'à ce qu'un certain seuil du critère d'arrêt utilisé soit atteint.

Il faut noter que la différence entre les méthodes de discrétisation présentées dans le chapitre 1, section 1.1.2 et les méthodes de discrétisation en apprentissage supervisé, est que ces dernières intègrent l'information de classe afin de guider le processus pour obtenir une discrétisation discriminante.

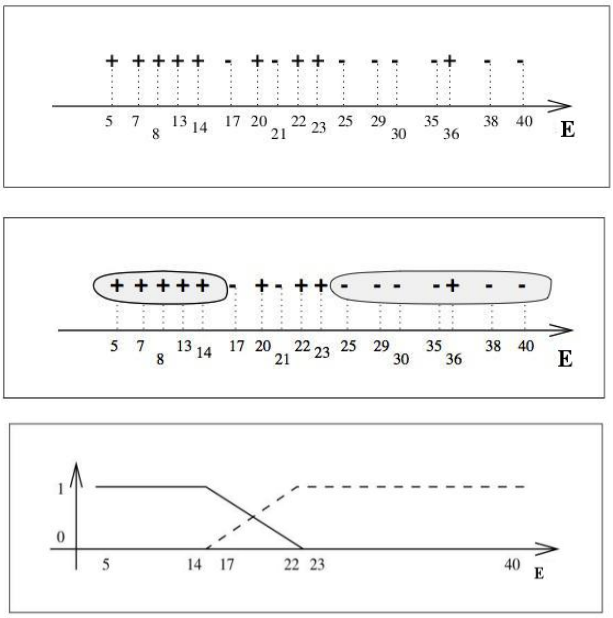


Figure 4.1 – Principe de la méthode de discrétisation floue proposée par Marsala et al. (1995) : en haut, discrétisation des valeurs et des étiquettes associées, au centre, les noyaux identifiés et en bas, la partition floue.

4.1.2 Identification de partitions floues

Dans un cadre de construction d'arbre de décision flou, c'est-à-dire un cadre d'apprentissage supervisé, Marsala et Bouchon-Meunier (1996) proposent une approche permettant la construction d'une partition floue sur un univers de valeurs numériques respectant au mieux la répartition des classes. Cette approche s'appuie sur les outils de la morphologie mathématique dont les principes sont présentés plus en détail dans la section 4.4.1.

Le principe général de cette approche est la création de groupes de valeurs appartenant à une même classe, tout en tolérant certaines valeurs associées à des classes différentes.

À titre d'exemple, on peut considérer un attribut numérique E dont les valeurs décrivant une base de données sont associées aux étiquettes $+$ et $-$ de la façon suivante X_E : $\{(5,+),\ (7,+),\ (8,+),\ (13,+),\ (14,+),\ (17,-),\ (20,+),\ (21,-),\ (22,+),\ (23,+),\ (25,-),\ (29,-),\ (30,-),\ (35,-),\ (36,+),\ (38,-),\ (40,+)\}$, comme illustré sur la figure 4.1.

La première étape de la méthode identifie les noyaux des sous-ensembles flous en appliquant un filtre morphologique (voir section 4.4.1, page 104) pour regrouper les zones homogènes de la distribution des classes. Pour l'exemple de la figure 4.1, elle conduit par exemple aux deux intervalles disjoints $[5;14]$ et $[25;35]$. La partition floue est ensuite déterminée par construction de fonctions d'appartenance trapézoïdales ayant ces noyaux.

Plus précisément, la première étape est basée sur une transformation de la représentation numérique de l'univers continu des valeurs en une représentation symbolique induisant un *mot* (ensemble de caractères) où chaque symbole représente une classe. Un opérateur de filtrage

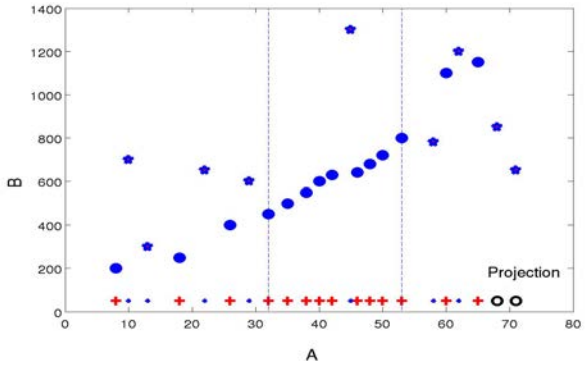


Figure 4.2 – Exemple de caractérisation d'un motif graduel, conduisant à « plus A, plus B, surtout si $A \in [32; 53]$ ».

est alors appliqué pour transformer ce mot en une séquence de séries de symboles homogènes, comme détaillé dans la section 4.4.1. Dans le cadre d'apprentissage supervisé traitant des ensembles de données étiquetés, le filtre permet de lisser l'ensemble d'apprentissage pour déduire une partition floue.

La méthode que nous proposons pour extraire une caractérisation de motifs graduels suit des principes similaires et se place dans le même cadre. La différence principale vient de ce que l'on ne dispose pas de données étiquetées et qu'une transposition au type de données considéré et à l'objectif est nécessaire comme détaillé ci-dessous.

4.2 Formalisation et principe de la méthode proposée

Nous présentons ci-dessous notre contribution fondée sur les outils de morphologie mathématique pour l'extraction automatique de motifs graduels caractérisés. Dans un premier temps, nous présentons le principe et l'interprétation de la caractérisation des motifs graduels en l'illustrant sur un exemple. Ensuite nous présentons la formalisation proposée.

4.2.1 Motivations

La figure 4.2 représente un ensemble de données décrit par deux attributs pour lesquels le motif graduel $I = A^{\geq}B^{\geq}$ est supporté par le chemin représenté par •. Son support graduel est $14/23 = 60\%$. Or, on peut observer que la co-variation entre A et B a lieu particulièrement dans la partie centrale du graphique, tandis que les données qui ne sont pas en accord avec le motif se trouvent surtout dans les parties où A prend des valeurs basses ou élevées. Plus précisément, si les données sont limitées aux objets pour lesquels A prend des valeurs dans l'intervalle $[32; 53]$, graphiquement délimité par les lignes verticales sur la figure 4.2, alors le support du motif augmente à $9/10 = 90\%$. Ceci motive l'extraction du motif graduel caractérisé $A^{\geq}B^{\geq}$ *; surtout si* $A \in [32; 53]$.

Plus généralement, nous proposons d'interpréter la caractérisation des motifs graduels comme validité accrue, quand les données sont restreintes aux objets satisfaisant la clause de caractérisation. Cependant, pour qu'elle soit informative, une telle caractérisation ne doit pas limiter les données drastiquement : il est facile d'atteindre 100% de support, par exemple en restreignant les données à un unique couple de points satisfaisant l'ordre induit par le motif graduel considéré. Pourtant, la caractérisation résultante serait trop spécifique et non pertinente. Selon le même principe, dans l'exemple précédent, restreindre les données à l'intervalle plus petit $[32; 42]$ augmente le support à 100%, mais conduit à une caractérisation trop spécifique.

Le principe des motifs graduels caractérisés est donc de trouver un compromis entre un support élevé et un nombre élevé d'objets lors de la restriction des données à un sous-ensemble de données définie par les intervalles considérés.

4.2.2 Formalisation

Définition 4.1 (Motifs graduels caractérisés)**.** Un motif graduel caractérisé est défini par l'expression linguistique « I *, surtout si* $J \in R$ », où I est un motif graduel, J un ensemble d'attributs qui apparaissent dans I et R un ensemble d'intervalles.

Il faut souligner que, dans le cas d'un intervalle V associé à un attribut A_V très étroit, il est pertinent de remplacer l'expression « surtout si A_V est dans $[\min(V), \max(V)]$ » par « surtout si A_V égal à val », où val est la valeur centrale de l'intervalle. La définition du seuil définissant si un intervalle est étroit ou non peut être donnée par l'utilisateur, pour lui permettre de faire des résumés adaptés à ses besoins et préférences.

L'ensemble d'intervalles R définit une région qui induit une restriction $\mathcal{D}'$ de l'ensemble de données $\mathcal{D}$, en considérant uniquement les données satisfaisant la contrainte de valeur exprimée par R.

Le principe exposé dans la section précédente consiste alors à maximiser à la fois le support du motif considéré I sur les données restreintes $\mathcal{D}'$, et le nombre d'objets satisfaisant les contraintes d'ordre sur $\mathcal{D}'$, c'est-à-dire

$$\max_R |\mathcal{D}'| \tag{4.1}$$

$$\max_R SG_{\mathcal{D}'}(I) \tag{4.2}$$

où SG représente le support graduel rappelé dans l'équation (1.7), page 40.

Un compromis doit être trouvé entre ces deux objectifs qui peuvent être contradictoires : en effet, une augmentation de la taille du sous-ensemble $\mathcal{D}'$ peut conduire à la diminution de la proportion d'objets compatibles avec l'ordre induit par le motif considéré. Ce phénomène peut être illustré avec l'exemple précédent : l'intervalle $[32; 53]$ satisfait l'objectif de l'équation (4.1), mais il ne satisfait pas l'objectif de l'équation (4.2). Au contraire, l'intervalle $[32; 42]$ satisfait l'objectif de l'équation (4.2) et non celui de l'équation (4.1).

Définition 4.2 (Motif graduel caractérisé valide)**.** Un motif graduel caractérisé I est dit *valide* si est seulement si le motif graduel sous-jacent est valide et le résultat de l'équation (4.2) est supérieur à un seuil s_c, fixé par l'utilisateur.

4.2.3 Principe général

Afin d'identifier les motifs graduels caractérisés valides, une approche naïve et trop coûteuse pourrait consister à réaliser une discrétisation préalable. Selon cette approche, on pouvait introduire des modalités binaires pour chaque clause de caractérisation possible (valant 1 si la valeur appartient à l'intervalle candidat, 0 sinon) sur lesquelles on pourrait appliquer l'approche du renforcement. Cependant, cela conduit à un grand nombre de possibilités et ne permet donc pas d'identifier les modalités pertinentes de manière efficace.

Aussi, une méthode intégrée recherchant directement des intervalles associés est proposée. Celle-ci est basée sur l'utilisation d'outils morphologiques mathématiques détaillés dans la section 4.4.1 et décompose la tâche d'identification des attributs caractéristiques et de leurs intervalles d'intérêt associés, en considérant successivement chaque attribut composant le motif graduel considéré I ainsi que chaque chemin supportant I : le calcul du support graduel restreint $SG_{\mathcal{D}'}(I)$ et l'intervalle d'intérêt sont basés sur la restriction des chemins associés à I. Pour cela, nous identifions d'abord, pour chaque chemin, la restriction candidate : nous proposons de passer de la représentation numérique à une représentation symbolique où chaque objet appartenant au chemin considéré est représenté par des + et où les objets restant sont représentés avec l'un des deux symboles $\{-, \circ\}$. Cette phase correspond à la phase de transcription des données, décrite dans la section 4.3. La représentation symbolique est ensuite traitée par un filtrage morphologique, détaillé dans la de section 4.4. La restriction que nous souhaitons identifier correspond à la plus grande séquence de + induite par le processus de filtrage morphologique. Les opérateurs de morphologie mathématique reflètent de manière pertinente la recherche des meilleures restrictions et conviennent pour assurer le compromis entre les deux équations (4.1) et (4.2). Une fois qu'une restriction est identifiée pour chaque chemin, un post-traitement est effectué sur les différents chemins : les restrictions sont combinées pour sélectionner les limites optimales qui correspondent aux limites de la plus grande restriction identifiée. Cette phase est détaillée dans la section 4.5.

4.3 Représentation symbolique des données : transcription

Cette section présente le processus de passage de la représentation numérique des données à une représentation symbolique, en détaillant les règles de transcription. Elle illustre ensuite le calcul de support graduel à partir de cette nouvelle représentation. Enfin, la dernière partie de cette section prend en compte l'information de la densité des données et présente de nouvelles règles de transcription permettant de considérer cette information.

4.3.1 Règles de transcription

Définition 4.3 (Règles de transcription)**.** La transcription des données de $\mathcal{D}$ pour un motif graduel I, un chemin D et un attribut A pour lequel un intervalle d'intérêt est recherché, est définie par le mot composé des symboles $\{+, -, \circ\}$ tel que le $i$ème caractère est obtenu, selon les règles suivantes :

— $o \rightarrow +$ ssi $o \in D$

— $o \rightarrow -$ ssi $(o \notin D) \wedge (A_{mD} \leq A(o) \leq A_{MD})$

— $o \rightarrow \circ$ sinon

où A_{mD} et A_{MD} représentent respectivement les valeurs minimale et maximale de l'attribut A observées pour les objets dans D : $A_{mD} = \min_{o \in D} A(o)$ et $A_{MD} = \max_{o \in D} A(o)$.

Le symbole $\circ$ code les données en dehors des limites du chemin traité ; il est nécessaire pour traiter le cas de plusieurs chemins maximaux, comme détaillé dans la section 4.5.1.

Les données de la figure 4.2 conduisent par exemple au mot v représenté sur la partie inférieure de la figure et redonné ci-dessous

$$v = +\text{-}-+\text{-}+\text{-}+++++\text{-}+++++\text{-}+\text{-}+\circ\circ$$

4.3.2 Calcul du support graduel à partir de la représentation symbolique

L'objectif formalisé dans les équations (4.1) et (4.2) peut alors être transposé à la représentation d'un chemin sous la forme d'un mot : la restriction de l'ensemble de données correspond à une sous-partie du mot, et $|\mathcal{D}'|$ à sa longueur. Le support restreint $SG_{\mathcal{D}'}(I)$ est défini par le nombre d'objets compatibles, qui sont exactement les objets du chemin ayant été transcrits comme + . On peut donc définir $SG_{\mathcal{D}'}(I)$ comme le nombre de +, normalisé par le nombre total d'éléments contenus dans cette sous-partie. Dans ce qui suit, pour un mot v, on note $l(v)$ sa longueur et $NP(v)$ le nombre de + qu'il contient.

Définition 4.4 (Expression symbolique du support graduel)**.** Le support d'un motif graduel est étendue à un mot v comme :

$$SG(v) = \frac{NP(v)}{l(v)} \tag{4.3}$$

Le support le plus élevé est obtenu lorsque la sous-partie considérée du mot est une séquence de + qui ne contient pas de symbole −, conduisant à $SG_{D'} = 1$. La plus longue séquence de + identifiée dans v, notée $S(v)$, a pour taille $l(S(v))$ et pour support $SG(S(v)) = 1$.

La question est alors d'étendre la taille d'une telle séquence, $S(v)$, en tolérant quelques symboles −, de manière à augmenter la taille de l'ensemble de données restreint, sans trop dégrader la proportion des + dans la sous-partie considérée. On peut, par exemple, avoir

dans v, deux séquences de $+$, s_1 et s_2, plus courtes que $S(v)$, et séparées seulement par une courte séquence de $-$, notée s_-. Dans ce cas, la sous-partie du mot composée de la concaténation $s' = s_1 s_- s_2$ conduit à une longue séquence avec un nombre de $+$ qui reste élevé. Plus précisément, $l(s') = l(s_1) + l(s_-) + l(s_2)$ et $SG(s') = (l(s_1) + l(s_2))/l(s')$.

Le compromis entre la taille et le support équivaut à se demander si l'on préfère considérer le sous-ensemble de données correspondant à $S(v)$, qui maximise le support, ou plutôt celui induit par s', qui a une plus grande longueur au détriment d'un support inférieur. Pour cela, nous proposons d'exploiter les outils de morphologie mathématique que nous rappelons dans la section suivante. Le support est calculé à partir de la séquence finale qui sera identifiée par ces outils. Cette séquence appelée *séquence caractéristique* est définie ci-dessous.

Définition 4.5 (Séquence caractéristique)**.** Pour un mot v' résultant de l'application des outils de morphologie mathématique, une séquence caractéristique est définie comme par un sous-mot de v' représentant la plus longue séquence de $+$.

La séquence caractéristique est donc la symbolique des objets appartenant à la restriction de l'ensemble de données. L'intervalle d'intérêt est défini par ses limites qui sont représentées par les limites de la séquence caractéristique : il s'agit des valeurs minimale et maximale de l'attribut pris en compte.

4.3.3 Prise en compte de la densité

Le principe général de la caractérisation illustrée dans la section précédente ne tient pas compte de la densité des données. Cette section illustre la pertinence de la prise en compte de cette information pour la caractérisation de façon générale. Cet objectif est traduit par des règles de transcription modifiant légèrement celles introduites dans la définition 4.3.

Motivations

Dans le cas général, il peut arriver que deux sous-ensembles de données diffèrent par leur densité mais qu'ils soient de même cardinalité et donnent le même intervalle caractéristique, comme illustré sur la figure 4.3. Les deux cas représentent un sous-ensemble d'une base de données pour lequel le support est de 100% et qui conduit à l'intervalle caractéristique $[8, 42]$. Néanmoins, pour le cas de droite, il semble plus satisfaisant de restreindre encore l'intervalle, pour définir la clause *surtout si* $A \in [26; 29]$: le fait d'ignorer les deux premiers objets et le dernier, isolés du reste des objets, permet d'identifier une zone dense qui est en effet plus caractéristique du motif.

La densité est mesurée par le nombre d'objets rapporté à la taille de l'intervalle, et permet de différencier les deux intervalles caractéristiques.

En appliquant ce principe à l'exemple de la figure 4.2, on considère l'intervalle $[32; 53]$ plutôt que $[32; 65]$: les trois objets représentés à droite du chemin considéré ne sont pas pris en compte, parce qu'ils sont isolés du reste des objets de l'intervalle $[32; 65]$. Le support de cette nouvelle restriction est plus élevé que celui de la restriction précédente, qui est de 90%.

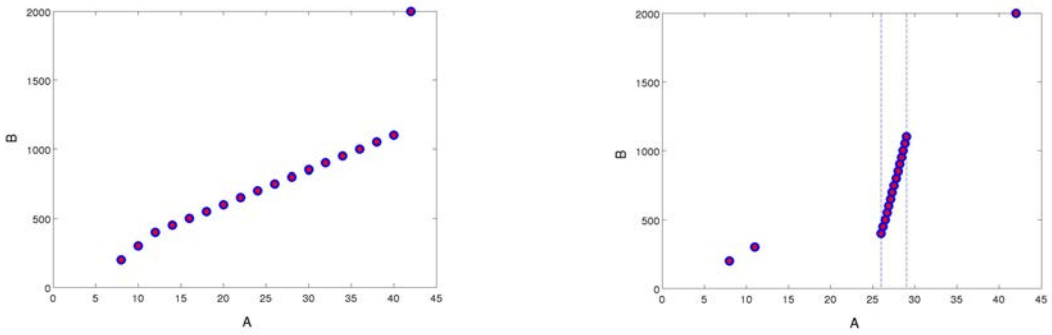


Figure 4.3 – Deux sous-ensembles de données de même cardinalité mais de densité différentes, donnant le même intervalle caractéristique.

En revanche, cette restriction contient moins d'objets que la précédente, ce qui signifie que le compromis effectué et souhaité est différent du compromis précédent.

Nous proposons d'intégrer l'information de densité dans l'extraction d'une séquence caractéristique : nous voulons des séquences denses, c'est-à-dire que nous ne souhaitons pas intégrer les + isolés dans une séquence caractéristique. La définition d'un symbole « isolé » est liée à la taille de l'écart qui le sépare du symbole voisin.

L'objectif consiste alors à trouver un compromis qui maximise à la fois le support du motif considéré I sur $\mathcal{D}'$, le nombre d'objets dans $\mathcal{D}'$ et la densité des données dans R.

Insertion et transcription des objets fictifs

Afin de prendre en compte la densité, nous proposons de générer des objets fictifs, insérés entre les objets de la base initiale, afin de garantir que l'écart entre deux valeurs successives observées pour l'attribut A soit inférieur ou égal à e, où e est un écart minimum fixé par l'utilisateur, appelé *écart de base*. Les règles permettant de passer à la représentation symbolique sont similaires à celles décrites dans la section 4.3.1, mais contiennent une règle supplémentaire qui permet de représenter les objets fictifs insérés.

Définition 4.6 (Règles de transcription avec présence d'objets fictifs)**.** Pour une base de données $\mathcal{D}'$ composée de données de $\mathcal{D}$ et des objets fictifs insérés, un motif graduel I, un chemin D et un attribut A pour lequel un intervalle d'intérêt est recherché, la transcription de $\mathcal{D}'$ est définie par le mot composé des symboles $\{+, -, \circ\}$ tel que le $i$ème caractère est obtenu, selon les règles suivantes :

- $o \rightarrow -$ **si** o **est un objet fictif**
- $o \rightarrow +$ si $o \in D$
- $o \rightarrow -$si $(o \notin D) \wedge (A_{mD} \leq A(o) \leq A_{MD})$
- $o \rightarrow \circ$ sinon

où A_{mD} et A_{MD} représentent respectivement les valeurs minimale et maximale de A sur D.

Ainsi, pour les données situées à droite de la figure 4.3, si l'on considère que l'écart minimal entre deux objets successifs est de 2, alors 2 objets fictifs sont insérés entre le premier et le deuxième objet, 8 objets fictifs sont insérés entre le deuxième et le troisième objet de gauche à droite et 6 objets sont insérés entre le dernier et l'avant dernier objet. En appliquant la nouvelle transcription, on obtient le mot représenté ci-dessous.

$$v = +\text{--}+\text{--------}+++++++++++++++++\text{------}+$$

Calcul du support d'un mot

La nouvelle définition du support du mot doit prendre en compte les objectifs fictifs : rappelons que la restriction de l'ensemble de données correspond à une sous-partie du mot, et $|\mathcal{D}'|$ à sa longueur. Le support restreint d'un motif graduel (I), $SG_{\mathcal{D}'}(I)$, est défini par le nombre d'objets compatibles, qui sont les objets du chemin ayant été transcrits comme $+$. On peut donc définir le $SG_{\mathcal{D}'}(I)$ comme le nombre de $+$, normalisé par le nombre total d'éléments dans cette sous-partie moins le nombre d'objets fictifs insérés.

