Représentation condensée de motifs émergents

Arnaud Soulet, Bruno Crémilleux, François Rioult

GREYC, CNRS - UMR 6072, Université de Caen
Campus Côte de Nacre
14032 Caen Cedex France
{Prenom.Nom}@info.unicaen.fr

Résumé. Les motifs émergents sont des associations de caractéristiques fortement présentes dans une classe et rares dans les autres. Ils font ressortir les distinctions entre classes et se révèlent particulièrement efficaces pour construire des classifieurs et apporter une aide au diagnostic. À cause de la forte combinatoire du problème, la recherche et la représentation des motifs émergents restent des tâches complexes pour de grandes bases de données. Nous proposons ici une représentation condensée exacte des motifs émergents (i.e., les motifs *et* leurs taux de croissance sont directement obtenus depuis la représentation condensée). L'idée principale est de s'appuyer sur les récents résultats relatifs aux représentations condensées de motifs fermés fréquents. À partir de cette représentation, nous donnons aussi une méthode aisée à mettre en oeuvre pour obtenir les motifs émergents ayant les meilleurs taux de croissance. Ces motifs, appelés motifs émergents forts, ont été exploités avec succès dans une collaboration avec la société Philips.

Mots clés : motifs émergents, représentations condensées, motifs fermés, caractérisation de classes.

1 Introduction

La caractérisation de classes et la classification sont d'importants domaines de recherche en fouille de données et apprentissage. Initialement introduits dans [Dong et Li, 1999], les motifs émergents sont des motifs dont la fréquence varie fortement entre deux classes. Ils caractérisent les classes de manière quantitative et qualitative. De par leur capacité à faire ressortir les distinctions entre classes, les motifs émergents permettent de construire des classifieurs ou de proposer une aide au diagnostic. Ils sont à l'origine de travaux variés et ils sont, entre autres, utilisés dans la réalisation de classifieurs performants [Dong *et al.*, 1999, Li *et al.*, 2000]. Dans un cadre plus applicatif, on peut citer différents travaux sur la caractérisation de propriétés biochimiques ou de données médicales [Li et Wong, 2001].

La recherche de tous les motifs émergents dans les grandes bases de données est une tâche difficile car le nombre de motifs candidats est très élevé. Nous verrons à la section 2.2 que les élagages utilisés par les algorithmes par niveaux [Mannila et Toivonen, 1997] souvent utilisés en fouille de données sont inadaptés. Les méthodes les plus classiques utilisent des manipulations de bordures [Dong et Li, 1999]. Sous un angle plus général,

le problème peut être vu comme la recherche des motifs vérifiant la conjonction d'une contrainte anti-monotone et d'une contrainte monotone [De Raedt *et al.*, 2002].

Dans cet article, nous nous intéressons à l'extraction et à la caractérisation des motifs émergents dans des grands jeux de données et nous proposons une nouvelle méthode de production de motifs émergents. L'originalité de notre démarche consiste à s'appuyer sur les récents progrès sur les repré-sentations condensées de motifs [Pasquier *et al.*, 1999, Boulicaut *et al.*, 2003]. À l'aide de ce nouveau regard, cet article propose une triple contribution à la problématique de l'extraction et de la caractérisation des motifs émergents. Premièrement, nous mettons en évidence une nouvelle propriété caractérisant des motifs émergents particuliers, les jumping emerging patterns qui font l'objet d'actives recherches. Deuxièmement, nous proposons une représentation condensée exacte des motifs émergents d'une base de données. Contrairement aux techniques par manipulation de bordures qui fournissent une minoration du taux de croissance des motifs émergents, cette représentation condensée permet de connaître facilement et avec exactitude le taux de croissance de chaque motif émergent. De plus, il existe des algorithmes pour extraire efficacement celle-ci. Enfin, nous proposons une méthode qui permet d'extraire aisément les motifs émergents ayant les meilleurs taux de croissance possibles (nous les appelons "motifs émergents forts"). En effet, ce travail est aussi motivé par des demandes d'applications réelles et on constate en pratique la présence d'un grand nombre de motifs émergents. Les motifs émergents forts sont particulièrement utiles pour présenter des résultats plus synthétiques et plus exploitables aux fournisseurs des données. Nous donnons ici les résultats obtenus grâce à l'utilisation des motifs émergents forts pour la détection de lots défectueux de plaques de silicium.