Définition 4.7 (Support d'un mot)**.** En notant $NF(v)$ le nombre d'objets fictifs contenus dans le mot v, le support d'un motif graduel est étendu à un mot comme :

$$SG(v) = \frac{NP(v)}{(l(v) - NF(v))} \tag{4.4}$$

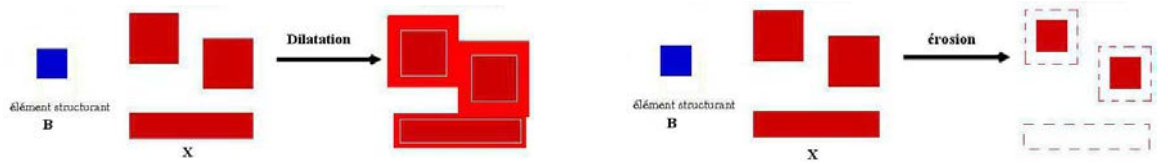
On peut noter que la formule donnée dans l'équation (4.3) en constitue bien un cas particulier quand aucun objet fictif n'est ajouté.

4.4 Filtrage morphologique

L'objectif formulé dans la section 4.2.2 peut être atteint en appliquant une méthode qui consiste à filtrer les mots obtenus à l'issue de la transcription, de telle sorte que les symboles $-$ isolés n'empêchent pas de construire de grands ensembles de données restreints. En effet, il est alors possible d'augmenter la taille de la séquence considérée, avec une diminution limitée de la proportion de $+$. De tels effets de filtrage peuvent être obtenus par l'application des opérateurs de morphologie mathématique appropriés. Le principe consiste à appliquer un opérateur φ sur un mot v, conduisant à $v' = \varphi(v)$ afin de combler l'écart entre les séquences de $+$ dans v, tout en identifiant la plus longue séquence de $+$ dans v', $S(v')$, et en évaluant la séquence correspondante dans v, $S_{v'}(v)$, avec une longueur $l(S(v'))$ et un support

$$SG(v') = \frac{NP(S_{v'}(v))}{l(S_{v'}(v))}$$

Cette section a pour but de définir l'opérateur de filtrage appliqué au résultat du processus de transcription pour répondre à cet objectif. Tout d'abord, nous présentons un rappel de morphologie mathématique. Nous décrivons ensuite l'opérateur proposé, ainsi que l'analyse de ses propriétés et de sa pertinence.



(a) Dilatation (b) Érosion

Figure 4.4 – Opérateurs de dilatation et d'érosion

4.4.1 Rappels de morphologie mathématique

La morphologie mathématique (Coster & Chermant, 1985; Serra, 1988), notée MM dans ce qui suit, a été largement utilisée pour le traitement d'images et l'analyse fonctionnelle. Son idée première est de comparer une forme à traiter à une autre forme géométrique fixée appelée *élément structurant*. Au-delà du traitement d'images, la morphologie mathématique a apporté une contribution importante dans des domaines variés, tels que, par exemple, la programmation de jeu de go (Bouzy, 1995) pour isoler les blocs d'éléments similaires dans un univers, ou encore pour la construction de partitions floues (Marsala & Bouchon-Meunier, 1996).

Les opérateurs utilisés ici sont des transpositions des opérateurs classiques au cas unidimensionnel, et s'appliquent à des mots obtenus à l'issue d'une transcription de l'univers numérique.

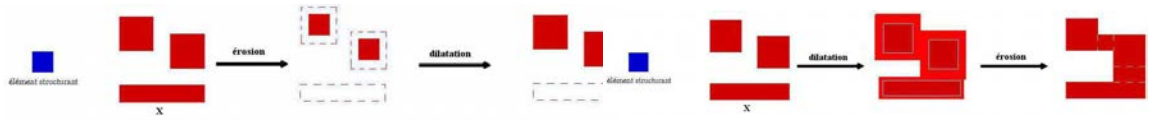
Opérateurs de base : dilatation et érosion

Les transformations morphologiques de base sont la dilatation et l'érosion. La dilatation est définie comme l'union avec l'élément structurant que l'on fait glisser sur l'image. Sous l'effet de la dilatation, tous les objets « grossissent » d'une partie correspondant à la taille de l'élément structurant. S'il existe des trous dans les objets, c'est-à-dire, dans le cas des images, des « morceaux » de fond à l'intérieur des objets, ils sont comblés et si des objets sont situés à une distance moins grande que la taille de l'élément structurant, ils fusionnent. L'effet de cet opérateur est illustré sur la figure 4.4(a) : la dilatation permet la fusion des formes proches.

Définition 4.8 (Dilatation)**.** Pour un espace E, un élément structurant B et une forme X, l'opérateur de dilatation est défini comme : $Di_B = \{x \in E, B(x) \cap X \neq \emptyset\}$.

L'érosion est l'opération duale de la dilatation : sous l'effet de l'érosion, les objets de taille inférieure à celle de l'élément structurant disparaissent, les autres sont « amputés » d'une partie correspondant à la taille de l'élément structurant. S'il existe des « trous » dans les objets, c'est-à-dire des « morceaux » de fond à l'intérieur des objets, ils sont accentués, et les objets reliés entre eux peuvent être séparés. Cette effet est illustré sur la figure figure 4.4(b) : l'érosion permet la suppression des petites formes.

Définition 4.9 (Érosion)**.** Pour un espace E, un élément structurant B et une forme X, l'opérateur d'érosion est défini comme : $Er_B(X) = \{x \in E, B(x) \subseteq X\}$.



(a) Opérateur d'ouverture (b) Opérateur de fermeture

Figure 4.5 – Opérateurs d'ouverture et de fermeture

Opérateurs d'ouverture et de fermeture

Ces opérateurs de base sont ensuite combinés pour définir des opérateurs plus complexes. Une ouverture est la composition d'une érosion suivie d'une dilatation. Elle permet la destruction des petites formes, relativement à la taille de l'élément structurant. Une ouverture permet d'adoucir les contours, coupe les isthmes étroits, supprime les petites îles et adoucit les caps étroits (Coster & Chermant, 1985). Cette opération est illustrée sur la figure 4.5(a) : la forme rectangulaire est supprimée par l'effet de cette opération.

Définition 4.10 (Ouverture)**.** Une ouverture est définie comme :

$$Ouv_B = Di_B \circ Er_B$$

Une fermeture est la composition d'une dilatation suivie d'une érosion. Elle permet de fusionner les formes proches dans l'espace. L'effet d'une fermeture est de boucher les canaux étroits, supprimer les petits lacs et les golfes étroits (Coster & Chermant, 1985). Cette opération est illustrée sur la figure 4.5(b) : les trois formes de l'espace sont fusionnées par l'effet de cette opération.

Définition 4.11 (Fermeture)**.** Une fermeture est définie comme :

$$Fer_B = Er_B \circ Di_B$$

Le filtre alterné

On note n un entier désignant le nombre de fois où un opérateur morphologique est appliqué. Un *filtre alterné* d'ordre n est composé de n ouvertures successives suivies de n fermetures successives, appliquées à une forme de l'espace, avec le même élément structurant B. Il permet la destruction des petites formes, en fonction de la taille de B, tout en fusionnant les formes proches. Une grande valeur de n ne laisse que les formes de taille suffisante et élimine les aspérités et les vides de taille importante. Le résultat de l'application du filtre d'ordre 1 sur la forme X illustrée avec les opérateurs précédents est donné sur la figure 4.6.

Définition 4.12 (Filtre alterné d'ordre n)**.** Un filtre alterné d'ordre n est défini par :

$$\begin{array}{ll} n = 1 & Filt_1 = Fer_1 \circ Ouv_1 \\ n > 1 & Filt_n = Fer_n \circ Ouv_n \circ Filt_{n-1} \end{array}$$

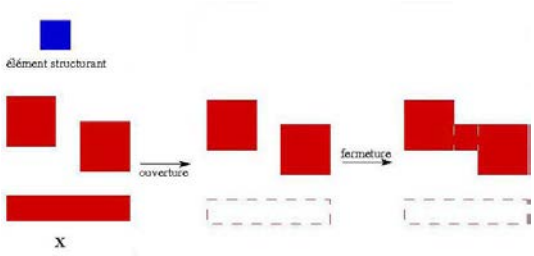


Figure 4.6 – Filtre alterné d'ordre 1

La morphologie mathématique unidimensionnelle, 1DMM

Les outils de la morphologie mathématique, rappelés ci-dessus, sont généralement appliqués en traitement d'images sur des espaces à deux dimensions. Marsala et Bouchon-Meunier (1996) ont considéré un univers unidimensionnel et ont proposé des outils 1DMM pour obtenir des effets de filtrage permettant de construire des partitions floues dans le cadre de l'apprentissage supervisé, comme indiqué dans la section 4.1.2. Ces outils s'appliquent à un mot défini sur un vocabulaire binaire, noté $\{+,-\}$, où chaque symbole est associé à une classe et où il existe une symétrie entre les deux symboles. Les auteurs ont de plus introduit un symbole particulier, noté u, pour marquer les zones du mot modifiées lors de l'application des opérateurs. Elles sont interprétées ensuite comme des séquences incertaines, c'est-à-dire des séquences où, après le processus de filtrage d'un mot, les classes sont très mélangées. Ces opérateurs transforment donc un mot défini sur le vocabulaire binaire, $\{+,-\}$, en un mot défini sur un vocabulaire ternaire $\{+,-,u\}$. Un opérateur de filtrage s'appliquant à un mot est également défini par les deux opérateurs d'ouverture et de fermeture, eux-mêmes construits à l'aide d'opérateurs d'érosion et de dilatation.

Lorsque le filtre d'ordre n est appliqué à un mot en utilisant un élément structurant de taille 1 (un + ou un −), les petites séquences de moins de $2n$ symboles (c'est-à-dire inférieur strictement à $2n$, sont transformées en séquences incertaines (représentées par des u), et les séquences composées de mêmes symboles séparées par moins de $2n$ symboles différents de ceux qu'elles contiennent, sont regroupées, et ainsi nommées séquences certaines. Ces deux types de séquences sont utilisés pour construire la partition floue associée à l'attribut en question. Les séquences certaines correspondent aux noyaux des sous-ensembles flous de la partition.

Nous illustrons ces principes sur l'exemple de la base d'apprentissage X_E présenté dans la section 4.1.2. L'opérateur de transcription remplace chaque valeur observée par son étiquette, + ou − et conduit au mot

$$v = + + + + + - + - + + - - - - + - -$$

En appliquant un filtre d'ordre 1 sur le mot $v = + + + + + - + - + + - - - - + - -$, on obtient le mot

$$v' = u + + + uuuuuuu - - uuuu$$

Ainsi, deux séquences incertaines apparaissent dans le mot filtré, ainsi qu'une séquence de + et une séquence de −. Ces deux dernières séquences sont utilisées comme noyaux des sous-ensembles flous de la partition floue.

4.4.2 Opérateurs proposés

Dans le cas de la caractérisation des motifs graduels, les mots modifiés sont définis sur le même ensemble de symboles que celui des mots initiaux. Ainsi, contrairement à l'approche de Marsala et Bouchon-Meunier (1996), nous n'introduisons pas de nouveau symbole qui conserve la trace des positions du mot où l'opérateur conduit à des modifications. D'autre part, cet ensemble est ternaire, dès le début, de par le symbole ∘ qui encode les données en dehors des limites du chemin traité. Le symbole ∘ n'est pas modifié par un opérateur considéré. Une autre spécificité de la caractérisation est l'absence de symétrie entre le symbole + et le symbole − : l'effet est entièrement focalisé sur les séquences de +, alors que dans le cas des partitions floues, les séquences de − jouent des rôles équivalents aux séquences de +.

Cette section décrit les opérateurs proposés pour effectuer le filtrage souhaité, qui constituent des transpositions des opérateurs classiques définis par la MM pour le cas unidimensionnel.

L'opérateur d'érosion, noté Er_1, diminue la taille des séquences + en remplaçant les + externes par des −.

Définition 4.13 (Érosion)**.** Pour tout $m \geq 0$, l'opérateur d'érosion s'applique en suivant les règles suivantes :

$$\begin{array}{cccccccccc} - & +^{m+2} & - & \longrightarrow & - & - & +^{m} & - & - \\ \circ & +^{m+1} & - & \longrightarrow & & \circ & +^{m} & - & - \\ - & +^{m+1} & \circ & \longrightarrow & - & - & +^{m} & \circ & \end{array}$$

avec le cas limite : $- \quad + \quad - \quad \longrightarrow \quad - \quad - \quad - \;.$

Les deux dernières lignes explicitent la spécificité du symbole ∘.

Ainsi, pour $v =$ - + + + - - - +
$Er_1(v) =$ - - + - − − − -

Er_n désigne la combinaison de n érosions successives. On peut observer que l'application de Er_n transforme toutes les séquences de + de longueur inférieure à $2n$, dont chaque élément est progressivement remplacé par −.

Réciproquement, l'opérateur de dilatation, notée Di_1, diminue les séquences de − et développe les séquences de +.

Définition 4.14 (Dilatation)**.** Pour tout $m \geq 0$, l'opérateur de dilatation s'applique ensuivant les règles suivantes :

$$\begin{array}{cccccccccc} + & -^{m+2} & + & \longrightarrow & + & + & +^{m} & + & + \\ \circ & -^{m+1} & + & \longrightarrow & & \circ & -^{m} & + & + \\ + & -^{m+1} & \circ & \longrightarrow & + & + & -^{m} & \circ & \end{array}$$

avec le cas limite : $+ \quad - \quad + \quad \longrightarrow \quad + \quad + \quad +$.

Par exemple, pour le mot précédent v, $Di_1(v)$ produit $Di_1(v) = +++++-++$.

Di_n est la combinaison de n dilatations successives. L'application de Di_n transforme toutes les séquences de $-$ de longueur inférieure à $2n$ en séquences de $+$.

Définition 4.15 (Ouverture)**.** L'opérateur d'ouverture est défini de même qu'en morphologie mathématique classique, c'est-à-dire

$$Ouv_n = Di_n \circ Er_n$$

Par exemple, avec le mot

$$\begin{aligned} v &= \texttt{--++++++-++++-+} \\ \text{on a } Ouv_1(v) &= \texttt{--++++++-++++--} \\ \text{et } Ouv_2(v) &= \texttt{--++++++-------} \end{aligned}$$

L'effet de l'érosion est d'élargir les séquences de $-$. La dilatation permet de les réduire à nouveau, sauf dans les régions où l'étape d'érosion a supprimé tous les symboles $+$, puisqu'il n'y a plus de symbole $+$ à propager. Cela signifie que, par rapport au mot initial, l'opérateur d'ouverture comble les vides entre les séquences de $-$ séparées par moins de $2n$ symboles $+$.

Définition 4.16 (Fermeture)**.** Réciproquement, l'opérateur de fermeture est défini comme

$$Fer_n = Er_n \circ Di_n$$

Par exemple, à partir du mot utilisé dans l'exemple précédent,

$$\begin{aligned} \text{on a} \quad Fer_1(v) &= \texttt{--++++++++++-} \\ \text{et} \quad Fer_2(v) &= \texttt{--++++++++++--} \end{aligned}$$

L'opérateur de fermeture comble le vide entre les séquences de $+$ séparées par moins de $2n$ symboles $-$.

Définition 4.17 (Filtre alterné)**.** Le filtre alterné est la combinaison récursive des opérations d'ouverture et de fermeture :

$$\begin{aligned} n = 1 \quad & Filt_1 = Fer_1 \circ Ouv_1 \\ n > 1 \quad & Filt_n = Fer_n \circ Ouv_n \circ Filt_{n-1} \end{aligned}$$

Pour tout n donné, l'étape d'ouverture, Ouv_n, supprime d'abord les séquences de $+$ courtes, de longueur inférieure à $2n$, en comblant les vides entre les séquences de $-$. Les séquences de $+$ restantes, qui sont donc de longueur supérieure à $2n + 1$, peuvent être regroupées par l'opérateur de fermeture si elles sont séparées par moins de $2n$ symboles $-$.

De plus, le filtre alterné étant défini de manière récursive, ce comportement est appliqué à la suite des filtres précédents, $Filt_{n-1} \circ Filt_{n-2} \circ \cdots \circ Filt_1$.

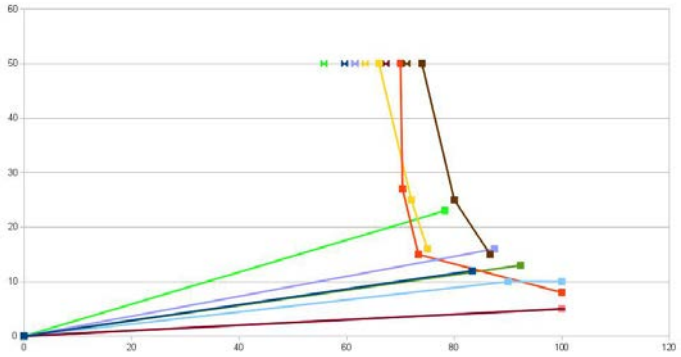


Figure 4.7 – Support (abscisse) et longueur (ordonnée) des séquences extraites à partir de 10 mots aléatoires chacun représenté par une couleur et par des filtres alternés de $Filt_1$ à $Filt_{10}$: les points notés ■ représentent les résultats du filtre et les points notés ⋈ représentent le support des mots initiaux.

4.4.3 Propriétés du filtre

Dans cette section, nous discutons des propriétés des opérateurs proposés, et plus particulièrement de celles qui nous permettent d'examiner les caractéristiques du sous-mot extrait à partir du mot initial. Dans un premier temps, nous présentons les propriétés du comportement de compromis et d'asymétrie du filtre, puis nous illustrons ces propriétés sur des exemples.

Comportement de compromis

Il faut d'abord souligner que lorsque n augmente, le filtre alterné est à la fois plus tolérant et plus exigeant : il permet en effet de remplacer des séquences de $-$ plus longues dans des séquences de $+$. Il est donc plus tolérant aux séquences de $-$ dans les séquences de $+$. Néanmoins, pour effectuer de telles modifications, le filtre alterné prend uniquement des séquences de $-$ entourées par des séquences de $+$ longues : il est donc plus exigeant pour fusionner les séquences de $+$. C'est la raison pour laquelle il met en œuvre un compromis entre la longueur et la proportion de symboles $+$, fournissant ainsi un outil intéressant pour extraire les intervalles d'intérêt.

Afin d'illustrer cette propriété, nous avons généré de manière aléatoire des mots de taille fixe, égale à 50. Chaque mot est représenté par une couleur sur la figure 4.7. Les points représentés avec les symboles ■ décrivent les résultats du filtre de taille n allant de 1 à 10. Les points représentés avec les symboles ⋈ indiquent le support des mots initiaux.

Le filtre est appliqué sur ces mots dans le but d'observer pas à pas le comportement du filtre. La figure 4.7 montre le support et la longueur des séquences extraites à partir des 10 mots considérés. Deux types de mots peuvent être observés. Certains sont transformés en mots sans séquence de $+$, ce qui conduit à un support nul et une longueur nulle. Dans ces mots, les séquences de $+$ sont trop courtes et un filtre d'ordre faible doit être appliqué si on souhaite les garder et parfois même en utilisant un ordre de filtre très faible, par exemple

égale à 1, ne suffit pas. Un second type de mots montre un compromis entre taille et support : pour de faibles valeurs d'ordre du filtre, les séquences de + sont courtes et pures (c'est-à-dire un support de 100%), alors que pour des ordres du filtre élevés, plus de symboles − sont effacés, ce qui conduit à des séquences plus longues mais de support inférieur.

On peut également souligner que, généralement, après application des filtres alternés $Filt_n$, soit aucun symbole + n'apparaît dans le mot obtenu, soit la séquence de + contient au moins $2n + 1$ symboles +. Cela confère une garantie sur la taille des séquences extraites lors de l'application d'un filtre alterné.

Propriété d'asymétrie

D'après les observations exposées dans le paragraphe précédent, on peut aussi préciser que la combinaison d'ouverture et de fermeture conduit à une asymétrie intéressante du filtre $Filt_n$. En effet, après l'application de $Filt_n$,

— les séquences de + d'une longueur inférieure à $2n$ sont remplacées par des séquences de − ;

— les séquences de − d'une longueur inférieure à $2n$ et entourées de séquences de + de longueur supérieure à $2n + 1$ sont remplacées par des séquences de +.

Dans les commentaires ci-dessous, les définitions de « courte » et « longue » sont données par rapport à la valeur du filtre $2n$.

Cette propriété d'asymétrie du filtre montre que les séquences de + courtes sont inconditionnellement remplacées par des séquences de −, alors que le remplacement de courtes séquences de − impose des conditions sur la longueur des séquences de + qui les entourent.

Cette propriété est très pertinente dans le contexte de la caractérisation de motifs graduels. Elle se concentre sur les symboles +, et est liée à l'obligation de ne pas dégrader la valeur du support défini dans l'équation (4.3), lorsque les séquences de + sont fusionnées en ajoutant quelques symboles −. En revanche, un opérateur de fermeture fusionne les séquences de + indépendamment de leur longueur, ce qui peut conduire à de longues séquences ayant un support faible. Le filtre alterné permet de fusionner les séquences de − uniquement si elles sont entourées de séquences de + plus longues.

Illustrations

Nous considérons tout d'abord l'exemple du pire cas pour l'opération de $Fer_n \circ Ouv_n$, c'est-à-dire le cas où la séquence obtenue a le support le plus faible : il correspond à la séquence obtenue à partir d'un mot qui contient le maximum de symboles −. Rappelons que, pour fusionner deux séquences de +, celles-ci doivent être séparées par au maximum $2n$ symboles − et doivent elles-mêmes contenir au minimum $2n + 1$ symboles +. Le pire cas contient alors le maximum de −, $2n$, et le minimum de +, $2n + 1$, de chaque coté de la séquence de −. Ce mot correspond à $v = +^{2n+1} -^{2n} +^{2n+1}$. La combinaison $Fer_n \circ Ouv_n$ transforme en une séquence de + le mot initial $v = +^{2n+1} -^{2n} +^{2n+1}$. $S_{v'}(v)$ est de longueur $6n + 2$ et a un support de $(4n + 2)/(6n + 2)$. Ce dernier est donc toujours supérieur à 0.66.

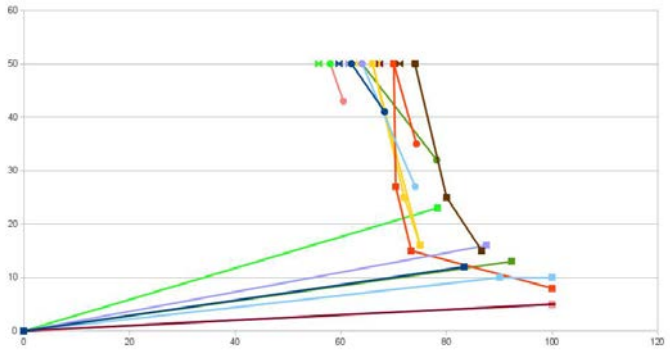


Figure 4.8 – Support (abscisse) et longueur (ordonnée) des séquences extraites à partir de 10 mots aléatoires chacun représenté par une couleur, par des filtres alternés de $Filt_1$ à $Filt_{10}$ et par des fermetures de Fer_1 à Fer_{10} : les points notés ■ représentent les résultats du filtre, les points notés ⋈ représentent le support des mots initiaux et les points notés • représentent les résultats de la fermeture appliquée aux 10 mots.

Pour une opération réduite à une fermeture de taille n, le pire cas est obtenu pour le $v = +-^{2n}-$, qui est transformé en $v' = +^{2n+2}$. Il est donc de longueur $2n+2$, et son support est de $2/(2n+2)$. Le support peut donc être très faible. Un tel opérateur peut ainsi identifier de longues séquences de $+$, mais avec des supports très faibles. En revanche, un opérateur de filtrage transforme le mot initial v en une séquence de $-\tilde{}$: le résultat du filtre est le mot $v' = -^{2n+2}$ de longueur $2n+2$.

Le filtre n'identifie donc pas de longues séquences avec de très faibles supports. L'utilisation du filtre alterné est donc justifiée par rapport à l'utilisation d'une simple fermeture.

À titre illustratif, nous appliquons aux 10 mots choisis aléatoirement, représentés sur la figure 4.7, la fermeture de taille allant de 1 à 10. La figure 4.8 montre les mêmes points que la figure 4.7, auxquels on ajoute les résultats de la fermeture appliquée sur ces mots, illustrés avec les points notés •. Chaque couleur représente donc un mot avec ses trois résultats : son support initial, le support et la taille de la séquence obtenue après application d'un filtre de taille n, et le support et la taille de la séquence obtenue après application d'une fermeture de taille n. On peut observer que, pour les 10 mots, les séquences obtenues à l'issue de l'application de l'opérateur de fermeture ont des tailles importantes, allant même jusqu'à la taille maximale (50), mais avec des supports très faibles par rapport à ceux obtenus avec application de l'opérateur de filtrage.

L'examen du pire des cas du filtre alterné est plus complexe, en raison de la définition récursive qui peut conduire à un support inférieur à la valeur calculée ci-dessus. En effet, comme $Filt_n(v) = Fer_n \circ Ouv_n \circ Filt_{n-1}(v)$, le pire cas de $Fer_n \circ Ouv_n$, c'est-à-dire la séquence $+^{2n+1}\,-^{2n}\,+^{2n+1}$ qui conduit à la séquence $+^{2n+1}$ peut être obtenue comme le résultat de $Filt_{n-1}$, c'est-à-dire correspondre à moins de $+$ encore dans le mot initial : elle peut en effet avoir été construite par fusion de sous-séquences de $+$, ou par comblement, par l'effet du filtre précédent. Il est donc possible que moins de symboles $+$ se réfèrent au mot initial.

Cet effet de consolidation peut être vu comme suit : la séquence construite par $Filt_1$ avec le plus petit nombre de symboles + initiaux est $u_1 = +^3 -^2 +^3$, de longueur 8. Ainsi, la séquence construite par $Filt_2$ avec le plus petit nombre de symboles + est $u_2 = u_1 -^4 u_1$. Plus généralement, en notant u_n la séquence construite par $Filt_n$ avec le plus petit nombre de symboles +, on a la relation récursive $u_{n+1} = u_n -^{2n} u_n$. Les tailles et les supports de ces séquences, notées respectivement C_n et S_n, vérifient

$$\left\{ \begin{array}{l} C_1 = 8 \\ C_n = 2C_{n-1} + 2n \end{array} \right. \quad \left\{ \begin{array}{l} SNN_1 = 6 \\ SNN_n = SNN_{n-1} \end{array} \right. \quad \left\{ S_n = \frac{SNN_n}{C_n} \right.$$

La combinaison $Fer_n \circ Ouv_n$ transforme le mot initial en une séquence de + à savoir dans le cas général, en une séquence de longueur $C_n = 2C_{n-1}+2n$ et en un support $S_n = \frac{SNN_n}{2C_{n-1}+2n}$, qui est supérieur à 0.66.

4.5 Étape d'agrégation

La méthode décrite dans les sections précédentes présente l'extraction d'une séquence caractéristique pour un chemin donné. Or, dans le cas général, un motif graduel est vérifié par plusieurs chemins complets, qui peuvent correspondre à plusieurs séquences caractéristiques. Cette section décrit l'opérateur d'agrégation proposé pour combiner les résultats obtenus à partir de ces chemins, puis discute des chemins à prendre en compte.

4.5.1 Opérateur proposé

Principe

Un motif graduel peut être vérifié par plusieurs chemins, chacun conduisant à une séquence caractéristique. Par exemple, pour les données représentées sur la figure 4.2, page 95, le motif « plus A, plus B » est vérifié par deux chemins maximaux. La représentation symbolique de ces derniers conduit aux mots v_1 et v_2 représentés sur la figure 4.9, qui chacun à leur tour conduisent à une séquence caractéristique, représentée respectivement par S_{v_1} et S_{v_2}. Dans ce cas simple, un accord élevé entre les deux séquences identifiées est observé, et la séquence caractéristique résultante, S_c, est leur intersection. En effet, dans le cadre de la caractérisation considérée, seuls les éléments les plus représentatifs sont souhaités, ce qui justifie une fonction d'agrégation sévère.

Fonction d'agrégation proposée

La fonction d'agrégation, Agg, s'applique à des mots définis sur $\{+, -, \circ\}$, ayant la même longueur, égale à la somme du nombre d'objets dans le jeu de données $\mathcal{D}$ et du nombre d'objets fictifs ajoutés. Elle s'applique successivement à chaque élément du mot et fournit en sortie un mot défini sur $\{+, \varnothing\}$. Le symbole $\varnothing$ représente les valeurs sur lesquelles le motif

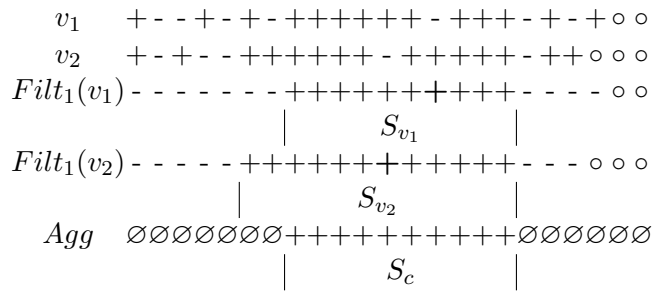


Figure 4.9 – Agrégation des séquences caractéristiques obtenues pour plusieurs chemins.

n'est pas caractérisé. La fonction Agg proposée est symétrique et définie comme suit, pour toutes les paires possibles de symboles :

$$\begin{array}{rcccc} Agg & : & \{+,-,\circ\}^2 & \rightarrow & \{+,\varnothing\} \\ & & (s_1,s_2) & \mapsto & s \end{array}$$

s_1	$+ + + - - \circ$
s_2	$+ \circ - \circ - \circ$
$Agg(s_1,s_2)$	$+ + \varnothing\varnothing\varnothing\varnothing$

Ainsi, les valeurs en dehors d'un chemin, notées $\circ$, sont neutres et n'ont pas d'influence sur les résultats ; les valeurs qui sont exclues du chemin sont associées au symbole $\varnothing$, et sont donc exclues du résultat final. En effet, le symbole $\varnothing$ indique les objets transcrits en $-$ ou en $\circ$, qui correspondent aux objets neutres ou non compatibles avec le motif considéré. Seuls les sous-mots composés de symboles $+$ sont donc conservés, car la majorité des symboles $+$ de ces sous-mots correspondent aux objets compatibles transcrits en $+$ lors du processus de transcription. Ces symboles $+$ représentent alors les valeurs sur lesquelles le motif est caractérisé. Ceci est compatible avec l'objectif de la caractérisation : seuls les éléments importants et représentatifs doivent être pris en considération.

Une fois la séquence caractéristique agrégée, l'intervalle d'intérêt de caractérisation est identifié, défini par les valeurs numériques des caractères limites de la séquence caractéristique : ce sont les valeurs minimale et maximale de l'attribut considéré pour la séquence de $+$ identifiée.

4.5.2 Chemins considérés

Les chemins considérés pour la transcription et l'agrégation sont les chemins maximaux valides, c'est-à-dire les éléments de $\mathcal{L}_s^*(M)$. En effet, la prise en compte de tous les chemins complets $\mathcal{L}(M)$ pourrait générer trop de contre-exemples, c'est-à-dire de symboles $-$ qui représentent les objets non compatibles, et ainsi conduire à un résultat vide. Ceci peut être expliqué par le fait que, dans le processus de transcription, lorsqu'un chemin est transformé en représentation symbolique, tous les objets qui ne lui appartiennent pas sont transcrits en

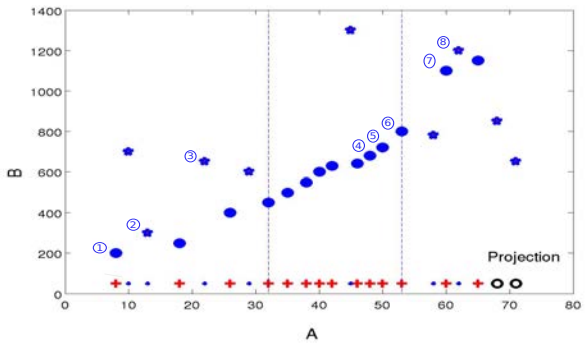


Figure 4.10 – Chemin maximal (•) et chemin complet (objets numérotés de ① à ⑧) pour le motif « plus A, plus B ».

v_1	+	−	−	+	−	+	−	+	+	+	+	−	+	+	+	+	+	−	+	−	+	∘	∘
v_2	+	−	+	−	+	−	−	−	−	−	−	−	−	−	+	+	+	−	+	+	∘	∘	∘
$Filt_1(v_1)$	−	−	−	−	−	−	−	+	+	+	+	+	+	+	+	+	+	−	−	−	−	∘	∘
$Filt_1(v_2)$	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+	−	−	−	∘	∘	∘
Agg	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	+	+	+	∅	∅	∅	∅	∅	∅

Figure 4.11 – Agrégation des séquences caractéristiques obtenues pour un chemin maximal et le chemin complet représentés sur la figure 4.10.

−, même si ces objets appartiennent à un autre chemin du motif considéré. Ce phénomène est illustré sur l'exemple de la figure 4.10, qui représente les mêmes données que celles représentées sur la figure 4.2 illustrée dans la section 4.5.1. On considère v_1 le mot correspondant au chemin maximal D de cardinal 14 représenté avec • sur la figure 4.10 et v_2 le mot correspondant au chemin complet valide de cardinal 8, représenté par les objets numérotés de 1 à 8, extrait de l'exemple de la même figure.

En appliquant un filtre d'ordre 1 sur les deux mots, on obtient une séquence de longueur 10 pour v_1 et une séquence de longueur 3 pour v_2. Comme, après agrégation, les objets transcrits en + dans le mot v_1 sont transcrits en − dans v_2, ils sont donc représentés par des ∅. Les résultats obtenus après filtrage des deux mots et le résultat de l'agrégation sont présentés dans le tableau 4.11.

Le résultat final est une séquence caractéristique de longueur inférieure à celle obtenue avec le mot issu du chemin maximal. Cela signifie qu'une séquence caractéristique de longueur 10 aurait été identifiée si le chemin complet valide n'avait pas été considéré.

4.6 Discussion sur les paramètres de la méthode proposée

La méthode proposée est basée sur quatre paramètres : le seuil de support graduel utilisé pour sélectionner les motifs graduels fréquents, s, le seuil de support graduel caractérisé utilisé pour sélectionner les motifs graduels caractérisés fréquents, s_c, l'ordre du filtre, n, et l'écart de base e.

Nous discutons ici principalement les rôles et relations des paramètres s_c et n : nous étudions leurs valeurs optimales et discutons la nécessité de la présence de ces deux paramètres dans l'approche. Une indication pour choisir la valeur de l'ordre du filtre n et pour fixer le seuil de support de caractérisation s_c est ensuite fournie.

4.6.1 Rôle individuel des paramètres

Dans cette section, nous discutons des rôles individuels des paramètres cités ci-dessus en nous basant sur leurs propriétés.

Ordre du filtre n

L'ordre du filtre n est un paramètre essentiel de notre méthode. En effet, la double condition, de par sa propriété détaillée dans la section 4.4.3, imposée par un filtre d'ordre n répond au double objectif fixé dans la section 4.2.2 : un filtre d'ordre n influence le nombre de symboles $-$, ce qui permet d'augmenter le nombre d'objets de $\mathcal{D}'$ dans l'équation (4.1). Il influence également le nombre de symboles $+$, ce qui permet d'augmenter la valeur du support dans l'équation (4.2) : une augmentation du nombre de $-$ s'accompagne donc d'une augmentation des $+$.

Il est important de noter que la satisfaction du double objectif ne signifie pas forcément l'utilisation d'un filtre d'ordre élevé. En effet, les séquences de $+$ peuvent êtres séparées par très peu de symboles $-$, ce qui rend possible leur fusion en utilisant un filtre d'ordre faible. De plus, l'application d'un filtre d'ordre très élevé peut être drastique, dans le sens où aucune séquence caractéristique ne peut être identifiée à l'issue de son application et que par conséquent aucun motif ne sera caractérisé. Ceci peut être expliqué par la propriété d'asymétrie du filtre indiquée dans la section 4.4.3.

Nous en déduisons qu'il n'existe pas de corrélation entre le fait de maximiser les objectifs des équations (4.1) et (4.2) et la valeur de l'ordre du filtre. Le choix d'une valeur optimale de n n'est pas une tâche triviale, c'est pourquoi des indications pour fixer cette valeur sont fournies dans la section 4.6.2.

Rôle de s_c

Étant donné que l'application d'un filtre d'ordre n induit une séquence contenant au moins $2n + 1$ symboles $+$ qui représentent les objets compatibles (voir section 4.4.3, page 108), on

pourrait envisager de considérer que le seuil de support de caractérisation peut être fixé comme $(2n+1)/|\mathcal{D}|$. Toutefois, ceci ne convient pas.

En effet, si ce paramètre est fixé à une valeur faible et qu'il représente le seuil du support, alors cela peut engendrer l'identification des séquences caractéristiques très courtes ayant des supports très élevés, qui ne devraient pas être valides, puisque leur longueur est très faible. Mais comme le paramètre n lui-même représente le seuil du support, alors le motif en question est validé avec une séquence caractéristique de longueur très faible.

Si, au contraire, n est fixé à une valeur très élevée et considéré comme seuil de support, alors deux types de problèmes se présentent : soit aucun motif n'est validé puisqu'il y a peu de chance d'identifier des séquences caractéristiques, soit des motifs sont identifiés avec des séquences caractéristiques très courtes ayant des supports élevés, en particulier, dans le cas de multiples chemins où la séquence caractéristique finale peut devenir très courte par effet d'agrégation.

Pour l'ensemble de ces raisons, la présence de s_c est nécessaire et n ne peut pas jouer le rôle de s_c.

Rôle de l'écart de base e

La méthode proposée dépend du paramètre e, qui détermine le nombre d'objets fictifs introduits, et donc le niveau de prise en compte de la densité : plus la valeur de e est faible, plus la contrainte imposée par la densité est importante. Si e est inférieur à l'écart minimal e_{min} observé entre deux données consécutives, aucune clause de caractérisation ne peut être identifiée : les objets transcrits par des + sont tous séparés par des objets fictifs transcrits par des −, et aucune séquence de + n'est alors identifiée. Si e est supérieur à l'écart maximal, e_{max}, entre deux valeurs successives, aucun objet fictif n'est introduit et la densité n'a pas d'influence sur le résultat.

4.6.2 Discussion sur la relation entre les paramètres n et s_c

L'objectif dans cette section est double : il s'agit tout d'abord de souligner les facteurs qui peuvent avoir un impact sur le choix de la valeur optimale de n, et de déterminer ensuite la relation qui peut exister entre n et s_c.

Le paramètre n dépend directement de l'ensemble de données utilisé. Plus précisément, il dépend nécessairement de la longueur du chemin considéré. Ainsi, une première indication pour choisir la valeur de n est de ne pas fixer celui-ci à une valeur supérieure ou égale à la longueur du chemin maximal considéré lors de la transcription des données. En effet, seul ses objets sont transcrit en +. De plus, comme la propriété du filtre (voir section 4.4.3) précise qu'après application d'un filtre d'ordre n, une séquence ayant au moins $2n+1$ symboles + représentant les objets compatibles est obtenue, n peut être fixé à une valeur plus faible que la taille du chemin considéré, comme par exemple une valeur égale à taille du chemin divisée par 2. Cependant, cette indication n'est pas pertinente pour déterminer clairement une valeur optimale pour n. En effet, la longueur d'un chemin peut être différente d'un motif à un autre.

On peut toutefois prendre en compte la longueur la plus faible des chemins de tous les motifs. Celle-ci ne peut être estimée dans le cas général, mais elle est contrainte par la valeur du seuil s_c. On peut donc envisager une valeur inférieure ou égale à $n = |\mathcal{D}| \times s_c/2$.

Le paramètre s_c peut avoir un double rôle : intervenir dans la validation des motifs caractérisés et déterminer la valeur de n. Si s_c est fixé à une valeur faible, alors n doit avoir une valeur plus faible que celui-ci. En effet, le sous-ensemble d'objets caractéristiques obtenu à l'issue d'un filtre d'ordre n à partir duquel le support caractéristique est calculé contient au moins $2n + 1$ objets compatibles (voir propriétés du filtre, section 4.4.3). n peut alors être fixé, par exemple, à une valeur égale à $\frac{s_c}{2}$, où s_c représente le nombre minimum d'objets compatibles, et non pas en pourcentage.

Si, au contraire, s_c est fixé à une valeur élevée, alors l'ordre du filtre peut aussi être fixé à une valeur élevée. Cependant, pour que le filtre soit efficace, la valeur de n ne doit pas dépasser $n = |\mathcal{D}| \times s_c/2$. Toutefois, il est préférable de fixer le paramètre n à une valeur inférieure à $n = |\mathcal{D}| \times s_c/2$. En effet, nous n'avons pas d'information a priori sur la manière dont les symboles + sont distribués tout au long du mot à filtrer : on peut avoir des grandes séquences de + séparées de petites séquences de −, comme on peut avoir des petites séquences de + suivies de séquences séries de −. Or, dans ce dernier cas, comme il a été déjà indiqué ci-dessus, un filtre d'ordre n élevé, égal par exemple à $n = |\mathcal{D}| \times s_c/2$, peut ne pas identifier de séquences caractéristiques, car il transforme toutes les petites séquences de + en séquences de −.

4.7 Expérimentations et résultats

Nous avons effectué une étude expérimentale de la méthode de caractérisation proposée. L'analyse des résultats est basée sur la comparaison des supports des motifs graduels avant et après caractérisation ainsi que le nombre de motifs extraits dans chacun des cas. En considérant la densité des données, une analyse du comportement de l'approche en variant l'écart de base est effectuée. Puis une comparaison des résultats obtenus avec le cas où la densité des données n'est pas prise en compte est réalisée.

Les expérimentations menées avec exploitent les mêmes données que celles utilisées dans l'approche proposée dans le chapitre 3. Ces données sont décrites en annexe A.1, page 145.

Pour ces expérimentations, nous fixons un seuil de support graduel minimal à $s = s_c = 20\%$ et l'ordre du filtre à $n = 4$.

4.7.1 Motifs caractérisés extraits

Parmi les 835 motifs extraits par GRITE, 509 sont enrichis par une clause de caractérisation, soit plus de 60% des motifs extraits. Les motifs caractérisés peuvent être illustrés par les exemples suivants, ayant un SG_c parmi les plus élevés :

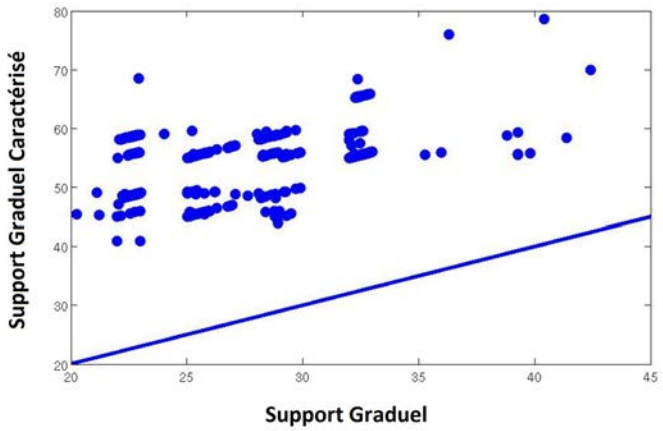


Figure 4.12 – Comparaison des supports graduels avant et après caractérisation.

— Plus la température est élevée, moins la vitesse du vent est élevée, surtout si la vitesse du vent $\in [1, 10]$, $SG = 36.3\%$, $SG_c = 78.6\%$. Les valeurs de la « vitesse du vent » varient entre 1 et 14.8.

— Plus la pression est élevée, plus la température est élevée, surtout si la température $\in [13, 19.2]$, $SG = 22\%$, $SG_c = 76\%$. Les valeurs de la « température » varient entre 3.1 et 21.8.

— Moins l'humidité est élevée, moins la température est élevée, surtout si la température $\in [8.1, 12.2]$, $SG = 22.8\%$, $SG_c = 70\%$.

La figure 4.12 compare les supports graduels des motifs avant et après caractérisation. Elle montre qu'après caractérisation, tous les supports graduels obtenus sont supérieurs à ceux obtenus avant caractérisation, ce qui confirme la validité accrue des motifs graduels. Le support graduel le plus élevé avant caractérisation est de 42.4%. En revanche, le support graduel le plus élevé après caractérisation est de 78.6%, ce qui est bien supérieur à celui obtenu avant caractérisation.

4.7.2 Prise en compte de la densité

Dans cette expérimentation, l'écart de base de chaque attribut présent dans la base de données est fixé à l'écart moyen entre deux valeurs successives présentes dans la base de données.