L'article est organisé de la manière suivante. La section 2 introduit les notations nécessaires et traite des travaux relatifs aux motifs émergents. La section 3 définit une nouvelle caractérisation des jumping emerging patterns. Puis, elle détermine une représentation condensée exacte des motifs émergents et propose les motifs émergents forts qui sont aisément obtenus à partir de la représentation condensée. Enfin, la section 4 présente une utilisation, avec succès, des motifs émergents forts dans le cadre d'une collaboration que nous menons avec la société Philips.

2 Contexte et travaux relatifs

2.1 Notations et définitions

Soit $\mathcal{D}$ un jeu de données, comme par exemple celui présenté par le tableau 1 qui est un extrait des données que nous utilisons pour la recherche de défaillance dans une chaîne de production (cf. section 4). Ce tableau (qui est une simplification de la réalité du problème) sert d'exemple élémentaire pour présenter les différentes notions de cet article.

Chaque ligne (ou *transaction*) du tableau 1 représente un lot (noté $L_1, \ldots, L_8$) décrit par des caractéristiques (ou *items*) : $A, \ldots, E$ codent le cheminement du lot au sein de la chaîne de production et C_1, C_2 les valeurs de classes. $\mathcal{D}$ est ici partitionné en deux jeux de données $\mathcal{D}_1$ et $\mathcal{D}_2$ ($\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$), $\mathcal{D}_1$ contenant les lots dont le rendement est normal et $\mathcal{D}_2$ les lots défectueux. Les transactions possédant l'item C_1 (resp. C_2)

$\mathcal{D}$

Lot	Items						
L_1	C_1		A	B	C	D	
L_2	C_1		A	B	C	D	
L_3	C_1		A	B	C		
L_4	C_1		A			D	E
L_5		C_2	A	B	C		
L_6		C_2		B	C	D	E
L_7		C_2		B	C		E
L_8		C_2		B			E

TAB. 1 – Exemple d'une base transactionnelle

appartiennent à $\mathcal{D}_1$ (resp. $\mathcal{D}_2$). Un *motif* est un ensemble d'items quelconques (e.g., $\{A, B, C\}$ noté sous forme de chaîne ABC). Une transaction t contient le motif X si et seulement si $X \subseteq t$. Enfin, $|\mathcal{D}|$ (où $|.|$ dénote la cardinalité d'un ensemble) est le nombre de transactions de $\mathcal{D}$.

La notion de motif émergent est liée à celle de fréquence. La fréquence d'un motif X dans un jeu de données $\mathcal{D}$ (notée $\mathcal{F}(X, \mathcal{D})$) est le nombre de transactions de $\mathcal{D}$ contenant X (par exemple, $\mathcal{F}(ABC, \mathcal{D}) = 4$). À partir de cette fréquence absolue, on peut calculer la fréquence relative qui est $\mathcal{F}(X, \mathcal{D})/|\mathcal{D}|$. Dans la suite, sauf mention contraire, toutes les fréquences considérées sont absolues. Notons que par définition des bases partielles $\mathcal{D}_i$ associées aux identifiants de classe C_i, nous avons la relation $\mathcal{F}(X, \mathcal{D}_i) = \mathcal{F}(XC_i, \mathcal{D})$.

Intuitivement, un motif émergent est un motif dont la fréquence varie fortement entre deux classes. La capture du contraste entre les classes apporté par un motif se mesure par son taux de croissance. Le *taux de croissance* d'un motif X de $\mathcal{D}_2$ dans $\mathcal{D}_1$, noté $GR_1(X)$, est défini par :

$$\begin{cases} 0, & \text{si } \mathcal{F}(X, \mathcal{D}_1) = 0 \text{ et } \mathcal{F}(X, \mathcal{D}_2) = 0 \\ \infty, & \text{si } \mathcal{F}(X, \mathcal{D}_1) \neq 0 \text{ et } \mathcal{F}(X, \mathcal{D}_2) = 0 \\ \frac{|\mathcal{D}_2| \times \mathcal{F}(X, \mathcal{D}_1)}{|\mathcal{D}_1| \times \mathcal{F}(X, \mathcal{D}_2)}, & \text{sinon} \end{cases}$$

On définit ainsi de façon formelle un motif émergent (ou *emerging pattern*, noté *EP* en abrégé) :