En prenant en compte ainsi la densité des données, 461 motifs graduels caractérisés sont extraits, ce qui correspond à plus de 55% des motifs extraits (parmi les 835 motifs graduels), c'est-à-dire 48 de moins que lorsque la densité n'est pas prise en compte. Cela s'explique par l'insertion des objets fictifs dans les mots transcrits et qui sont considérés comme des $-$ lors de l'application du filtre.

Le tableau 4.1 compare les trois meilleurs motifs graduels caractérisés obtenus dans la section précédente avec l'approche de caractérisation sans prise en compte de la densité,

Motif	SG %	Sans prise en compte		Avec prise en compte	
		clause de caractérisation	SG_c %	clause de caractérisation	SG_c %
M_1	36.3	la vitesse du vent $\in [1, 10]$	78.6%	la vitesse du vent $\in [1, 7.4]$	86.2
M_2	22	la température $\in [13, 19.2]$	76	la température $\in [13, 19.2]$	76
M_3	22.8	la température $\in [8.1, 12.2]$	70	la température $[9, 11.8]$	76.5

Tableau 4.1 – Comparaison des trois meilleurs motifs graduels caractérisés obtenus sans et avec prise en compte de densité.

Tableau 4.2 – Nombre de motifs caractérisés extraits en fonction de l'écart utilisé e pour l'attribut « vitesse du vent »

e	$e_{max} = 0.1$	0.01	0.005	$e_{min} = 0.003$
# motifs	51	37	7	0

avec ceux obtenus dans cette section avec prise en compte de la densité. Les motifs graduels classiques correspondant à ces motifs caractérisés sont M_1, M_2 et M_3 donnés ci-dessous :

— M_1 : plus la température est élevée, moins la vitesse du vent est élevée.

— M_2 : plus la pression est élevée, plus la température est élevée.

— M_3 : moins l'humidité est élevée, moins la température est élevée.

La figure 4.13 compare les supports graduels des motifs et la figure 4.14 la taille des séquences caractéristiques, avec et sans prise en compte de la densité. Les résultats sans prise en compte de la densité sont représentés sur l'axe des abscisses et les résultats avec prise en compte de la densité sur l'axe des ordonnées.

On constate que la méthode avec prise en compte de la densité extrait des motifs avec des supports plus élevés : le support le plus élevé est de 86.2%, tandis qu'il vaut seulement 78.6% sans prise en compte de la densité. En revanche, les longueurs des séquences caractéristiques obtenues sont plus faibles : la séquence la plus longue obtenue sans prise en compte de la densité est de 779 contre 777 avec prise en compte de la densité.

Résultats des variations de l'écart de base

Le tableau 4.2 montre le nombre de motifs caractérisés extraits en fonction de la valeur de e, pour l'attribut « vitesse du vent ». Ce dernier est associé à des valeurs comprises entre 1 et 14.8 et son écart de base est $e = 0.01$.

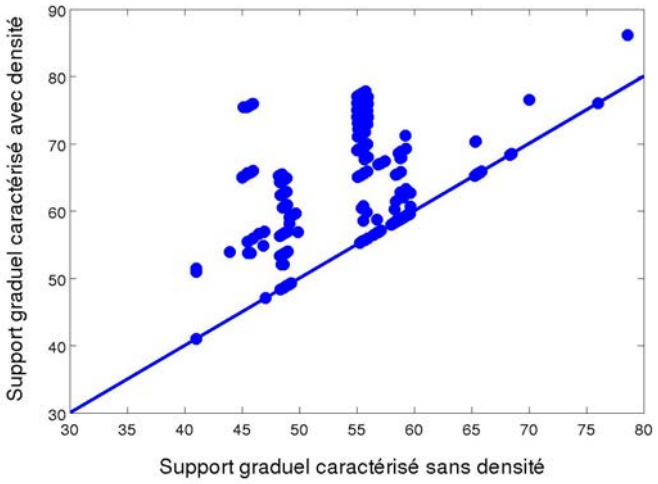


Figure 4.13 – Comparaison des supports graduels des motifs avec et sans prise en compte de la densité.

Comme attendu d'après la discussion sur le rôle de e détaillé dans la section 4.6.1, en utilisant l'écart maximum, on obtient les mêmes motifs que sans prise en compte de la densité. Leur nombre diminue quand e diminue. En effet, seulement 7 motifs parmi les 37 obtenus avec $e = 0.01$ sont retenus. De plus, ces 7 motifs sont caractérisés par des séquences de longueur très faible et des supports élevés. Ainsi, le premier exemple donné dans la sous-section 4.7.1 est caractérisé par un intervalle très étroit $[4.6, 5]$ et la longueur de sa séquence caractéristique est de 57 avec un support de 100%, tandis que la longueur de la séquence caractéristique pour le même motif en utilisant $e_{max} = 0.1$ est de 640. En utilisant un écart encore plus faible, aucun motif caractérisé avec l'attribut considéré n'est obtenu.

Lorsqu'on prend un écart supérieur ou égal à l'écart maximal, l'approche extrait les mêmes motifs que sans prise en compte de la densité, c'est-à-dire les 461 motifs.

4.7.3 Évaluation des performances

En ce qui concerne la consommation en termes de temps et de mémoire, le temps d'extraction dépend très fortement du nombre d'attributs et d'objets de la base de données, ainsi que du seuil de support graduel utilisé. Les expérimentations montrent clairement que la phase de caractérisation est peu coûteuse. Pour la base de données réelles météorologiques (voir annexe A.1, page 145), le temps d'extraction des motifs graduels caractérisés est d'environ une minute. Si on regarde le temps nécessaire pour, à la fois la phase d'extraction des motifs graduels fréquents et la phase de caractérisation, alors le temps d'extraction peut être long et ne permet pas d'extraction en temps réel. Toutefois, la phase de caractérisation elle-même montre que le temps qu'elle nécessite est très acceptable et beaucoup plus faible que celui de la phase d'extraction de motifs graduels fréquents avec l'algorithme GRITE. Plus précisément, 90% du temps total nécessaire à l' extraction est utilisé dans l'étape d'extraction des motifs graduels classiques, 10% seulement est utilisé dans l'étape de caractérisation.

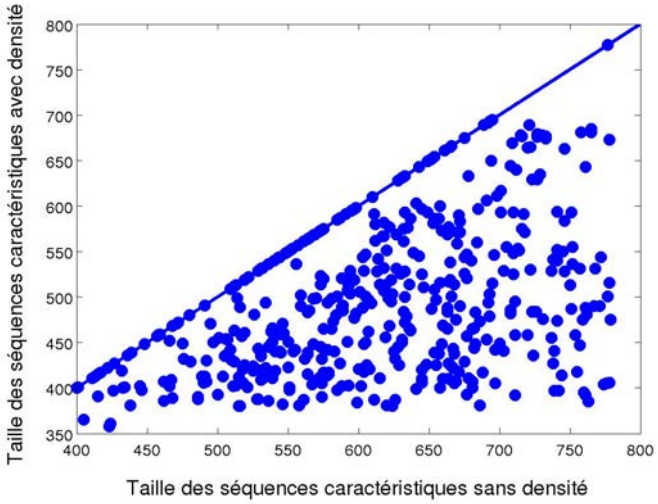


Figure 4.14 – Comparaison des taille des séquences caractéristiques des motifs avec et sans prise en compte de la densité.

4.8 Conclusion

Dans ce chapitre, nous avons proposé une nouvelle approche permettant la caractérisation des motifs graduels, en utilisant une clause linguistiquement introduite par l'expression « surtout si », afin d'extraire plus d'informations résumant un ensemble de données. Cette approche repose sur l'identification d'intervalles d'intérêt pour les attributs apparaissant dans le motif considéré et vérifiant une contrainte de densité des données. La caractérisation de motifs graduels conduit à une meilleure interprétation des motifs extraits.

Nous avons proposé un support graduel de caractérisation pour mesurer la pertinence des motifs graduels caractérisés basée sur une interprétation comme validité accrue du motif. L'approche proposée est basée sur les outils de morphologie mathématique.

La caractérisation que nous avons proposée extrait une information distincte de celle extraite dans le chapitre précédent : en effet, l'objectif de la caractérisation est la validité accrue des motifs extraits, tandis que l'objectif du traitement de la contradiction est de lever l'ambiguïté entre motifs. De plus, la caractérisation est locale (indépendante des autres motifs) et elle se base sur l'identification d'intervalles d'intérêt, alors que le traitement de la contradiction est global (concerne un sous-ensemble de motifs) et se base sur l'identification de sous-ensembles d'objets propres à chacun des motifs contradictoires. Nous avons illustré la pertinence de l'approche avec différentes expérimentations sur des données réelles.

Nous définissons aussi les *motifs graduels accélérés*, qui qualifient les corrélations entre les valeurs d'attributs et contextualisent les motifs graduels par l'expression linguistique « rapidement »

5

Motifs graduels accélérés

Sommaire

Introduction

Les motifs graduels traduisent une variation des valeurs d'attributs, mais ils n'expriment pas d'information sur la manière dont les valeurs des attributs varient par rapport aux autres. Nous présentons dans ce chapitre un nouveau type d'enrichissement dans le cas de données numériques, permettant de capturer un nouveau type d'information : la rapidité avec laquelle certains attributs varient par rapport aux autres. Nous proposons d'exprimer cette information avec l'expression linguistique *rapidement*, et introduisons la sémantique des *motifs graduels accélérés*. De tels motifs peuvent être illustrés par l'exemple « plus la vitesse du

vent augmente, plus la distance parcourue par le vent augmente rapidement » ou « plus la température d'un gaz augmente, plus sa pression augmente rapidement ».

Le principe de l'accélération est naturellement compris comme une augmentation de la variation de la vitesse, qui peut être traduite comme une contrainte de convexité de la fonction sous-jacente associée aux attributs considérés. Cette contrainte peut être modélisée comme une contrainte de covariation supplémentaire, conduisant à la définition d'un nouveau critère de qualité permettant d'évaluer la validité de ces motifs graduels accélérés, que nous appelons *support graduel accéléré*.

Le chapitre est organisé de la façon suivante : la section 5.1 présente la motivation et le principe des motifs graduels accélérés dans le cas de motifs de longueur 2. Dans la section 5.2, nous définissons le support graduel accéléré qui permet d'évaluer la pertinence de tels motifs candidats et illustrons son calcul sur un exemple. La section 5.3 détaille l'algorithme qui permet leur extraction et la section 5.4 présente la généralisation de ces motifs à des longueurs supérieures. Enfin, la section 5.5 illustre et analyse les résultats expérimentaux obtenus sur une base de données réelles.

Ces travaux ont été publiés dans (Oudni et al. 2014).

5.1 Motivation et formalisation

Cette section présente l'interprétation et le principe des motifs graduels accélérés, en l'illustrant sur un exemple, ainsi que la formalisation proposée.

5.1.1 Principe et interprétation des motifs graduels accélérés

Motivation

La figure 5.1 représente deux ensembles de données de même cardinalité décrits par deux attributs, A en abscisse et B en ordonnée. Ils conduisent tous les deux à un même motif graduel $M = A^{\geq}B^{\geq}$ dont le support graduel est 100% dans les deux cas. Sa caractérisation par un intervalle donne le même attribut caractéristique A, et le même intervalle caractéristique, égal à l'ensemble du domaine de A, c'est-a-dire $[0, 18]$. Or, on peut observer que la covariation entre A et B est différente : un effet d'accélération des valeurs de B par rapport à celles de A est observé pour l'ensemble de droite, alors qu'il n'est pas valable pour l'ensemble de gauche. Ceci motive donc l'extraction d'un motif graduel dit accéléré dans le cas de la figure de droite sous la forme « plus A augmente, plus B augmente rapidement ».

Il faut souligner que les motifs graduels accélérés ne sont pas symétriques et distinguent le cas « plus A augmente, plus B augmente rapidement » du cas « plus B augmente, plus A augmente rapidement », alors que, dans les deux cas, le motif graduel est « plus A augmente, plus B augmente ».

Il faut souligner également que les motifs graduels accélérés s'appliquent à des données numériques, et non à des données catégorielles ou floues.

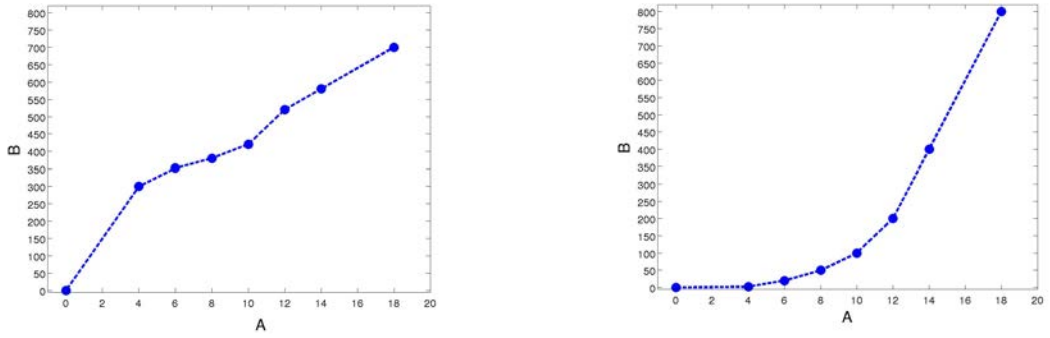


Figure 5.1 – Deux ensembles de données, conduisant à « *plus A augmente, plus B augmente* » où un effet d'accélération est observé pour l'ensemble de droite et non pour celui de gauche.

Formulation mathématique et interprétation

Mathématiquement, l'effet d'accélération correspond à une propriété de convexité de la fonction qui associe les valeurs de B aux valeurs de A, imposant que son graphe soit « tourné vers le haut » comme illustré sur la partie droite de la figure 5.1 : le segment de droite situé entre deux points du graphe de la fonction se situe entièrement au-dessus du graphe. Une croissance convexe signifie une augmentation à un rythme de plus en plus élevé (mais pas nécessairement proportionnelle à la valeur courante), qui est équivalent à l'effet d'accélération souhaité. Pour les fonctions dérivables, la propriété de convexité est équivalente à une condition sur la dérivée : une fonction dérivable est convexe si et seulement si sa dérivée est monotone croissante.

Comme les ensembles de données à partir desquels les motifs graduels accélérés doivent être extraits ne donnent pas accès à la fonction mathématique liant les valeurs de A et B, sa dérivée ne peut donc être calculée. Par conséquent, nous proposons de considérer une discrétisation, définie comme le quotient des différences successives $\left(\frac{\Delta B}{\Delta A}\right)$ lorsque les données sont ordonnées par rapport à leurs valeurs de A. L'attribut sur lequel porte l'adverbe « rapidement » est au numérateur.

Nous proposons donc d'interpréter l'effet de l'accélération comme une augmentation de $\left(\frac{\Delta B}{\Delta A}\right)$. Il est important de noter que cette interprétation ne tient pas compte de la forme de la fonction convexe : elle ne fait pas de différence par exemple entre le fait que la fonction sous-jacente soit quadratique ou exponentielle. De plus le cas linéaire, pour lequel le quotient est constant, est un cas limite d'augmentation.

Originalité

La principale différence entre les motifs graduels accélérés et les enrichissements par renforcement ou par caractérisation présentés dans les chapitres 2 et 4 respectivement provient de la nature de la clause additionnelle : pour la caractérisation et le renforcement, la clause d'enrichissement a une sémantique présentielle, dans la mesure où la présence de cette contrainte supplémentaire conduit à une restriction de données sur laquelle la validité du motif doit augmenter. En revanche, comme discuté plus en détail dans les sections suivantes, la sémantique

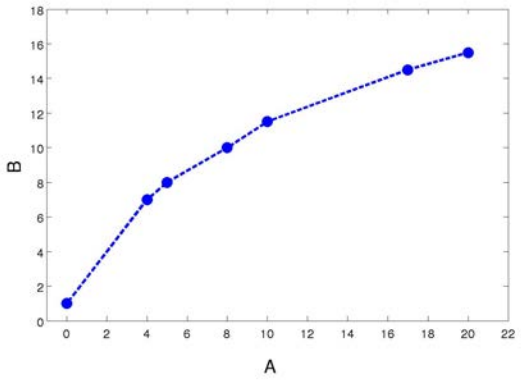


Figure 5.2 – $A^{\geq}B^{\geq}\left(\frac{\Delta B}{\Delta A}\right)^{\leq}$ avec un effet de décélération.

de la clause d'accélération est graduelle, c'est-à-dire de même nature que le motif graduel considéré.

5.1.2 Formalisation de l'accélération

Définition 5.1 (Motif graduel accéléré)**.** Un motif graduel accéléré de longueur 2 est défini comme un triplet : $A^{*_1}B^{*_2}\left(\frac{\Delta B}{\Delta A}\right)^{*_3}$, où $A^{*_1}B^{*_2}$ représente un motif graduel, et $\left(\frac{\Delta B}{\Delta A}\right)^{*_3}$ représente la clause d'accélération qui compare les variations de B à celles de A.

$*_1$ détermine si « plus A augmente » ($*_1 = \geq$) ou « plus A diminue » ($*_1 = \leq$), et $*_2$ joue le même rôle pour B. $*_3$ détermine si une accélération ou une décélération est considérée : $*_3 = \geq$ conduit à « plus B augmente rapidement » et $*_3 = \leq$ conduit à « moins B augmente rapidement » soit aussi « plus B augmente lentement ».

Notre travail est focalisé sur l'effet d'accélération, c'est-à-dire sur les attributs pour lesquels les valeurs augmentent « rapidement », soit le cas où $*_3 = \geq$. Ce cas représente le cas d'une courbe convexe. Le cas où $*_3$ représente $\leq$ correspond à un effet de décélération, comme illustré sur la figure 5.2, qui peut être décrit comme « plus A augmente, plus B augmente lentement ». Il faut noter que cela est équivalent à « plus B diminue, plus A augmente rapidement », à savoir $A^{\geq}B^{\geq}\left(\frac{\Delta A}{\Delta B}\right)^{\geq}$. Considérer uniquement le cas $*_3 = \geq$ n'est donc pas une limitation.

5.2 Évaluation de l'effet d'accélération

Un motif graduel accéléré de longueur 2 MM_a contient deux composantes : le motif graduel classique $M = A^{*_1}B^{*_2}$ et la clause d'accélération $M_a = \left(\frac{\Delta B}{\Delta A}\right)^{*_3}$. Il doit donc être évalué en fonction de ces deux éléments. Sa qualité est mesurée à la fois par le support graduel classique (équation (1.7), page 40) et un support graduel accéléré qui mesure la qualité de l'accélération, défini dans cette section.

5.2.1 Pré-ordre induit par la clause d'accélération

Le motif graduel M induit un pré-ordre sur des objets, comme défini dans le chapitre 1 ; la clause d'accélération $\left(\frac{\Delta B}{\Delta A}\right)^{*_3}$ induit également un pré-ordre noté $\preceq_{M_a}$, défini sur des paires d'objets.

Définition 5.2 (Ordre induit par la clause d'accélération)**.** Pour tout o_1, o_2, o_3 et $o_4 \in \mathcal{D}$

$$(o_1, o_2) \preceq_{M_a} (o_3, o_4) \Leftrightarrow \frac{B(o_2) - B(o_1)}{A(o_2) - A(o_1)} *_3 \frac{B(o_4) - B(o_3)}{A(o_4) - A(o_3)} \ . \tag{5.1}$$

où $A(o)$ et $B(o)$ représentent respectivement les valeurs des attributs A et B pour l'objet o.

5.2.2 Définition du support graduel accéléré

Le support défini dans cette section permet d'évaluer la validité de l'accélération de motifs graduels accélérés. La qualité du motif graduel accéléré candidat MM_a est élevée s'il existe un sous-ensemble de données qui satisfait simultanément l'ordre induit par M et celui induit par M_a. Par conséquent, la qualité de l'accélération nécessite d'abord d'identifier un sous-ensemble qui ne satisfait que $\preceq_M$. Pour cela, l'algorithme GRITE (Di Jorio et al., 2009) peut être utilisé pour identifier les motifs graduels candidats ainsi que l'ensemble de leurs chemins complets maximaux $\mathcal{L}^*(M)$ (voir chapitre 1, page 39).

Pour tout chemin D, le calcul du support graduel accéléré consiste alors à identifier un sous-ensemble de D de sorte que la contrainte $\left(\frac{\Delta B}{\Delta A}\right)^{*_3}$ soit également vérifiée.

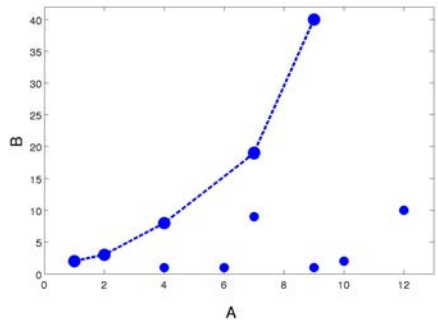
Notons φ la fonction qui permet d'identifier le sous-ensemble d'objets de D tel que $\forall o_1, o_2, o_3 \in \varphi(D)$, $o_1 \preceq_M o_2 \preceq_M o_3 \Rightarrow (o_1, o_2) \preceq_{M_a} (o_2, o_3)$

Définition 5.3 (Support graduel accéléré)**.** Le support graduel accéléré est défini comme :

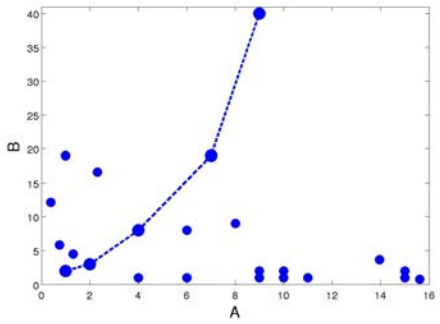
$$SG_a = \frac{1}{|D| - 1} \max_{D \in \mathcal{L}^*(M)} |\varphi(D)| \tag{5.2}$$

où $|D|$ représente la taille de tout chemin complet dans $\mathcal{L}^*(M)$, puisque, par définition de $\mathcal{L}^*(M)$, tous les chemins de $\mathcal{L}^*(M)$ ont la même taille. $|D| - 1$ est alors la valeur maximale possible de $\varphi(D)$ et représente donc le facteur de normalisation. En effet, $\left(\frac{\Delta B}{\Delta A}\right)$ n'est pas un motif graduel classique, car il ne possède pas le même ensemble de définition : il s'applique à des paires d'objets.

La prise en compte des chemins complets maximaux $\mathcal{L}^*(M)$ et non de tous les chemins complets $\mathcal{L}(M)$ dans le calcul du support graduel accéléré donné dans l'équation (5.2) est justifié de la façon suivante : si un motif graduel accéléré est extrait en utilisant un chemin plus court que le chemin maximal, alors un support graduel plus faible est associé au motif graduel à enrichir et la proportion de données sur lesquelles le support graduel accéléré est calculé est très faible, ce qui peut induire un support graduel accéléré très élevé. Le problème



(a) $A^{\geq}B^{\geq}(\frac{\Delta B}{\Delta A})^{\geq}$ avec un support graduel élevé.



(b) $A^{\geq}B^{\geq}(\frac{\Delta B}{\Delta A})^{\geq}$ avec un support graduel faible.

Figure 5.3 – Deux ensembles de données pour lesquels le motif $A^{\geq}B^{\geq}(\frac{\Delta B}{\Delta A})^{\geq}$ a un SG différent (45% à gauche et 22% à droite) et des $SG_a = 100\%$ identiques.

qui se pose alors est le fait d'obtenir un motif accéléré ayant un support d'accélération très élevé mais uniquement sur une très petite proportion de données. Cela peut poser des difficultés de lisibilité pour l'utilisateur. Pour remédier à cela, prendre en compte uniquement les chemins maximaux est une solution efficace, car on peut garantir que le sous-ensemble de données pouvant être considéré lors du calcul du support graduel accéléré représente la plus grande proportion de données vérifiant le motif à enrichir et pouvant être pertinente pour l'accélération. Si celle-ci ne permet pas d'obtenir un support d'accélération suffisant, alors une proportion plus faible est considérée comme moins satisfaisante pour valider un motif graduel accéléré.

5.2.3 Combinaison des critères de qualité

La définition de la validité classique est ensuite étendue pour intégrer la condition sur SG_a.

Définition 5.4 (Validité d'un motif graduel accéléré)**.** Un motif graduel accéléré MM_a est *valide* si $SG \geq s$ et $SG_a(MM_a) \geq s_a$ où s_a est le seuil pour le support graduel accéléré et s le seuil de support graduel classique.

Il faut souligner que les deux supports SG et SG_a sont nécessaires pour évaluer la qualité d'un motif graduel accéléré. La figure 5.3 illustre le cas de deux ensembles de données conduisant à un même $SG_a = 100\%$, mais avec des SG différents : il vaut 45% pour l'ensemble de gauche et 22% pour l'ensemble de droite. En effet, SG_a est calculé par rapport à la taille du chemin, tandis que SG prend en compte le nombre total d'objets.

En combinant les deux composantes, une priorité est donnée à SG : pour un niveau de SG donné, les motifs graduels accélérés sont comparés selon leur SG_a.

5.2.4 Exemple illustratif

Nous illustrons ici le calcul du support graduel accéléré donné dans l'équation (5.2) sur la base de données de la figure 5.4 contenant $n = 10$ objets décrits par 2 attributs numérotés

Id.	A	B
1	102	100
2	106	110
3	138	110
4	50	170
5	102	180
6	200	1090
7	400	1090
8	500	1490
9	900	1790
10	1100	1900

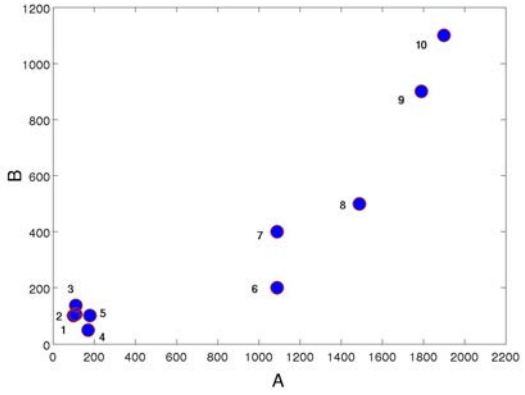


Figure 5.4 – Exemple illustratif

D_1	A	B	a_1	a_2	$\varphi_{a_1}(D_1)$	$\varphi_{a_2}(D_1)$
1	170	50	0.19	5.2	1	
2	180	102	15.17	0.07	2	2
6	1090	200	1.33	0.75		6
8	1490	500	0.75	1.33		8
9	1790	900	0.55	1.82		9
10	1900	1100				

Tableau 5.1 – Calcul de $\varphi_{a_1}(D_1)$ et $\varphi_{a_2}(D_1)$ pour les deux motifs graduels accélérés M_{a_1} et M_{a_2}.

D_2	B	A	a_1	a_2	$\varphi_{a_1}(D_2)$	$\varphi_{a_2}(D_2)$
4	100	102	2.5	0.4	4	
5	110	106	10.43	0.1	5	5
6	1090	200	1.33	0.75		6
8	1490	500	0.75	1.33		8
9	1790	900	0.55	1.82		9
10	1900	1100				

Tableau 5.2 – Calcul de $\varphi_{a_1}(D_2)$ et $\varphi_{a_2}(D_2)$ pour les deux motifs graduels accélérés M_{a_1} et M_{a_2}.

selon les valeurs de l'attribut A. On a pour le motif graduel $M = A^{\geq}B^{\geq} : \mathcal{L}^*(M) = \{D_1, D_2\}$ avec $D_1 = \{1, 2, 6, 8, 9, 10\}$ et $D_2 = \{4, 5, 6, 8, 9, 10\}$. Ils ne diffèrent que par les deux premiers objets. Dans cet exemple, le seuil de support graduel accéléré est fixé à 50%.

On note $a_1 = \dfrac{\Delta A}{\Delta B}$ et $a_2 = \dfrac{\Delta B}{\Delta A}$. Les tableaux 5.1 et 5.2 donnent leurs valeurs pour chacun des deux chemins maximaux vérifiant M.

On note $M_{a_1} = A^{\geq}B^{\geq}\left(\dfrac{\Delta A}{\Delta B}\right)^{\geq}$, $M_{a_2} = A^{\geq}B^{\geq}\left(\dfrac{\Delta B}{\Delta A}\right)^{\geq}$ les deux motifs graduels accélérés candidats. $\varphi_{a_1}(D_1)$, $\varphi_{a_1}(D_2)$ les chemins associés à M_{a_1} et $\varphi_{a_2}(D_1)\}$, $\varphi_{a_2}(D_2)\}$ les chemins associés à M_{a_2}.

On constate que les sous-ensembles d'objets de D_1 et de D_2 qui vérifient la contrainte d'ordre $a_1^{\geq}$ sont de taille 2 seulement, soit $SG_a = 2/5 = 40\%$, ce qui ne permet pas de valider le motif graduel accéléré M_{a_1}. Au contraire $\varphi_{a_2}(D_1)\}$, $\varphi_{a_2}(D_2)\}$ sont tous les deux de longueur 4, soit $SG_a = 4/5 = 90\%$ validant M_{a_2}. Ces résultats sont compatibles avec l'observation de la représentation graphique de la figure 5.4 : une accélération des valeurs de l'attribut B par

rapport à celles de A est vérifiée sur le sous-ensemble d'objets $\{6, 8, 9, 10\}$ commun aux deux chemins maximaux D_1 et D_2.

5.3 Algorithme d'extraction

Nous décrivons dans cette section l'approche que nous proposons pour la génération et la sélection des motifs graduels accélérés composés de deux attributs. Elle exploite le paradigme générer-élaguer et opère en deux étapes, appliquées au résultat de l'extraction des motifs graduels par GRITE.

1. Calcul des quotients des écarts successifs pour chaque chemin.
2. Identification des clauses d'accélération.

Le pseudo-code de la deuxième étape est présenté dans l'algorithme 4 : il prend en entrée l'ensemble de motifs graduels fréquents associés à leurs chemins maximaux, ainsi que le seuil de support graduel d'accélération s_a. En sortie, il renvoie les motifs graduels accélérés.

Algorithm 4 Génération des motifs graduels accélérés

Input : $\mathcal{M}$ ensemble de motifs graduels fréquents, s_a : seuil de support graduel accéléré.
Output : $\mathcal{M}_a$ l'ensemble de motifs graduels accélérés
for all $M = A^{*_1}B^{*_2} \in \mathcal{M}$ **do**
  $valCourante = 0$
  for all $D \in \mathcal{L}^*(M)$ **do**
    calculer $a_1 = \left(\frac{\Delta A}{\Delta B}\right)$ et $\varphi_{a_1}(D)$
    calculer $a_2 = \left(\frac{\Delta B}{\Delta A}\right)$ et $\varphi_{a_2}(D)$
    $m = \max(|\varphi_{a_1}(D)|, |\varphi_{a_2}(D)|)$
    if $m > valCourante$ **then**
      $\text{clauseAcc} = \underset{\{a_1,a_2\}}{\text{argmax}}(|\varphi_{a_1}(D)|, |\varphi_{a_2}(D)|)$
      $valCourante = m$
    end if
  end for
  if $\frac{valCourante}{|D|} > s_a$ **then**
    $\mathcal{M}_a \leftarrow \mathcal{M}_a \cup clauseAcc^{\geq}$
  end if
end for

L'algorithme commence par calculer, pour chaque paire d'objets successifs appartenant à un chemin maximal D, les variations de l'attribut A par rapport à B, $\left(\frac{\Delta A}{\Delta B}\right)$, et celles de B par rapport à A, $\left(\frac{\Delta B}{\Delta A}\right)$. Il calcule ensuite les supports graduels accélérés des deux motifs graduels accélérés $M\ \left(\frac{\Delta A}{\Delta B}\right)^{\geq}$et $M\ \left(\frac{\Delta B}{\Delta A}\right)^{\geq}$ et conserve celui ayant le support graduel accéléré le plus élevé si celui-ci dépasse le seuil s_a. L'idée sous-jacente à ce choix est d'éviter le conflit qui peut survenir entre deux motifs présentant un effet d'accélération de chacun des attributs qui le composent. Considérons ainsi l'exemple illustré dans la section précédente : si on fixe un seuil de support graduel accéléré égal à 30%, alors les deux motifs $A^{\geq}B^{\geq}\left(\frac{\Delta A}{\Delta B}\right)^{\geq}$ et $A^{\geq}B^{\geq}\left(\frac{\Delta B}{\Delta A}\right)^{\geq}$ auraient été extraits simultanément conduisant à une notion d'accélération

contradictoire. Celle-ci pourrait être traitée avec les principes proposés dans le chapitre 3, pour garantir un sous-ensemble d'objets propres à chacun de ces motifs graduels accélérés.

Ce traitement est laissé comme perspective à court terme et non effectué dans l'algorithme 4 qui sélectionne parmi les deux candidats celui qui maximise le support graduel accéléré.

5.4 Généralisation

5.4.1 Définitions

La définition précédente considère les motifs graduels composés de deux attributs. Dans le cas général, le motif graduel à enrichir peut être composé de plusieurs attributs, de même pour la clause d'accélération.

Définition 5.5 (Motif graduel accéléré général)**.** Un motif graduel accéléré général est de la forme $M = M_1 M_2 M_a$ et tel que

$$\begin{cases} M_1 \text{ et } M_2 \text{ sont des motifs graduels classiques} \\ M_1 \cap M_2 = \emptyset \\ M_a = \left\{ \left(\frac{\Delta B_m}{\Delta A_n} \right)^{\geq}, A_n \in M_1, B_m \in M_2 \right\} \end{cases}$$

Cette définition générale présente trois cas particuliers, selon la taille des motifs graduels M_1 et M_2 : le cas où $|M_1| = |M_2| = 1$, qui correspond au cas détaillé précédemment ; le cas où $|M_1| = 1$ et $M_2 \geq 2$, et le cas où $|M_1| \geq 2$ et $M_2 = 1$. Ces deux derniers cas sont commentés en détails dans la section 5.4.3.

Dans le cas général, la définition ci-dessus impose que chaque attribut de M_2 présente un effet d'accélération pour chacun des attributs de M_1 : un tel motif graduel général est *valide* si et seulement si

Définition 5.6 (Validité d'un motif graduel accéléré général)**.** Un motif graduel accéléré est *valide* s'il existe deux sous-ensembles de données D et $D' \subset D$ tels que tous les items graduels présents dans M_1 et dans M_2 sont vérifiés sur D et $\frac{|D|}{|\mathcal{D}|} > s$ et tous les items graduels présents dans M_a sont vérifiés sur D' et $\frac{|D'|}{|\mathcal{D}|} > s_a$. Formellement, $\forall o_1, o_2, o_3 \in D'$

$$(o_1 \preceq_{M_1 M_1} o_2 \preceq_{M_1 M_1} o_3) \Rightarrow \forall A_n \in M_1, \forall B_m \in M_2, (o_1, o_2) \preceq_{\frac{\Delta B_m}{\Delta A_n}} (o_2, o_3)$$

Le support graduel accéléré du motif $M = M_1 M_2 M_a$ est calculé avec l'équation (5.2), page 125.

Linguistiquement, un tel motif est interprété comme : plus A_1 augmente, ..., plus A_n augmente, plus B_1 augmente rapidement et ... et plus B_m augmente rapidement. Un tel motif peut être illustré par les exemples comme « plus l'altitude est basse, plus il y a d'air,

plus la pression augmente rapidement, plus la température augmente rapidement » ou « plus la vitesse des rafales de vent augmente, plus la vitesse du vent augmente, plus la distance parcourue par le vent augmente rapidement, plus l'humidité augmente rapidement ».

5.4.2 Discussion sur la contrainte imposée sur la clause d'accélération

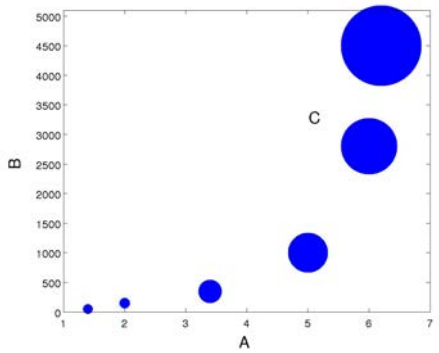
La contrainte imposant que tous les items graduels de $M_1 \cup M_2$ doivent apparaître dans M_a soit au numérateur, soit au dénominateur, est justifiée par le fait que, sans cette contrainte, on peut rencontrer deux problèmes illustrés ci-dessous sur des cas particuliers : un problème d'interprétation et l'existence d'une variante plus simple et meilleure qui vérifie la contrainte.

En effet, si on autorise le cas d'un motif de la forme $A^{\geq}B^{\geq}C^{\geq}\left(\frac{\Delta C}{\Delta A}\right)^{\geq}$ (qui ne vérifie pas la définition car il manque $\frac{\Delta C}{\Delta B}^{\geq}$ ou $\frac{\Delta B}{\Delta A}^{\geq}$ ou $\frac{\Delta A}{\Delta B}^{\geq}$), alors ce motif est traduit par la phrase « plus A augmente, plus B augmente, plus C augmente rapidement ». Toutefois, l'enrichissement est porté uniquement sur le motif graduel composé des attributs A et C, à savoir que la rapidité d'augmentation des valeurs de l'attribut C est observée uniquement par rapport à celles de A. Or d'après la phrase ci-dessus, on comprend que cette augmentation est également observée par rapport à celles de B. Ce type de motif pose donc un problème d'interprétation et de formulation linguistique.

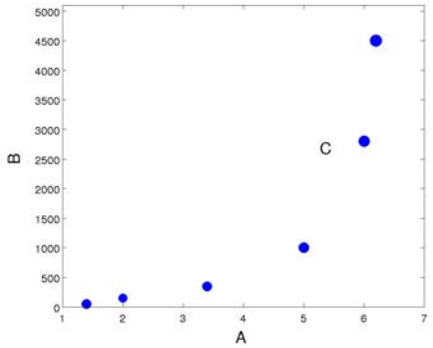
De plus, si l'information sur la rapidité d'augmentation des valeurs de l'attribut C par rapport à celles de A est observée, alors elle est extraite, dans une variante simple, à partir du motif graduel composé des deux attributs, A et C, sous la forme $A^{\geq}C^{\geq}\left(\frac{\Delta C}{\Delta A}\right)^{\geq}$, traduite par la phrase « plus A augmente, plus C augmente rapidement ». Cette formulation linguistique est sans ambiguïté et plus adaptée à l'information exprimée. La forme $A^{\geq}B^{\geq}C^{\geq}\left(\frac{\Delta C}{\Delta A}\right)^{\geq}$ est en réalité traduite par le motif graduel accéléré $A^{\geq}C^{\geq}\left(\frac{\Delta C}{\Delta A}\right)^{\geq}$ et par le motif graduel classique $A^{\geq}B^{\geq}C^{\geq}$: ces deux motifs expriment toute l'information sans ambiguïté.

L'exemple précédent montre que chaque item graduel de M_1 ou M_2 doit apparaître dans la clause d'accélération. Il faut noter que celle-ci est plus contraignante encore, car elle impose qu'il apparaît en combinaison avec tous les items graduels : chaque attribut apparaissant au numérateur, B, doit présenter un effet d'accélération pour *tous* les attributs A apparaissant au dénominateur.

Ainsi, un motif de type $A^{\geq}B^{\geq}C^{\geq}D^{\geq}\left(\frac{\Delta A}{\Delta C}\right)^{\geq}\left(\frac{\Delta B}{\Delta D}\right)^{\geq}$ n'est pas autorisé : car il manque les items accélérés $\left(\frac{\Delta A}{\Delta D}\right)^{\geq}$ et $\left(\frac{\Delta B}{\Delta C}\right)^{\geq}$. Pour cet exemple, la contrainte selon laquelle tous les items graduels doivent apparaître soit au numérateur, soit au dénominateur de la clause d'accélération est vérifiée. La contrainte qui n'est cependant pas vérifiée est le fait que chaque attribut du numérateur doit présenter un effet d'accélération pour tous les attributs apparaissant au dénominateur : si l'attribut A accélère par rapport à l'attribut C, le motif n'exprime pas d'accélération de A par rapport à D et de même l'attribut B accélère par rapport à l'attribut D, mais on n'a pas d'accélération de B par rapport à C. Un tel motif pose le même problème d'interprétation que celui de l'exemple précédent : il n'existe pas d'expression linguistique permettant de conserver toute l'information. Il peut toutefois être remplacé simplement par les deux motifs graduels accélérés $A^{\geq}C^{\geq}\left(\frac{\Delta A}{\Delta C}\right)^{\geq}$, $B^{\geq}D^{\geq}\left(\frac{\Delta B}{\Delta D}\right)^{\geq}$ et le motif graduel classique $A^{\geq}B^{\geq}C^{\geq}D^{\geq}$, qui expriment toute l'information sans ambiguïté.



(a) Clause d'accélération concernant 2 attributs : *plus A augmente, plus B augmente rapidement, plus C augmente rapidement* .



(b) Clause d'accélération concernant 1 attribut : *plus A augmente, plus C augmente, plus B augmente rapidement.*

5.4.3 Cas particuliers des motifs graduels accélérés généralisés

Un motif graduel général tel que $|M_1| = 1$ et $|M_2| \geq 2$ Correspond au cas où la clause d'accélération concerne plusieurs attributs B, dont chacun présente un effet d'accélération pour un même attribut, A, qui compose M_1. Autrement dit, en notant $M_1 = A$, $\forall B \in M_2, \left(\frac{\Delta B}{\Delta A}\right)^{\geq}$.

La clause d'accélération est donc interprétée comme une augmentation de $\left(\frac{\Delta B_1}{\Delta A}\right)$ et ... et une augmentation de $\left(\frac{\Delta B_{|M_2|}}{\Delta A}\right)$ sur un même ensemble de données.

Linguistiquement, un tel motif est interprété comme : « plus A augmente, ..., plus B_1 augmente rapidement et ... et plus B_m augmente rapidement ». Un tel motif peut par exemple être extrait des données artificielles représentées sur la figure 5.5(a), décrites par trois attributs A, B et C, respectivement indiqués par l'abscisse, l'ordonnée et la taille des points., avec $M_1 = A^{\geq}$ et $M_2 = B^{\geq}C^{\geq}$, linguistiquement exprimé comme « plus A augmente, plus B augmente rapidement, plus C augmente rapidement ». On observe en effet son support graduel accéléré est de 100% : un effet d'accélération des valeurs des attributs B et C par rapport à celles de A pour toutes les données.

Un motif graduel général tel que $|M_1| \geq 2$ et $|M_2| = 1$ Correspond au cas où la clause d'accélération concerne un seul attribut B, pour lequel un effet d'accélération est présent pour chacun des attributs composant M_1. Autrement dit, $M_2 = B$ et $\forall A \in M_1, \left(\frac{\Delta B}{\Delta A}\right)^{\geq}$.

La clause d'accélération est donc interprétée comme une augmentation de $\left(\frac{\Delta B}{\Delta A_1}\right)$ et ... et une augmentation de $\left(\frac{\Delta B}{\Delta A_{|M_1|}}\right)$ sur un même ensemble de données.

Linguistiquement, un tel motif est interprété comme dans le cas d'un motif composé de deux attributs : « plus A_1 augmente, ..., plus A_n augmente, plus B augmente rapidement ». Un tel motif peut être illustré par l'exemple « plus l'altitude est élevée, plus il fait froid, plus la pluie coule rapidement ».

Ce cas particulier peut être illustré par l'exemple du motif graduel $M_1M_2 = A^{\geq}B^{\geq}C^{\geq}$, en utilisant l'ensemble de données de la figure 5.5(b) : la figure présente ici une accélération

des valeurs du seul attribut B par rapport aux valeurs des deux autres attributs A et C. Elle représente donc le motif graduel accéléré, qui illustre ce cas particulier, « plus A augmente, plus C augmente, plus B augmente rapidement ».

5.4.4 Formulation avec fonction convexe à plusieurs variables

Mathématiquement, la notion de fonction convexe est également définie pour les fonctions à plusieurs variables, et est basée sur les propriétés de leurs matrices hessiennes.

La définition précédente s'interprète comme la convexité de toutes les fonctions f_i, telles que $B_i = f_i(A_1, ..., A_{|M_1|})$ définies sur les mêmes univers de A_i.

Comme dans le cas des motifs graduels composés de deux attributs, une discrétisation basée sur les données considérées peut être calculée pour le cas général des motifs graduels accélérés.