Définition 1 (motif émergent) *Pour un seuil $\rho > 1$, un motif X est appelé* motif émergent *de $\mathcal{D}_2$ dans $\mathcal{D}_1$ si $GR_1(X) \geq \rho$.*

Donnons quelques exemples à partir des données du tableau 1. Pour $\rho = 3$, les motifs A, ABC et $ABCD$ sont des EPs de $\mathcal{D}_2$ dans $\mathcal{D}_1$. En effet, $GR_1(A) = 4/1 = 4$, $GR_1(ABC) = 3/1 = 3$ et $GR_1(ABCD) = 2/0 = \infty$. A l'inverse, le motif BCD n'est pas un EP car $GR_1(BCD) = 2/1 = 2$. Lorsque le motif X est absent de la classe $\mathcal{D}_2$ (i.e. $\mathcal{F}(X, \mathcal{D}_2) = 0$), on a alors $GR_1(X) = \infty$ et un tel motif est appelé *jumping emerging pattern* (JEP). Dans l'exemple, le motif $ABCD$ est un JEP de $\mathcal{D}_1$. De même, DE est un JEP de $\mathcal{D}_2$.

On constate que $\mathcal{D}_2 = \mathcal{D} \backslash \mathcal{D}_1$. En particulier, dans la définition du taux de croissance, $\mathcal{F}(X, \mathcal{D}_2)$ est équivalente à $\mathcal{F}(X, \mathcal{D}) - \mathcal{F}(X, \mathcal{D}_1)$ (de même, $|\mathcal{D}_2| = |\mathcal{D}| - |\mathcal{D}_1|$). À partir de cette observation, la généralisation de la définition des motifs émergents aux problèmes possédant plus de deux classes est immédiate. Soit $\mathcal{D}$ un jeu de données partitionné en k parties notées $\mathcal{D}_1, \ldots, \mathcal{D}_k$ (i.e. $\mathcal{D} = \bigcup_i \mathcal{D}_i$). Les items $C_1, \ldots, C_k$ désignent respectivement l'appartenance d'une transaction à une base $\mathcal{D}_1, \ldots, \mathcal{D}_k$. $\forall i \in \{1, \ldots, k\}$, le taux de croissance de $\mathcal{D} \backslash \mathcal{D}_i$ dans $\mathcal{D}_i$ est :

$$GR_i(X) = \underbrace{\frac{|\mathcal{D}| - |\mathcal{D}_i|}{|\mathcal{D}_i|}}_{noté\ \alpha_i} \times \frac{\mathcal{F}(X, \mathcal{D}_i)}{\mathcal{F}(X, \mathcal{D}) - \mathcal{F}(X, \mathcal{D}_i)} \quad (1)$$

2.2 Travaux relatifs

La recherche de tous les EPs dans de grandes bases est une tâche difficile car le nombre de motifs candidats est exponentiel en fonction du nombre d'items. L'énumération naïve de l'ensemble des motifs avec leurs fréquences échoue rapidement. D'autre part, la définition des EPs ne fournit pas de contraintes anti-monotones par rapport à la spécialisation qui permettraient aux algorithmes par niveaux [Mannila et Toivonen, 1997] souvent employés en fouille de données de puissants élagages dans l'espace de recherche. Aussi, différents auteurs ont proposé d'autres voies.

Nous avons déjà mentionné l'approche par manipulation de bordures introduite par Dong et al. [Dong et Li, 1999]. Il s'agit de rechercher les motifs les plus fréquents dans une classe et les moins fréquents dans l'autre à l'aide de deux bordures : la première constituée des motifs infréquents minimaux et la seconde, des motifs fréquents maximaux. L'intervalle décrit par ces deux bordures correspond aux EPs. Cette méthode présente l'avantage de donner une description simple des motifs émergents. En revanche, elle nécessite de répéter pour tous les $\mathcal{D}_i$ le calcul des intervalles et elle ne fournit pas pour chaque EP son taux de croissance. Cette technique est particulièrement efficace pour la recherche des JEPs étant donné la convexité de leur espace de recherche [Li et Ramamohanarao, 2000]. Néanmoins, Bailey et al. [Bailey *et al.*, 2002] proposent une nouvelle structure de données utilisant un arbre pour représenter la base. La recherche directe des JEPs est alors de 2 à 10 fois plus rapide qu'avec la technique de manipulation de bordures.