5.4.5 Méthodes d'extraction : a posteriori et intégrée

Nous présentons ici l'extraction de motifs graduels accélérés composés de plusieurs attributs. Nous proposons pour cela deux approches suivant les mêmes principes que ceux des approches de traitement des motifs contradictoires proposées dans le chapitre 3, page 3.4.1.

La première approche consiste à considérer l'ensemble des motifs fournis par l'algorithme GRITE, puis à chercher les clauses d'accélération à partir des motifs fournis : M étant un motif graduel fréquent, il s'agit de générer tous les motifs accélérés $M_1 M_2 \frac{\Delta M_2}{\Delta M_1}$ tels que $M_1 \subseteq M$, $M_2 \subseteq M$, $M_1 \cap M_2 = \emptyset$ et $M_1 \cup M_2 = M$. Le but de la seconde approche est d'éviter de générer des motifs graduels de longueur $k+1$ s'appuyant sur des motifs de longueur k pour lesquels aucune clause d'accélération n'a été identifiée, et qui peuvent donc être éliminés, afin de filtrer davantage en cours de génération. En d'autres termes, la génération des motifs du niveau $k+1$ s'appuie uniquement sur les motifs du niveau k dont le SG_a vérifie la condition de support minimum. Cette information est intégrée dans le processus de génération, et plus précisément dans la matrice de concordance (Di Jorio et al., 2009) : celle-ci doit représenter des chemins vérifiant les motifs graduels accélérés et non l'ensemble des chemins supports des motifs graduels fréquents.

Approche a posteriori : algorithme efficace pour la génération des clauses d'accélération

L'algorithme permettant d'extraire les motifs graduels accélérés généralisés de la forme $M_1 M_2 \frac{\Delta M_2}{\Delta M_1}$ s'applique successivement à chaque chemin maximal supportant le motif à enrichir, et cherche à identifier toutes les conjonctions entre attributs impliqués dans ce motif qui peuvent présenter un effet accélérateur pour le reste des attributs du motif. Nous considérons donc dans ce paragraphe la question de l'identification efficace de ces clauses, afin de limiter le coût de calcul des supports d'accélération. Pour cela, nous proposons une procédure exploitant la propriété d'anti-monotonie du support graduel accéléré par rapport à la longueur

des clauses d'accélération : ainsi, en notant $M_2 = B$, si l'on sait que « M_1, plus B augmente rapidement » ne vérifie pas le seuil de support graduel d'accélération, alors on peut éliminer a priori toutes les clauses d'accélération contenant l'attribut B au numérateur.

La procédure n'envisage donc pas de générer toutes les clauses d'accélération composées (qui contiennent au moins deux attributs) pour ensuite sélectionner celles qui vérifient le seuil de support graduel d'accélération. Si une clause d'accélération de longueur k ne vérifie pas le seuil de support graduel d'accélération, alors toutes les clauses de longueur supérieure à k, contenant celles de longueur k, ne le vérifient pas.

Ainsi, avec cette procédure, si B présente un effet d'accélération pour un attribut A apparaissant dans M_1 (ou pour plusieurs mais pas tous les attributs apparaissant dans M_1), alors le motif graduel accéléré « M_1, plus B augmente rapidement » ne peut donc pas être extrait, mais l'information de l'accélération de B par rapport à A est exprimée par le motif graduel accéléré « plus A augmente, plus B augmente rapidement » extrait au niveau inférieur, sous condition qu'il vérifie le seuil de support.

Approche intégrée : construction de la matrice de concordance

La construction de cette matrice est basée sur les graphes de précédence, comme proposé par Di Jorio et al. (2009) (voir section 1.2.3, page 38). Les données sont représentées dans un graphe dont les nœuds sont les couples d'objets successifs de la base de données selon la contrainte d'ordre du motif considéré, et les arcs expriment les relations de précédence induites par les attributs impliqués dans les motifs considérés. Il faut noter que ce sont des paires d'objets considérées ici, et non des objets comme dans l'algorithme de base (Di Jorio et al., 2009), car on s'intéresse aux attributs graduels non classiques de la forme $(\frac{\Delta B}{\Delta A})$. Ces attributs ne possèdent pas le même ensemble de définition que les attributs graduels classiques : mais s'appliquent à des paires d'objets successifs.

Ce graphe est représenté par sa matrice d'adjacence, sous forme d'une matrice binaire $\frac{n(n-1)}{2}$: s'il existe un ordre entre le couple d'objets (o_1, o_2) et un couple d'objets (o_2, o_3), alors le bit correspondant à la ligne (o_1, o_2) et à la colonne (o_2, o_3) vaut 1, et 0 sinon.
Autrement dit, pour un motif graduel accéléré $M = \{A_n^{*n} B_m^{*m}; \left(\frac{\Delta B_m}{\Delta A_n}\right)^{\geq}, n = 1..k, m = 1...p\}$, le coefficient correspondant à la paire de couples d'objets $((o_1, o_2), (o_3, o_4))$ vaut 1, si $\forall\, n \in [1, k]$, $\forall m \in [1, p]$, on a

$$\begin{cases} A_n(o_1) *_n A_n(o_2) *_n A_n(o_3) *_n A_n(o_4) \\ \qquad\qquad \text{ET} \\ B_m(o_1) *_m B_m(o_2) *_m B_m(o_3) *_m B_m(o_4) \\ \qquad\qquad \text{ET} \\ \left(\frac{\Delta B_m}{\Delta A_n}\right)(o_1, o_2) \leq \left(\frac{\Delta B_m}{\Delta A_n}\right)(o_4, o_3) \end{cases}$$

et vaut 0 sinon.

Le support d'un motif graduel accéléré peut ensuite être obtenu à partir de la longueur maximale des chemins obtenus. Il est à noter que cette approche ne s'intéresse pas aux chemins eux-mêmes, mais uniquement à leur longueur.

Il convient également de noter que si la complexité spatiale de la mémoire semble importante, elle est réduite de manière considérable par l'élimination des nombreuses lignes et des colonnes nulles.

La pertinence de cette approche vient, d'une part, de sa très grande efficacité pour générer des motifs graduels de longueur $l+1$ à partir de motifs de taille l, et d'autre part, de sa capacité à identifier tous les motifs graduels accélérés dont le support graduel accéléré dépasse le seuil de support minimum.

En termes de performances de consommation mémoire et de temps de calcul des deux approches, nous pouvons préciser, avant d'effectuer les expérimentations, que la consommation de mémoire est plus élevée pour l'approche intégrée que pour la méthode a posteriori. Cela est dû à la manipulation des matrices de concordance qui ont une taille plus élevée que celle des matrices de concordances manipulées pour l'extraction des motifs graduels classiques et leur sauvegarde à chaque niveau d'extraction. En revanche, le temps de calcul peut être moins long que celui de l'approche a posteriori, puisque la méthode intégrée évite de générer des motifs qui seront ensuite éliminés du fait de leurs faibles supports.

La méthode intégrée apparaît donc comme plus coûteuse en termes de consommation mémoire, mais plus rapide en termes de temps de calcul. De plus, elle peut générer plus de motifs graduels accélérés que la méthode a posteriori.

5.5 Expérimentations et résultats

Cette section décrit les expériences menées en utilisant l'approche proposée pour l'extraction des motifs graduels accélérés sur l'ensemble de données réelles météorologiques. Ces données sont les mêmes que celles données en annexe A.1, page annexe :BDExp, mais elles décrivent des observations réalisées pendant une période différente, du 15 au 22 décembre 2013. Elle contient 2164 observations.

L'analyse des résultats est basée sur le nombre de motifs graduels extraits et leur qualité.

En fixant le seuil de support à $s = 20\%$, 153 motifs graduels sont extraits (de longueur maximale 3). La figure 5.5 représente le support graduel accéléré de tous les motifs graduels identifiés. On peut observer que les motifs graduels avec SG_a inférieur à 20% ne sont pas nombreux et qu'environ 30% d'entre eux ont un SG_a supérieur à 50%. En fixant le seuil du support graduel accéléré à $s_a = 20\%$, représenté par la ligne horizontale sur la figure 5.5, 130 motifs sont considérés comme enrichis par une clause d'accélération, ce qui correspond à plus de 85%. Aucun motif graduel accéléré de longueur supérieure à 2 n'a été extrait, faute d'avoir un SG_a suffisant.

Selon la combinaison des critères avec priorité discutée dans la section 5.2.3, page 126, le motif graduel accéléré le plus intéressant est alors celui correspondant au point A sur le

graphe. Il représente le motif graduel « plus la vitesse du vent augmente, plus la distance parcourue par le vent augmente rapidement » : son SG est de 100% et son SG_a est de 90%. Il correspond à un résultat attendu de la définition proposée des motifs graduels accélérés : la relation linéaire entre ces deux attributs correspond au cas limite de l'accélération, et obtient ainsi un support accéléré élevé.

Les autres motifs graduels accélérés les plus intéressants sont alors les deux points situés dans la région B du graphe, qui correspondent respectivement aux motifs graduels

— plus la température diminue, plus la pluie accumulée augmente rapidement : $SG = 100\%$ et $SG_a = 32\%$.

— plus l'humidité diminue, plus la température augmente rapidement : $SG = 94.73\%$, $SG_a = 34\%$.

Les points centraux de la région C dans le graphe montrent un compromis entre SG et SG_a. Ils correspondent à

— plus la vitesse des rafales de vent augmente, plus la distance parcourue par le vent augmente rapidement : $SG = 54.9\%$ and $SG_a = 51\%$.

— plus la vitesse des rafales de vent augmente, plus la vitesse du vent augmente rapidement : $SG = 57.3\%$ et $SG_a = 48\%$.

Enfin, on peut observer que la majorité des motifs graduels accélérés extraits ont un support graduel légèrement au-dessus du seuil de 20%, et que beaucoup d'entre eux ont un support graduel accéléré élevé. On trouve des exemples ayant un SG_a élevé dans la région D dans le graphe, qui comprennent

— plus l'humidité augmente, plus la distance parcourue par le vent augmente rapidement : $SG = 22.69\%$ et $SG_a = 81\%$.

— plus la pression diminue, plus l'humidité augmente rapidement : $SG = 20.93\%$ et $SG_a = 88\%$.

— plus la température ressentie augmente, plus la température augmente rapidement : $SG = 22.88\%$ et $SG_a = 86\%$.

Il est également intéressant d'examiner un exemple sans effet d'accélération : le motif graduel représenté par le point E correspond à

— plus la pluie accumulée diminue, plus la température ressentie augmente rapidement : $SG = 94.72\%$ et $SG_a = 10\%$.

5.6 Conclusion

Dans ce chapitre, nous avons étudié et formalisé une nouvelle information intéressante qui résume les données : nous avons ainsi proposé d'enrichir les motifs graduels par qualification

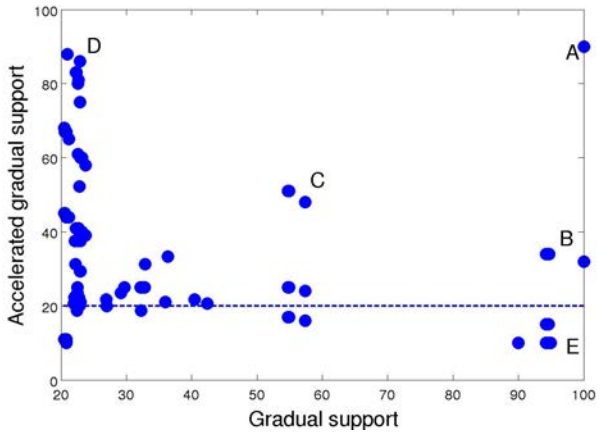


Figure 5.5 – Support graduel et support graduel accéléré, pour chacun des 153 motifs graduels extraits.

du mode de dépendance graduelle entre les valeurs d'attributs impliqués dans le motif considéré. Pour cela, nous avons proposé une nouvelle forme de motifs graduels que nous avons appelée motifs graduels accélérés.

L'extraction de ces motifs graduels accélérés repose sur l'identification d'attributs inclus dans le motif graduel considéré, dont les valeurs augmentent rapidement par rapport aux valeurs d'autres d'attributs. La contrainte est interprétée en termes de convexité et conduit à la définition d'un critère de qualité efficace permettant de quantifier la validité de la nouvelle information extraite.

6

Conclusion générale

Synthèse des travaux effectués

Dans le cadre de la fouille de données, nous nous sommes intéressées à la contextualisation et à l'enrichissement des motifs graduels. Notre objectif était d'identifier de nouveaux contextes apportant aux motifs graduels extraits précision, meilleure validité et facilité d'interprétation.

Pour répondre à cet objectif, nous avons proposé plusieurs approches pour lesquelles nous avons mis en œuvre une méthodologie en quatre étapes. Dans un premier temps, nous avons étudié et formalisé la sémantique et l'interprétation souhaitées pour les différentes formes de motifs enrichis extraits à partir des données. Nous avons ensuite proposé des mesures de qualité pour évaluer et quantifier la validité des motifs proposés. Nous avons ensuite proposé et implémenté des algorithmes efficaces d'extraction automatique des motifs qui maximisent les critères de qualité proposés. Enfin, nous avons mené une étude expérimentale, à la fois sur des données jouets pour étudier et analyser le comportement des approches proposées, et sur des données réelles pour montrer la pertinence des approches et l'intérêt des motifs extraits. Ces derniers permettent de valider l'apport des différentes formes de motifs proposées, ainsi que leur interprétation associée.

Nous avons tout d'abord mis en œuvre cette méthodologie pour une contextualisation des motifs par intégration d'attributs complémentaires, afin d'extraire *les motifs graduels renforcés*, pour lesquels on utilise une clause de renforcement composée de plusieurs attributs flous et traduite par l'expression linguistique « d'autant plus que ». Nous avons examiné la transposition de cette notion de renforcement au cas des règles d'association classiques en discutant leurs interprétations possibles, en définissant les critères de qualité qui peuvent être utilisés pour mesurer leur pertinence et en étudiant leur intérêt par rapport à des règles d'association classiques. Nous avons montré que l'information qu'elles apportent peut être exprimée par des règles classiques : le renforcement des règles d'association ne représente ainsi pas d'apport particulier, à cause de la nature identique, qui est présentielle, de la clause de renforcement et de la règle d'association.

Nous avons également étudié une contextualisation par caractérisation des motifs graduels. Celle-ci considère des restrictions définies par des contraintes d'intervalles identifiés automatiquement, et non des modalités existant dans les données. La caractérisation est exprimée linguistiquement par des clauses introduites par l'expression « surtout si ». L'approche proposée prend en compte des contraintes de densité, afin de mettre en évidence les régions appropriées du domaine qui sont fortement peuplées. Nous avons mis en œuvre les quatre étapes de la méthodologie rappelée ci-dessus pour cette contextualisation.

Les clauses d'enrichissement par renforcement et par caractérisation ont toutes les deux des sémantiques présentielles, dans la mesure où la présence de cette information supplémentaire conduit à une restriction de données sur laquelle la validité du motif doit augmenter.

Nous avons mis en œuvre la même méthodologie dans l'ajout d'une information liée à un ensemble de motifs et non à un unique motif. Il s'agit de considérer ici un point de vue global et non individuel sur les motifs, comme cela était le cas dans les approches précédentes. En effet, les problèmes de la contradiction et de l'ambiguïté de l'information extraite par un motif graduel sont induits par un ensemble de motifs. Pour le traiter, nous avons été amenées à proposer un critère de qualité contraint, qui ne dépend pas uniquement du motif considéré, mais également de ses contradicteurs potentiels. La définition du nouveau critère est basée sur un sous-ensemble d'objets propres pour chacun des motifs contradictoires, qui constitue un contexte propre au motif considéré. Ce contexte évite toute ambiguïté entre motifs validés et répond bien au problème de la lisibilité et de l'interprétation des motifs contradictoires.

Enfin, nous avons mis en œuvre notre méthodologie afin de qualifier le mode de dépendance graduelle pouvant résumer un ensemble de données. Cette information permet de capturer la façon dont les valeurs de certains attributs varient par rapport aux autres, et plus précisément de montrer l'existence d'un effet d'accélération des valeurs d'attributs. Nous avons traduit cet effet en termes de convexité, et proposé l'expression linguistique « rapidement » comme traduction de cet effet. Cette nouvelle information est extraite dans ce que nous avons appelé les *motifs graduels accélérés*.

Les expérimentations réalisées sur des données jouets et réelles ont permis de mettre en évidence la pertinence de ces différentes contextualisations de motifs graduels qui formalisent des résumés enrichis, extrayant des connaissances pertinentes et facilement exploitables par l'utilisateur.

Perspectives

Les perspectives ouvertes par ces contributions sont nombreuses. Nous les organisons ci-dessous comme suit : un premier axe de perspectives vise à approfondir les modes d'enrichissement proposés, en particulier le traitement des contradictions, la caractérisation et l'accélération, ainsi qu'à les combiner pour extraire des contextes plus riches encore. Un autre axe concerne l'amélioration des coûts de calcul, problématique centrale en fouille de données et en particulier pour l'extraction de motifs fréquents. Une troisième direction de recherche à envisager porte sur la généralisation de l'exploitation de motifs graduels.

Caractérisation basée sur le chemin propre global dans le cas des motifs contradictoires

Dans ce manuscrit, nous avons proposé une approche visant à traiter la contradiction des motifs graduels. Cette approche consiste à distinguer une plage de valeurs pour chaque motif où il est le seul valide. Cette contrainte nous a amenée à définir le support graduel propre global, qui mesure la validité des motifs non ambigus.

Il serait intéressant de chercher à exploiter l'information du chemin propre global pour la contextualisation des motifs graduels.

En effet, il est possible d'identifier un intervalle propre à chacun des deux motifs contradictoires en deux étapes : la première consiste à identifier l'attribut qui caractérise le motif, la seconde à déterminer le sous-ensemble d'objets propres qui induit un intervalle propre au motif. Ces deux étapes sont applicables aux motifs graduels de longueur 2, mais deviennent complexes pour les motifs de longueur supérieure à 2.

Une telle approche de caractérisation suit un principe essentiellement différent de la méthode basée sur des outils de morphologie mathématique présentée au chapitre 4 : elle adapte en effet un point de vue global qui traite des ensembles de motifs, et non des motifs individuels. Une étude comparative des résultats expérimentaux respectifs est à réaliser pour caractériser leurs propriétés.

Qualification de l'accélération

Une perspective liée aux motifs graduels accélérés vise à les enrichir par l'intégration d'une information sur le type d'accélération, par exemple afin de distinguer leur force : il est en effet pertinent de faire la différence entre une relation quadratique et une relation exponentielle ou encore une relation linéaire qui correspond au cas limite de l'accélération.

Une des perspectives de nos travaux vise à étudier les différents types d'accélération qu'on peut obtenir. Cela peut se réaliser en prenant en compte les valeurs du quotient qui permet d'identifier l'accélération. Cette étape permet de filtrer la qualité de l'accélération de chaque motif graduel accéléré. On peut par exemple quantifier le degré de l'accélération avec les expressions linguistiques « faible accélération », « accélération presque constante » ou « forte accélération ».

Approfondissement de la caractérisation des motifs graduels

Concernant le contexte de l'approfondissement de la contextualisation, une autre perspective vise à caractériser les motifs avec un nouvel attribut non impliqué dans le motif à caractériser, par exemple « plus la température augmente, plus la pression diminue, surtout si l'humidité appartient à $[50, 65]$ % ». Ce principe serait particulièrement intéressant dans le cas où les données comporteraient un attribut temporel ou des attributs catégoriels dérivés de la date, comme par exemple « plus la température augmente, plus l'accumulation de pluie diminue, surtout en l'été ».

Cet axe de recherche soulève le problème du temps de calcul ainsi que de l'explosion combinatoire de l'espace de recherche à explorer. Elle requiert le développement de méthodes efficaces pour éliminer les motifs contextualisés non pertinents dès leur détection.

La principale piste de recherche liée à la caractérisation des motifs graduels concerne l'approfondissement théorique de la méthode proposée.

Une première direction à prendre vise à étudier formellement les propriétés du filtre et l'effet de consolidation sur lesquels est basée notre proposition : nous avons examiné ces propriétés sur des cas précis, et il serait intéressant d'étendre cette étude au cas général, qui est complexe en raison de la récursivité du filtre alterné proposé.

Nous avons de plus proposé une interprétation des motifs graduels en tant que validité accrue. Nous avons montré expérimentalement que cette propriété était vérifiée sur des bases de données réelles. Une perspective théorique intéressante consisterait à prouver formellement cette propriété, en tenant compte de la liaison existant entre les garanties sur le support minimum et la longueur minimale des séquences obtenues. Ces critères sont liés à l'ordre du filtre utilisé et ils sont agrégés implicitement par le filtre. Deux autres perspectives peuvent être envisagées. La première consiste à pondérer le rôle de chaque critère lors de son agrégation, bien que ceci ne soit pas évident, puisqu'on peut s'interroger dans ce cas sur la manière de pondérer ces critères. Dans l'hypothèse où une pondération serait trouvée, il faudrait alors déterminer quelle valeur d'ordre du filtre serait optimale pour vérifier la pertinence de la pondération proposée. Une autre voie consisterait à orienter notre méthode vers l'optimisation directe du double critère et non vers une optimisation individuelle.

Combinaison des modes de contextualisation des motifs graduels proposés

On peut envisager différentes combinaisons des modes de contextualisation des motifs graduels proposés. Nous présentons ici celles faisant naturellement suite aux travaux effectués.

Si les données numériques sont décrites par des modalités floues, on peut combiner les deux types d'enrichissements présentiels, le renforcement et la caractérisation. On peut par exemple combiner les clauses de renforcement avec celles de caractérisation pour extraire des motifs tels que « plus on est proche du mur, plus on freine fort, d'autant plus que la vitesse est élevée et surtout si la distance au mur est dans $[0, 50]$ m ».

Une autre combinaison à mettre en évidence est celle que l'on peut effectuer avec les clauses d'accélération et le traitement de la contradiction. Cette combinaison permet de lever l'ambiguïté quand les effets d'accélération et de décélération sont présents simultanément, en garantissant une plage de valeurs propre pour chaque effet. Par exemple, lors de l'extraction simultanée des motifs « plus la température augmente, plus l'accumulation de pluie diminue rapidement » et « plus la température augmente, plus l'accumulation de pluie diminue lentement », une clause de caractérisation permettrait alors d'identifier les sous-ensembles de données où ces motifs sont respectivement observés. Ainsi, la caractérisation de ces deux motifs en fonction de la saison où chacun est observé : « plus la température augmente, plus l'accumulation de pluie diminue rapidement, en été » et « plus la température augmente, plus l'accumulation de pluie diminue lentement, en hiver », permet donc de lever l'ambiguïté

de ces deux motifs et de les rendre lisibles. Nous introduisons alors un nouveau contexte concernant une information temporelle exprimant un comportement qui se produit dans des périodes de temps spécifiques.

Une autre perspective vise à combiner les différentes approches proposées dans cette thèse. Ceci peut être envisagé de différentes manières. En effet, il serait intéressant d'extraire un motif graduel enrichi à la fois par une clause de caractérisation et une clause d'accélération, après avoir traité la contradiction comme par exemple « plus la pression diminue, plus l'humidité augmente rapidement, surtout si la pression est entre $[500, 1000]$ hPa ». Ce type de combinaison a l'avantage de résumer différentes informations d'ordre sémantique en un seul motif compréhensible et lisible. Un autre avantage que peut présenter une telle combinaison est de réduire considérablement le nombre de motifs extraits, en particulier par la gestion des motifs contradictoires, puisque plusieurs contraintes sont imposées simultanément lors de leur extraction. Ainsi, la combinaison peut être considérée comme un mode de filtrage des motifs graduels extraits.

Après avoir défini une combinaison de clauses, on pourrait appliquer la même méthodologie que la précédente. On pourrait ainsi définir une nouvelle sémantique, une formulation linguistique, ainsi que des critères de qualité permettant d'évaluer la pertinence de la combinaison des clauses. Il serait toutefois nécessaire de préciser quel enrichissement s'applique à quoi, comme par exemple la caractérisation de l'accélération ou du motif non accéléré avec la conjonction des deux enrichissements. Cependant, ceci soulève le problème de la formulation linguistique permettant d'exprimer les faits de façon intuitive à l'utilisateur, ainsi que le problème de l'évaluation : la question est de savoir si la combinaison est évaluée avec tous les critères de qualité proposés pour les clauses combinées ou alors s'il faut définir un nouveau critère pouvant évaluer la nouvelle sémantique.

Amélioration des performances

Les problèmes de stockage en mémoire et de temps nécessaire pour les calculs sont les difficultés principales de la plupart des algorithmes de fouille de données. On peut envisager différentes pistes pour améliorer les performances des algorithmes que nous avons proposés.

Une première piste consiste par exemple à intégrer les représentations condensées dans nos algorithmes. Celles-ci constituent une avancée majeure dans le cadre de l'extraction de motifs fréquents (Mannila & Toivonen, 1996; Ayouni, 2012). Il serait intéressant de les étudier dans le cadre de nos approches, puisque dans notre cas, l'étape la plus coûteuse en termes de temps et de mémoire est également la génération de motifs fréquents. Dans ce contexte, on peut imaginer de nouveaux critères de qualité basés sur la nouvelle représentation. Une question théorique qui nécessite d'être étudiée est de savoir si le même principe pourrait être adapté à l'approche d'extraction de motifs graduels accélérés que nous avons proposée, où un attribut particulier est impliqué.

On pourrait par exemple adapter à nos approches l'algorithme FP-Growth (Han et al., 2000) qui utilise une représentation condensée et évite les scans coûteux de la base de données. Ce dernier possède l'avantage d'être rapide et moins coûteux que l'algorithme Apriori

en termes de consommation mémoire. En effet, cet algorithme apporte une solution au problème de la fouille de motifs fréquents dans une grande base de données en une structure compacte. Le même questionnement à propos de l'attribut particulier impliqué dans la clause d'accélération se pose également dans ce cas-ci.

Une autre direction permettant l'optimisation en temps de nos approches vise à profiter de l'architecture des processeurs multi-cœurs et à adapter par exemple les algorithmes introduits récemment dans la littérature qui proposent des optimisations basées à la fois sur la réduction de la base de données et sur le parallélisme multi-threads. C'est le cas notamment des travaux de thèse de Negrevergne (2011) et de Quintero Flores (2013). Toutefois, ces algorithmes sont puissants en ce qui concerne le temps de calcul, mais ils ne sont pas concentrés sur la parallélisation en mémoire. Il reste donc un travail de recherche à réaliser sur la parallélisation en mémoire partagée des algorithmes. On peut penser par exemple aux architectures à mémoire distribuée partagée. L'utilisation de ce type d'architecture mémoire pourrait cependant engendrer un coût très élevé lors des communications.

Il serait également intéressant d'étudier une formulation incrémentale de l'extraction de motifs graduels et de leurs variantes enrichies, et de comparer ces performances à celles des approches parallèles.

Généralisation de l'exploitation des motifs graduels

Plusieurs thématiques étudiées dans le cadre des règles d'association sont très intéressantes en fouille de données, dans le cas général, comme par exemple la fouille visuelle et les règles rares. Des perspectives demeurent largement ouvertes dans ces thématiques. Les étendre au contexte des motifs et règles graduels constituerait une perspective à long terme.

La grande quantité de résultats à évaluer constitue un problème de la fouille de données. Leur validation par des experts demande des efforts cognitifs importants. Cette problématique a engendré l'apparition de la fouille visuelle de données, dont le but est de proposer des outils de visualisation adaptés s'appuyant par exemple sur différentes techniques de représentation, qu'elles soient textuelles ou basées sur deux ou trois dimensions (Wong et al., 1999; Lehn, 2000; Blanchard et al., 2003; Ben Yahia & Mephu Nguifo, 2004; Kuntz et al., 2006; Couturier et al., 2008). Par exemple, Kuntz et al. (2006) ont proposé un outil où les règles sont représentées sous forme de graphes : chaque règle est modélisée par un arc ayant pour origine les attributs décrivant la prémisse de la règle, et pour extrémité tous les attributs intervenant dans la règle. Ce graphe est orienté sans circuit. Leur outil est interactif et permet de filtrer efficacement les règles les plus pertinentes.

L'outil de visualisation répond à deux objectifs : faciliter l'interprétation, ce qui permet aux experts et aux utilisateurs de mieux évaluer les résultats obtenus, ou mettre en évidence d'autres informations en utilisant les motifs extraits, permettant ainsi d'enrichir les informations déjà extraites.

Ces outils visuels ont été développés principalement pour gérer la masse de règles d'association. Il serait donc intéressant de chercher à adapter ces outils à la visualisation des motifs et règles graduels.

L'adaptation des outils existants n'est cependant pas une chose aisée. Si on opte pour un mode de visualisation sous forme de graphe, il serait difficile de visualiser l'ensemble des résultats. Il convient donc de fournir en premier lieu une vue d'ensemble des résultats, qui doit permettre à l'utilisateur d'identifier les informations importantes afin qu'il puisse décider où commencer son analyse. Cet outil doit également lui permettre d'explorer seulement certaines parties qu'il juge pertinentes, et d'en obtenir des informations détaillées. En effet, il est inutile pour l'utilisateur que tous les détails soient affichés en même temps.

Cet objectif, au confluent de la fouille de données et de l'interaction homme-machine, est une perspective à long terme.

Une seconde thématique qui mérite d'être étendue au cas des motifs graduels est la recherche de motifs rares. Jusqu'à présent, les études en fouille de données se sont principalement intéressées à l'extraction de motifs fréquents. Récemment, le problème de la recherche de motifs rares ou non fréquents a été posé dans le cadre de l'extraction de règles d'association et leur intérêt a été mis en évidence pour la tache de fouille de données (Maumus et al., 2005; Szathmary et al., 2006).

Il serait donc pertinent d'étudier leur transposition au cas des motifs graduels, en formalisant la notion de motifs graduels rares, et en prenant en compte les corrélations rares qui peuvent être observées dans un jeu de données. L'identification de ces motifs peut toutefois soulever les questions suivantes. Est-il suffisant d'identifier le complément des motifs graduels fréquents pour identifier les motifs rares, ou est-il nécessaire de définir le support maximum pour leur extraction ? Dans ce cas-ci, comment pouvons-nous fixer ce support maximum ? Ces questionnements nous amènent donc à étudier le rapport exact existant entre les motifs graduels fréquents et les motifs graduels rares. Caractériser ce rapport permettrait ainsi de déterminer si les motifs graduels fréquents dérivent des motifs graduels rares et si les motifs graduels fréquents permettent de calculer les motifs graduels rares.

A

Données expérimentales

Cette annexe présente les données utilisées pour les expérimentations réalisées au cours de la thèse : la section A.1 décrit des données réelles météorologiques et la section A.2 décrit des données artificielles utilisées dans les chapitres 4 et 5

A.1 Données réelles

Nous avons utilisé une base de données réelles météorologiques appelée *météo* téléchargée à partir du site *http ://www.meteo-paris.com/ile-de-france/station-meteo-paris/pro* : ces données proviennent de la station météorologique parisienne de Saint-Germain-des-Prés.

La base de données contient 2133 observations météorologiques réalisées pendant huit jours (du 23 au 30 novembre 2012), décrites par 22 attributs numériques tels que la température (℃), la pluie accumulée (mm), l'humidité (%), la pression (hPa), la vitesse du vent (km/h), la vitesse des rafales de vent (km/h) [2], la température ressentie (℃) ou la distance parcoure par le vent (km).

La figure A.1 montre l'évolution de la vitesse du vent et la figure A.2 l'évolution de l'humidité pendant 20 heures dans la période de collecte des données.

Dans toute la thèse, nous traduisons l'attribut « wind run » par « la distance parcourue par le vent ».

Résultats obtenus par GRITE En fixant le seuil de support graduel $s = 20\%$, 835 motifs graduels sont extraits par l'algorithme GRITE.

A.2 Données artificielles

Les données des tableaux A.2 et A.3 sont utilisées pour illustrer les exemples des figures 2.3 et 2.4 du chapitre 2.

2. http ://en.wikipedia.org/wiki/Wind_run

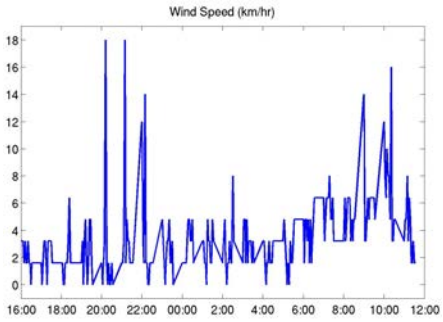


Figure A.1 – Évolution de la vitesse du vent.

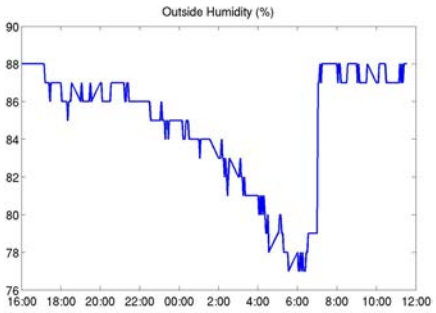


Figure A.2 – Évolution de l'humidité.

Objets \ Attribut	A	B	C
0	6.7	0.58	0.08
1	7.5	0.5	0.36
2	5.6	0.615	0
3	8.9	0.2	0.8
4	12	10	0
5	20	11	33
6	73	15	10
7	75	10	10
8	80	12	10
9	67	9	10
10	64	5	10
11	79	5	10
12	42.7	682	215
13	42	82	215
14	40	609	175.9
15	37.7	609	201
16	42.67	682	201.6
17	42.6	682	20
18	42.7	682	201.6
19	2.6	82	20
20	12	10	0
21	20	11	33
22	73	15	10
23	75	10	10
24	80	12	10
25	67	9	10
26	64	5	9
27	64	5	9
28	4.6	68	0
29	2.7	62	20
30	20	10	30

Tableau A.1 – Base de données utilisée pour illustrer le déroulement de la priorité d'ordre de traitement dans l'approche a posteriori (section 3.4.1) et la mise à jour des matrices de concordances dans l'approche intégrée (section 3.4.2).

A	B	C
0.1	0.1	0.14
0.2	0.05	0.1
0.2	0.2	0.02
0.3	0.35	0.16
0.35	15	0.12
0.4	0.45	0.05
0.5	0.5	0.14
0.6	0.7	0.05

Tableau A.2 – Données artificielles utilisées pour l'exemple de la figure 2.3, page 59.

A	B	C
0.1	0.1	1
0.2	0.05	0.1
0.2	0.2	1
0.3	0.35	0.16
0.35	15	0.92
0.4	0.45	0.92
0.5	0.5	01
0.6	0.7	1

Tableau A.3 – Données artificielles utilisées pour l'exemple de la figure 2.4, page 59.

Bibliographie

Agier, M., Petit, J.-M., & Suzuki, E. (2007). Unifying framework for rule semantics : Application to gene expression data. *Fundamental Informaticae*, *78*, 543–559.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proc. of the ACM Int. Conf. on SIGMOD*, *22*, 334–344.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo Han, A. I. (1996). Fast discovery of association rules. *Proc. of the Int. Conf. on Advances in knowledge discovery and data mining* (pp. 307–328).

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. *Proc. of the Int. Conf. on Very Large Data Sets* (pp. 487–499).

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. *Proc. of Int. Conf. on Data Engineering* (pp. 3–14).

Ahmad, A., & Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data and Knowledge Engineering*, *63*, 503–527.

Anwar, T., Beck, H., & Navathe, S. (1992). Knowledge mining by imprecise querying : a classification-based approach. *Proc. of Int. Conf on Data Engineering* (pp. 622–630).

Aumann, Y., & Lindell, Y. (2003). A statistical theory for quantitative association rules. *Journal of Intelligent Information Systems*, *20*, 255–283.

Ayouni, S. (2012). *Étude et extraction de règles graduelles floues : définition d'algorithmes efficaces*. Thèse de doctorat, Université Montpellier 2 et Université de Tunis El Manar.

Ayouni, S., Laurent, A., Ben Yahia, S., & Poncelet, P. (2010). Mining closed gradual patterns. *Artificial Intelligence and Soft Computing* (pp. 267–274).

Ayres, J., Flannick, J., Gehrke, J., & Yiu, T. (2002). Sequential pattern mining using a bitmap representation. *Proc. of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (pp. 429–435).

Azé, J., & Kodratoff, Y. (2002). Évaluation de la résistance au bruit de quelques mesures d'extraction de règles d'association. *Actes des 2èmes journées Extraction et Gestion des Connaissances* (pp. 143–154).

Bache, K., & Lichman, M. (2013). UCI machine learning repository.

Bastide, Y. (2000). *Data mining : algorithmes par niveau, techniques d'implantation et applications*. Thèse de doctorat, Université Blaise Pascal.

Bastide, Y., Taouil, R., Pasquier, N. ans Stumme, G., & Lakhal, L. (2002). Pascal : un algorithme d'extraction des motifs fréquents. *Technique et science informatiques*, *21*, 65–95.

Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., & Lakhal, L. (2000). Mining frequent patterns with counting inference. *SIGKDD Explorations*, *2*, 66–75.

Bayardo, R. J. (1998). Efficiently mining long patterns from databases. *Proc. of the ACM Int. Conf. of SIGMOD* (pp. 85–93).

Ben Yahia, S., & Jaoua, A. (2000). A top-down approach for mining fuzzy association rules. *Proc. of the Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 952–959).

Ben Yahia, S., & Mephu Nguifo, E. (2004). Emulating a cooperative behavior in a generic association rule visualization tool. *Proc. of the 16th IEEE Int. Conf. on Tools with Artificial Intelligence* (pp. 148–155).

Berry, M. J. A., & Linoff, G. S. (2004). *Data mining techniques : For marketing, sales and customer relationship management*. Wiley Computer Publishing. 2 édition.

Berzal, F., Blanco, I., Sánchez, D., Serrano, J., & Vila, M. (2005). A definition for fuzzy approximate dependencies. *Fuzzy Sets and Systems*, *149*, 105–129.

Berzal, F., Cubero, J. C., Sanchez, D., Miranda, M. A. V., & Serrano, J. M. (2007). An alternative approach to discover gradual dependencies. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 559–570.

Blanchard, J. (2005). *Un système de visualisation pour l'extraction, l'évaluation et l'exploration interactives des règles d'association*. Thèse de doctorat, École des mines de Nantes.

Blanchard, J., Guillet, F., & Briand, H. (2003). A user-driven and quality-oriented visualization for mining association rules. *Proc. of the 3rd IEEE Int. Conf. on Data Mining* (pp. 493–496).

Boros, E., Gurvich, V., Khachiyan, L., & Makino, K. (2002). On the complexity of generating maximal frequent and minimal infrequent sets. *Symposium on Theoretical Aspects of Computer Science* (pp. 133–141).

Bosc, P., Lietard, L., & Pivert, O. (1997). Gradualité, imprécision et dépendances fonctionnelles. *Bases de Données Avancées* (pp. 391–413).

Bosc, P., Pivert, O., Dubois, D., & Prade, H. (2001). On fuzzy association rules based on fuzzy cardinalities. *Proc. of the 10th IEEE Int. Conf. on Fuzzy Systems* (pp. 461–464).

Bosc, P., Pivert, O., & Ughetto, L. (1999). On data summaries based on gradual rules. *Proc. of the Int. Conf. on Computational Intelligence, Theory and Applications*.

Bouchon-Meunier, B., & Desprès, S. (1990). Acquisition numérique / symbolique de connaissances graduelles. *Proc. of the IEEE Int. Conf. on Fuzzy Systems* (pp. 127–138).

Bouchon-Meunier, B., Laurent, A., Lesot, M.-J., & Rifqi, M. (2010). Strengthening fuzzy gradual rules through "all the more" clauses. *Proc. of the IEEE Int. Conf. on Fuzzy Systems* (pp. 1–7).

Boulicaut, J., Bykowski, A., & Rigotti, C. (2003). Free-sets : A condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery*, *7*, 5–22.

Boullé, M. (2006). MODL : a Bayes optimal discretization method for continuous attributes. *Machine Learning*, *65*, 131–165.

Bouzy, B. (1995). *Modélisation congnitive du joueur de go.* Thèse de doctorat, Université Pierre et Marie Curie.

Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). *Classification and regression trees.* New York : Chapman & Hall.

Breuker, J., & Greef, P. (1993). Modelling system-user co-operation. *Proc. of the Int. Conf. on KADS : a Principled Approach to Knowledge Engineering* (pp. 47–70).

Brin, S., Motwani, R., & Silverstein, C. (1997a). Beyond market baskets : Generalizing association rules to correlations. *SIGMOD Rec.*, *26*, 265–276.

Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997b). Dynamic itemset counting and implication rules for market basket data. *SIGMOD Rec.*, *26*, 255–264.

Burdick, D., Calimlim, M., Flannick, J., Gehrke, J., & Yiu, T. (2005). MAFIA : A maximal frequent itemset algorithm. *IEEE Transactions on Knowledge and Data Engineering*, *17*, 1490–1504.

Bykowski, A., & Rigotti, C. (2001). A condensed representation to find frequent patterns. *Proc. of the 12th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (pp. 267–273).

Calders, T., & Goethals, B. (2002). Mining all non-derivable frequent itemsets. *Proc. of the Int. Conf. on Data Mining and Knowledge Discovery* (pp. 74–85).

Calders, T., & Goethals, B. (2007). Non-derivable itemset mining. *Data Mining and Knowledge Discovery*, *14*, 171–206.

Chan, K., & Au, W.-H. (1997). An effective algorithm for mining interesting quantitative association rules. *Proc. of the ACM Symposium on Applied Computing* (pp. 88–90).

Chen, G., Wei, Q., & Kerre, E. E. (2000). Fuzzy data mining : Discovery of fuzzy generalized association rules. In *Recent issues on fuzzy databases*, vol. 53, 45–66. Physica-Verlag HD.

Chiu, D.-Y., Wu, Y.-H., & Chen, A. (2004). An efficient algorithm for mining frequent sequences by a new strategy without support counting. *Proc. of the 20th Int. Conf. on Data Engineering* (pp. 375–386).

Choong, Y. W., Di Jorio, L., Laurent, A., Laurent, D., & Teisseire, M. (2009). Classification based on gradual patternss. *Proc. of the Int. Conf. on Soft Computing and Pattern Recognition* (pp. 7–12).

Chunyao, S., & Tingjian, G. (2013). Discovering and managing quantitative association rules. *Proc. of the ACM Int. Conf. on Information Knowledge Management* (pp. 2429–2434).

Coster, M., & Chermant, J. L. (1985). *Précis d'analyse d'images.* Presses du CNRS.

Couturier, O. (2005). *Contribution à la fouille de données : règles d'association et interactivité au sein d'un processus d'extraction de connaissances dans les données.* Thèse de doctorat, Université d'Artois, Lens.

Couturier, O., Dubois, V., & Hsu, T. Mephu Nguifo, E. (2008). Optimizing occlusion appearances in 3d association rule visualization. *Proc. of the IEEE Inter. Conf. on Intelligent Systems* (pp. 15–42).

Cover, T., & Hart, P. (2006). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*, 21–27.

Dârlea, G.-L. (2010). *Un système de classification supervisée à base de règles implicatives*. Thèse de doctorat, Université de Savoie.

Delgado, M., Marin, N., Sanchez, D., & Vila, M.-A. (2003). Fuzzy association rules : general model and applications. *IEEE Transactions on Fuzzy Systems*, *11*, 214–225.

Di Jorio, L., Laurent, A., & Teisseire, M. (2008). Fast extraction of gradual association rules : a heuristic based method. *Proc. of the Int. Conf. on Soft Computing as Transdisciplinary Science and Technology* (pp. 205–210).

Di Jorio, L., Laurent, A., & Teisseire, M. (2009). Mining frequent gradual itemsets from large databases. *Advances in Intelligent Data Analysis* (pp. 297–308).

Dieng, R., Corby, O., & Lapalut, S. (1993). Acquisition of gradual knowledge. In *Knowledge acquisition for knowledge-based systems*, vol. 723 of *Lecture Notes in Computer Science*, 407–426.

Dimitrov, B. N., & Rykov, V. (2004). On reliability of hierarchical systems with gradual failures. *Journal of Mathematical Sciences*, *123*, 3802–3815.

Döring, C., Borgelt, C., & Kruse, R. (2004). Fuzzy clustering of quantitative and qualitative data. *Proc. of the Annual Meeting of the North American Fuzzy Information Processing Society* (pp. 84–89).

Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. *Int. Conf. on Machine Learning* (pp. 194–202).