D'autres approches existent. Zhang et al [Zhang *et al.*, 2000] introduisent une contrainte anti-monotone pour pouvoir appliquer un algorithme par niveaux. Mais celle-ci élimine de nombreux motifs émergents et la complétude de la recherche est perdue. De plus, aucune représentation particulière n'est proposée pour stocker les motifs émergents extraits. De façon plus générale, ce problème peut être vu comme la recherche de motifs vérifiant la conjonction d'une contrainte anti-monotone et d'une contrainte monotone [De Raedt et Kramer, 2001, De Raedt *et al.*, 2002], ces travaux tirant leurs origines des espaces des versions [Mitchell, 1980].

2.3 Représentations condensées basées sur les motifs fermés

Comme indiqué en introduction, cet article revisite la problématique des motifs émergents à la lumière des récents résultats sur les représentations condensées. Aussi, nous rappelons brièvement les principales notions liées à celles-ci.

Une représentation condensée de motifs fournit une synthèse des données mettant en évidence certaines corrélations (par exemple, les motifs vérifiant la propriété de fréquence). Il existe plusieurs sortes de représentations condensées de motifs [Pasquier *et al.*, 1999, Boulicaut *et al.*, 2003], les plus courantes étant les représentations condensées à base de motifs fermés, libres (ou clés) ou encore les δ-libres. Un cadre général est présenté dans [Calders et Goethals, 2002].

Il existe un double avantage à utiliser les représentations condensées. D'une part, grâce à la mise en oeuvre de puissants critères d'élagage lors de l'extraction (e.g., contrainte de liberté), le passage par une représentation condensée rend plus efficace la réalisation de certaines tâches (comme la recherche des classiques règles d'association). D'autre part, la synthèse des données fournie par une représentation condensée est un solide point de départ pour la réalisation de différentes tâches d'exploration de données comme par exemple les règles informatives ou non redondantes [Zaki, 2000], les règles à prémisses minimales [Crémilleux et Boulicaut, 2002], le clustering à base d'associations [Han *et al.*, 1997],...

Dans la suite de cet article, nous nous intéressons plus particulièrement à la représentation condensée basée sur les motifs fermés. Un motif *fermé* pour $\mathcal{D}$ est un ensemble maximal d'items (au sens de l'inclusion) partagé par un ensemble de transactions. Cette notion est liée à la théorie des treillis et à la connexion de Galois. Dans notre exemple (cf. le tableau 1), AB n'est pas un motif fermé puisque l'ensemble des transactions qui contient AB contient aussi C. En revanche, ABC est un motif fermé puisque les transactions L_1, L_2, L_3 et L_5 n'ont pas d'autres items en commun. La notion de *fermeture* est liée à celle de fermé.

Définition 2 (fermeture) *La* fermeture *d'un motif X dans $\mathcal{D}$ est $h(X, \mathcal{D}) = \bigcap\{transaction\ t | X \subseteq t\}$.*

Une propriété importante sur la fréquence découle de cette définition. Un item A appartient à la fermeture de X dans $\mathcal{D}$ si et seulement si $\mathcal{F}(XA, \mathcal{D}) = \mathcal{F}(X, \mathcal{D})$. La fermeture de X est un motif fermé et $\mathcal{F}(X, \mathcal{D}) = \mathcal{F}(h(X, \mathcal{D}), \mathcal{D})$. Dans notre exemple, $h(AB, \mathcal{D}) = ABC$ et $\mathcal{F}(AB, \mathcal{D}) = \mathcal{F}(ABC, \mathcal{D})$. Ainsi, l'ensemble des motifs fermés est une représentation condensée de tous les motifs car la fréquence de n'importe quel motif peut être inférée à partir de sa fermeture.

3 Représentation condensée et motifs émergents forts

Cette section met en évidence une nouvelle propriété caractérisant les jumping emerging patterns, elle définit une représentation condensée exacte des motifs émergents et propose les motifs émergents forts.