Drummond, I., Godo, L., & et Sandri, S. (2002). Restoring consistency in systems of fuzzy gradual rules using similarity relations. *Proc. of the 16th Brazilian Symposium on Artificial Intelligence : Advances in Artificial Intelligence* (pp. 386–396).

Dubois, D., & Prade, H. (1992). Gradual inference rules in approximate reasoning. *Proc. of the IEEE Int. Conf. on Fuzzy Systems* (pp. 103–122).

Dubois, D., & Prade, H. (1996). What are fuzzy rules and how to use them. *Fuzzy Sets and Systems*, *84*, 169–185.

Dubois, D., & Prade, H. (2008). Gradual elements in a fuzzy set. *Soft Computing*, *12*, 165–175.

Dubois, D., Prade, H., & Grabisch, M. (1995). Gradual rules and the approximation of control laws. *Theoretical aspects of fuzzy control* (pp. 147–181).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). *Advances in knowledge discovery and data mining*, chapter From Data Mining to Knowledge Discovery : An Overview, 1–34. American Association for Artificial Intelligence.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). Data mining : Concepts and techniques. *Morgan Kaufmann Publishers* (pp. 37–54).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996c). Knowledge discovery and data mining : Towards a unifying framework. *AAAI Press* (pp. 82–88).

Fayyad, U., Weir, N., & Djorgovski, S. G. (1993). SKICAT : A machine learning system for automated cataloging of large scale sky surveys. *Proc. of the Int. Conf. on Machine Learning* (pp. 112–119).

Feyyad, U. (1996). Data mining and knowledge discovery : making sense out of data. *IEEE Expert*, *11*, 20–25.

Fiot, C., Laurent, A., & Teisseire, M. (2007). From crispness to fuzziness : Three algorithms for soft sequential pattern mining. *Proc. of the IEEE Int. Conf. on Fuzzy Systems*, *15*, 1263–1277.

Fiot, C., Masseglia, F., Laurent, A., & Teisseire, M. (2008). Gradual trends in fuzzy sequential patterns. *Proc. of the Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 1–8).

Fiot, C., Masseglia, F., Laurent, A., & Teisseire, M. (2009). TED and EVA : expressing temporal tendencies among quantitative variables using fuzzy sequential patterns. *Proc. of the IEEE Int. Conf. on Fuzzy Systems*.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases - an overview. *Artificial Intelligence Magazine*, *13*, 57–70.

Fukuda, T., Morimoto, Y., Morishita, S., & Tokuyama, T. (1996a). Data mining using tow-dimensional optimized association rules : Scheme, algorithm and visualisation. *Proc. of the ACM Int. Conf. on SIGMOD* (pp. 12–23).

Fukuda, T., Morimoto, Y., Morishita, S., & Tokuyama, T. (1996b). Mining optimized association rules for numeric attributes. *Proc. of the ACM Int. Conf. on SIGACT-SIGMOD-SIGART* (pp. 12–23).

Galichet, S., Dubois, D., & Prade, H. (2003). Fuzzy interpolation and level 2 gradual rules. *Proc. of the Int. Conf. of the European Society for Fuzzy Logic and Technology* (pp. 506–511).

Galichet, S., Dubois, D., & Prade, H. (2004). Imprecise specification of ill-known functions using gradual rules. *International Journal of Approximate Reasoning*, *35*, 205–222.

Ganter, B., & Wille, R. (1997). *Formal concept analysis : Mathematical foundations*. Springer-Verlag New York, Inc.

Gardarin, G., Pucheral, P., & Wu, F. (1998). Bitmap based algorithms for mining association rules. *14èmes journées Bases de données Avancées* (pp. 157–175).

Gasmi, G., Ben Yahia, S., Mephu Nguifo, E., & Slimani, Y. (2006). IGB : une nouvelle base générique informative des règles d'association. *Information - Interaction - Intelligence(I3)*, *6*, 31–67.

Geerts, F., Goethals, B., & Bussche, J. V. D. (2005). Tight upper bounds on the number of candidate patterns. *ACM Transactions on Database Systems*, *30*, 333–363.

Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining : A survey. *ACM Computing Surveys*, *38*, 1–31.

Goethals, B. (2007). Fimi repository website http ://fimi.cs.helsinki.fi/.

Goodman, R., & Smyth, P. (1988). Information theoretic rule-induction. *Proc. of the European Conference on Artificial Intelligence* (pp. 357–362).

Gouda, K., & Zaki, M. J. (2001). Efficiently mining maximal frequent itemsets. *Proc. of the IEEE Int. Conf. on Data Mining* (pp. 163–170).

Grahne, G., & Zhu, J. (2003). Efficiently using prefix-trees in mining frequent itemsets. *Proc. of the Int. Workshop on Frequent Itemset Mining Implementations.*

Gras, R. (1979). *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques.* Thèse de doctorat, Université Rennes 1.

Gunopulos, G., Mannila, H., & Saluja, S. (1997). Discouvering all most specific sentences by randomized algorithms. *Proc. of the Int. Conf. on Database Theory* (pp. 215–229).

Han, J., Cai, Y., & Cercone, Y. (1992). Knowledge discovery in databases : An attribute-oriented approach. *Proc. of the Int. Conf. on Very Large Databases* (pp. 547–559).

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *Proc. of the ACM Int. Conf. on SIGMOD* (pp. 1–12).

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining.* Cambridge, MA, USA : MIT Press.

Hand, D., Mannila, H., & Smyth, P. (2005). *Data mining : Concepts and techniques.* San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

Hilderman, J. R., & Hamilton, H. J. (2001). Evaluation of interestingness measures for ranking discovered knowledge. *Lecture Notes in Computer Science* (pp. 247–259).

Hipp, J., Güntzer, U., & Nakhaeizadeh, G. (2000). Algorithms for association rule mining - a general survey and comparison. *SIGKDD Exploration Newsletter*, *2*, 58–64.

Holsheimer, M., Kersten, M., Mannila, H., & Toivonen, H. (1995). A perspective on databases and data mining. *Proc. of the Int. Conf. on Knowledge Discovery and Data Mining* (pp. 150–155).

Hong, T.-P., Lin, K.-Y., & Wang, S.-L. (2003). Fuzzy data mining for interesting generalized association rules. *Fuzzy Sets Syst.*, *138*, 255–269.

Houtsma, M., & Swami, A. (1993). Set-oriented mining of association rules. *Research Report RJ 9567, IBM Almaden Research Center.*

Hüllermeier, E. (2001). Implication-based fuzzy association rules. *Principles of Data Mining and Knowledge Discovery* (pp. 241–252).

Hüllermeier, E. (2002). Association rules for expressing gradual dependencies. *Proc. of the Int. Conf. on Principles of Data Mining and Knowledge Discovery* (pp. 200–211).

Hüllermeier, E. (2007). *Why fuzzy set theory is useful in data mining*, chapter 1, 1–16. Successes and New Directions in Data Mining.

Huynh, X. H., Guillet, F., & Briand, H. (2005). ARQAT : An exploratory analysis tool for interstingness measures. *Proc. of Int. Conf. on Applied Stochastic Models and Data Analysis* (pp. 334–344).

Kendall, M. G., & Babington, S. (1939). The problem of m rankings. *The Annals of Mathematical Statistics*, *10*, 275–287.

Kerber, R. (1992). ChiMerge : Discretization of numeric attributes. *Proc. of the 10th National Conference on Artificial Intelligence* (pp. 123–128).

Koperski, K., & Han, J. (1995). Discovery of spatial association rules in geographic information databases. *Proc. of 4th Int. Symp. on Large Spatial Databases* (pp. 47–66).

Kuntz, P., Lehn, R., Guillet, F., & Pinaud, B. (2006). Découverte interactive de règles d'association via une interface visuelle. In *4ème atelier visualisation et extraction de connaissances (EGC)*, vol. RNTI-E-7, 113–125. Cépaduès.

Kuok, C. M., Fu, A., & Wong, M. H. (1998). Mining fuzzy association rules in databases. *SIGMOD Rec.*, *27*, 41–46.

Lallich, S. (2002). Mesure et validation en extraction des connaissances à partir des données. *Habilitation à diriger les recherches, Université Lyon 2*.

Lallich, S., & Teytaud, O. (2004). Evaluation et validation de l'intérêt des règles d'association. *Revue des Nouvelles Technologies de l'Information*, *1*, 193–218.

Lallich, S., Teytaud, O., & Prudhomme, E. (2007). Association rule interestingness : Measure and statistical validation. *Proc. of the Int. Conf. on Quality Measures in Data Mining* (pp. 251–275).

Laurent, A., Lesot, M.-J., & Rifqi, M. (2009). GRAANK : Exploiting rank correlations for extracting gradual itemsets. *Proc. of the Int. Conf. on FQAS* (pp. 382–393).

Laurent, A., Negrevergne, B., Sicard, N., & Termier, A. (2010). PGP-MC : Towards a multicore parallel approach for mining gradual patterns. *Proc. of the Int. Conf. on Database Systems for Advanced Publications* (pp. 78–84).

Lehn, R. (2000). *Un système interactif de visualisation et de fouille de règles pour l'extraction connaissances dans les bases de données*. Thèse de doctorat, Université de Nantes.

Lenca, P., Meyer, P., Vaillant, B., & Picouet, P. (2003). Aide multicritère à la décision pour évaluer les indices de qualité des connaissances. *Actes des 3ièmes journées Extraction et Gestion des Connaissances* (pp. 271–282).

Lenca, P., Meyer, P., Vaillant, B., Picouet, P., & Lallich, S. (2004). Evaluation et analyse multicritères des mesures de qualité des règles d'association. *Revue des Nouvelles Technologies de l'Information*, *1*, 219–246.

Lenca, P., Vaillant, B., Meyer, P., & Lallich, S. (2007). Association rule interestingness measures : Experimental and theoretical studies. *Proc. of the Int. Conf. on Quality Measures in Data Mining* (pp. 51–76).

Lent, B., Swami, A. N., & Widom, J. (1997). Clustering association rules. *Proc. of the Int. Conf. on Data Engineering* (pp. 220–231).

Lin, D., & Kedem, Z. M. (1998). PINCER-SEARCH : A new algorithm for discovering the maximum frequent sets. *Proc. of the Int. Conf. on Extending Database Technology* (pp. 105–119).

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs : General and Applied*, *61*, 1–49.

Lubinsky, D. J. (1989). Discovery from databases : A review of AI and statistical techniques. *Proc. of the Int. Workshop on Knowledge Discovery in Databases* (pp. 204–218).

Mannila, H., & Toivonen, H. (1996). Multiple uses of frequent sets and condensed representations (extended abstract). *Proc. of the Int. Conf. on KDD* (pp. 189–194).

Mannila, H., & Toivonen, H. (1997). Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, *1*, 241–258.

Mannila, H., Toivonen, H., & Verkamo, A. I. (1994). Efficient algorithms for discovering association rules. *Proc. of the Int. Workshop on Knowledge Discovery in Databases* (pp. 181–192).

Marsala, C., & Bouchon-Meunier, B. (1996). Fuzzy partitioning using mathematical morphology in a learning scheme. *Proc. of the IEEE Int. Conf. on Fuzzy Systems* (pp. 1512–1517).

Masseglia, F. (2002). *Algorithmes et applications pour l'extraction de motifs séquentiels dans le domaine de la fouille de données : de l'incrémental au temps réel.* Thèse de doctorat, Université de Versailles Saint-Quentin-en-Yvelines.

Masseglia, F., Cathala, F., & Poncelet, P. (1998). The PSP approach for mining sequential patterns. *Principles of Data Mining and Knowledge Discovery* (pp. 176–184).

Masseglia, F., Poncelet, P., & Teisseire, M. (2004). Pre-processing time constraints for efficiently mining generalized sequential patterns. *Proc. of the IEEE Int. Conf. Fuzzy Systems* (pp. 87–95).

Mata, J., Alvarez, J. L., & Riquelme, J. C. (2002). An evolutionary algorithm to discover numeric association rules. *Proc. of the ACM Int. Conf. on Symposium on Applied computing* (pp. 590–594).

Maumus, S., Napoli, A., Szathmary, L., & Visvikis-Siest, S. (2005). Exploitation des données de la cohorte STANISLAS par des techniques de fouille de données numériques et symboliques utilisées seules ou en combinaison. *Atelier Fouille de Données Complexes dans un Processus d'Extraction des Connaissances* (pp. 73–76).

Michalski, R. S., Kerschberg, L., Kaufman, K. A., & Ribeiro, J. S. (1992). Mining for knowledge in databases : The inlen architecture, initial implementation and first results. *Jour. of Intelligent Information Systems*, *1*, 85–113.

Miller, R. J., & Yang, Y. (1997). Association rules over interval data. *Proc. of the Int. Conf. on Management of Data* (pp. 452–461).

Miyazaki, J., Akutsu, S., Satow, N., Hirao, C., & Yao, Y. (2001). The gradual expression of troponin t isoforms in chicken wing muscles. *Journal of Muscle Research and Cell Motility*, *22*, 693–701.

Molina, C., Serrano, J., Sanchez, D., & Vila, M. (2007). Measuring variation strength in gradual dependencies. *Proc. of the Int. Conf. of the European Society for Fuzzy Logic and Technology* (pp. 337–344).

Negrevergne, B. (2011). *A generic and parallel pattern mining algorithm for multi-core architectures.* Thèse de doctorat, Université de Grenoble.

Negrevergne, B., Termier, A., Rousset, M.-C., & Mehaut, J.-F. (2014). Para-Miner : a generic pattern mining algorithm for multi-core architectures. *Data Mining and Knowledge Discovery*, *28*, 593–633.

Nortet, C., Salleb, A., Turmeaux, T., & Vrain, C. (2006). Mining quantitative association rules in a atherosclerosis dataset. *Proc. of the Int. Conf. on PKDD* (pp. 495–506).

Oudni, A., Lesot, M.-J., & Rifqi, M. (2012). Gestion de la contradiction dans l'extraction de motifs graduels. *Actes de la conférence LFA* (pp. 218–225).

Oudni, A., Lesot, M.-J., & Rifqi, M. (2013a). Caractérisation de motifs graduels par des clauses du type « surtout si ». *Actes de la conférence LFA* (pp. 251–258).

Oudni, A., Lesot, M.-J., & Rifqi, M. (2013b). Characterisation of gradual itemsets through « especially if » clauses based on mathematical morphology tools. *Proc. of the Int. Conf. of the European Society for Fuzzy Logic and Technology* (pp. 826–833).

Oudni, A., Lesot, M.-J., & Rifqi, M. (2013c). Processing contradiction in gradual itemset extraction. *Proc. of the IEEE Int. Conf. on Fuzzy Systems* (pp. 1–8).

Oudni, A., Lesot, M.-J., & Rifqi, M. (2014). Accelerating effect of attribute variations : Accelerated gradual itemsets extraction. *Proc. of the Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems*.

Özden, B., Ramaswamy, S., & Silberschatz, A. (1998). Cyclic association rules. *Proc. of Int. Conf. on Data Engineering* (pp. 412–421).

Park, J. S., Chen, M.-S., & Yu, P. (1995). An efficient hash based algorithm for mining associations rules. *Proc. of the ACM Int. Conf. of SIGMOD* (pp. 175–186).

Pasquier, N. (2000). Mining association rules using formal concept analysis. *Proc. of the Int. Conf on Conceptual Structures* (pp. 259–264).

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1998). Pruning closed itemset lattices for association rules. *Actes des 14èmes journées Bases de données avancées* (pp. 177–196).

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999a). Closed set based discovery of small covers for association rules. *Actes des 14èmes journées Bases de données avancées* (pp. 361–381).

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999b). Discovering frequent closed itemsets for association rules. *Proc. of the Int. Conf. on Database Theory* (pp. 398–416).

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999c). Efficient mining of association rules using closed itemset lattices. *Journal Information Systems*, *24*, 25–46.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems : Networks of plausible inference*. Morgan Kaufmann Publishers Inc.

Pei, J., Han, J., & Mao, R. (2000). Closet : An efficient algorithm for mining frequent closed itemsets. *Proc. of the ACM Int. Workshop on Research Issues in Data Mining and Knowledge Discovery* (pp. 21–30).

Pei, J., Han, J., Member, S., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., & Hsu, M. (2004). Mining sequential patterns by Pattern-Growth : The prefixspan approach. *IEEE Transactions on Knowledge and Data Engineering*, *16*, 1424–1440.

Piateski, G., & Frawley, W. (1991). *Knowledge discovery in databases.* Cambridge, MA, USA : MIT Press.

Piatetsky-Shapiro, G. (1991). Discovery, analysis and presentation of strong rules. *Proc. of the Int. Conf. on Knowledge Discovery in Databases* (pp. 229–248).

Quinlan, J. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106.

Quinlan, J. R. (1993). *C4.5 : Programs for machine learning.* Morgan Kaufmann Publishers Inc.

Quintero Flores, M. P. (2013). *Fouille de motifs graduels flous basée sur architectures multi-cœur.* Thèse de doctorat, Université de Montpellier.

Rabaséda-Loudcher, S., Sebba, M., & Rakotomalala, R. (1995). Discretization method of continuous attributes : a survey of methods. *Proc. of 2nd annual Joint Conference on Information Sciences* (pp. 164–166).

Rastogi, R., & Shim, K. (1999). Mining optimized support rules for numeric attributes. *Proc. of the IEEE Int. Conf. on Data Engineering Computer Society* (pp. 206–215).

Salleb-Aouissi, A., Vrain, C., Nortet, C., Kong, X., Rathod, V., & Cassard, D. (2013). Quant-miner for mining quantitative association rules. *Journal of Machine Learning Research*, *14*, 3153–3157.

Savasere, A., Omiecinski, E., & Navathe, S. (1995). An efficient algorithm for mining association rules in large databases. *Proc. of the Int. Conf. VLDB* (pp. 432–444).

Savasere, A., Omiecinski, E., & Navathe, S. B. (1998). Mining for strong negative associations in a large database of customer transactions. *Proc. of the 4th Int. Conf. on Data Engineering* (pp. 494–502).

Sebag, M., & Schoenauer, M. (1988). Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. *Proc. of the Int. Conf. of the European Knowledge Acquisistion Workshop* (pp. 1–28).

Serra, J. (1988). *Image analysis and mathematical morphology.* Theoretical Advances. Academic Press.

Simoudis, E. (1996). Reality check for data mining. *IEEE Expert*, *11*, 26–33.

Srikant, R., & Agrawal, R. (1996). Mining quantitative association rules in large relational tables. *SIGMOD Rec.*, *25*, 1–12.

Stonebraker, M., Agrawal, R., Dayal, U., Neuhold, E. J., & Reuter, A. (1993). DBMS research at a crossroads : The vienna update. *Proc. of the 19th Int. Conf. on Very Large Data Bases* (pp. 688–692).

Stumme, G., Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (2002). Computing iceberg concept lattices with titanic. *Data and Knowledge Engineering*, *42*, 189–222.

Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., & Lakhal, L. (2000). Fast computation of concept lattices using data mining techniques. *Proc. of the Int. Workshop on Knowledge Representation Meets Databases* (pp. 129–139).

Szathmary, L., Maumus, S., Pierre, P., Toussaint, Y., & Napoli, A. (2006). Vers l'extraction de motifs rares. *Actes des 6ièmes journées Extraction et gestion des connaissances* (pp. 499–510).

Szathmary, L., Valtchev, P., Napoli, A., & Godin, R. (2009). Efficient vertical mining of frequent closures and generators. In *Advances in intelligent data analysis viii*, vol. 5772, 393–404. Springer Berlin Heidelberg.

Toivonen, H. (1994). Sampling large databases for association rules. *Proc. of the Int. Conf. on VLDB* (pp. 134–145).

Uno, T., Asai, T., Uchida, Y., & Arimura, H. (2003). LCM : an efficient algorithm for enumerating frequent closed item sets. *Proc. of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations.*

Uno, T., Kiyomi, M., & Arimura, H. (2004). LCM ver. 2 : efficient mining algorithms for frequent/closed/maximal itemsets. *Proc. of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations.*

Uno, T., Kiyomi, M., & Arimura, H. (2005). LCM ver.3 : Collaboration of array, bitmap and prefix tree for frequent itemset mining. *Proc. of the Int. Conf. on Open Source Data Mining : Frequent Pattern Mining Implementations* (pp. 77–86).

Wang, J., Han, J., & Pei, J. (2003). Closet+ : searching for the best strategies for mining frequent closed itemsets. *Proc. of the ACM Int. Conf. on SIGKDD* (pp. 236–245).

Wang, K., Tay, S. H. W., & Liu, B. (1998). Interestingness-based interval merger for numeric association rules. *Knowledge Discovery and Data Mining* (pp. 121–128).

Webb, G. I. (2001). Discovering associations with numeric variables. *Proc. of the ACM Int. Conf. on SIGKDD* (pp. 383–388).

Wille, R. (2009). Restructuring lattice theory : An approach based on hierarchies of concepts. In *Formal concept analysis*, vol. 5548, 314–339. Springer Berlin Heidelberg.

Wong, P., Whitney, P., & Thomas, J. (1999). Visualizing association rules for text mining. *Proc. of the IEEE Symposium on Information Visualization* (pp. 120–127).

Zaki, M. J. (1998). *Scalable data mining for rules.* Thèse de doctorat, University of Rochester.

Zaki, M. J. (2001). SPADE : An efficient algorithm for mining frequent sequences. *Machine Learning Journal, special issue on Unsupervised Learning*, *42*, 31–60.

Zaki, M. J., & Hsiao, C.-J. (2002). CHARM : An efficient algorithm for closed itemset mining. *SIAM Int. Conf. on Data Mining* (pp. 33–43).

Zaki, M. J., & Hsiao, C.-J. (2005). Efficient algorithms for mining closed itemsets and their lattice structure. *IEEE Transactions on Knowledge and Data Engineering*, *17*, 462–478.

Zaki, M. J., Parthasarathy, S., Ogihara, M., & Li, W. (1997). New algorithms for fast discovery of association rules. *Proc. of the Int. Conf. on KDD* (pp. 283–286).

Zeidler, J., & Schlosser, M. (1996). Continuous-valued attributes in fuzzy decision trees. *Proc. of Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 395–400).

Zhang, T. (2000). Association rules. *Prroc. of the Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 245–256).

Zhang, Z., Lu, Y., & B., Z. (1997). *An effective partitioning-combining algorithm for discovering quantitative association rules.* Proc. of the Int. Conf. on Knowledge Discovery and Data Mining.