3.1 Caractérisation des JEPs

La définition 2 indique qu'un item A appartient à la fermeture de X dans $\mathcal{D}$ si et seulement si $\mathcal{F}(XA, \mathcal{D}) - \mathcal{F}(X, \mathcal{D}) = 0$. Cette propriété permet de caractériser les JEPs :

Propriété 1 (caractérisation des JEPs par les motifs fermés)

$$X \text{ est un } JEP \text{ de } \mathcal{D}_i \iff C_i \in h(X, \mathcal{D})$$

Preuve $C_i \in h(X, \mathcal{D}) \iff \mathcal{F}(XC_i, \mathcal{D}) = \mathcal{F}(X, \mathcal{D})$. Par définition de $\mathcal{D}_i$, $\mathcal{F}(X, \mathcal{D}_i) = \mathcal{F}(XC_i, \mathcal{D})$. Alors $\mathcal{F}(X, \mathcal{D}) = \mathcal{F}(X, \mathcal{D}_i)$ et le dénominateur de $GR_i(X)$ est nul (cf. équation 1) et X est un JEP.

Cette propriété a l'intérêt de permettre d'obtenir aisément les JEPs à partir de la connaissance des fermetures.

3.2 Représentation condensée exacte des motifs émergents

Pour connaître le taux de croissance d'un motif X quelconque, l'équation 1 montre qu'il suffit de calculer $\mathcal{F}(X, \mathcal{D})$ et $\mathcal{F}(X, \mathcal{D}_i)$. Ces fréquences peuvent être obtenues à partir de la représentation condensée des motifs fermés fréquents. En effet, $\mathcal{F}(X, \mathcal{D}) = \mathcal{F}(h(X, \mathcal{D}), \mathcal{D})$ (propriété de la fermeture) et par définition des bases partielles $\mathcal{D}_i$, $\mathcal{F}(X, \mathcal{D}_i) = \mathcal{F}(XC_i, \mathcal{D}) = \mathcal{F}(h(XC_i, \mathcal{D}), \mathcal{D})$. Malheureusement, ces relations font intervenir le calcul de deux fermetures ($h(X, \mathcal{D})$ et $h(XC_i, \mathcal{D})$) pour obtenir le taux de croissance, ce qui est peu efficace. La propriété suivante pallie cet inconvénient :

Propriété 2 *Soit un motif X et une base de données $\mathcal{D}_i$, on a $\mathcal{F}(X, \mathcal{D}_i) = \mathcal{F}(h(X, \mathcal{D}), \mathcal{D}_i)$*

Preuve Les propriétés de la fermeture assurent que pour toute transaction t, $X \subseteq t \iff h(X, \mathcal{D}) \subseteq t$. En particulier, les transactions de $\mathcal{D}_i$ contenant X sont identiques à celles contenant $h(X, \mathcal{D})$ et nous avons l'égalité des fréquences.

Il est maintenant simple de montrer que le taux de croissance peut être obtenu grâce à la seule connaissance de $h(X, \mathcal{D})$:

Propriété 3 *Soit un motif X, on a $GR_i(X) = GR_i(h(X, \mathcal{D}))$.*

Preuve Soit X un motif. En remplaçant $\mathcal{F}(X, \mathcal{D})$ par $\mathcal{F}(h(X, \mathcal{D}), \mathcal{D})$ et $\mathcal{F}(X, \mathcal{D}_i)$ par $\mathcal{F}(h(X, \mathcal{D}), \mathcal{D}_i)$ dans l'équation 1, on retrouve immédiatement l'expression du taux de croissance de $h(X, \mathcal{D})$.

Le taux de croissance d'un motif quelconque est donc celui de son fermé dans $\mathcal{D}$. Par exemple (cf. le tableau 1), on a $GR_1(AB) = GR_1(h(AB, \mathcal{D})) = GR_1(ABC) = 3$.

Cette propriété est importante car le nombre de motifs fermés est inférieur à celui des motifs quelconques [Calders et Goethals, 2002]. Les motifs fermés fréquents, accompagnés de leurs taux de croissance, suffisent donc pour synthétiser l'ensemble des motifs émergents fréquents et leur taux de croissance. On obtient ainsi une représentation

condensée *exacte* des motifs émergents (i.e. le taux de croissance de chaque motif émergent est connu avec exactitude). Rappelons que la technique de manipulation de bordures ne donne qu'une minoration des taux de croissance.

3.3 Motifs émergents forts

Le nombre de motifs émergents d'une base de données peut être rédhibitoire pour leur utilisation. Dans la pratique, il est judicieux de ne conserver que les motifs émergents les plus fréquents et ayant les meilleurs taux de croissance. Mais relever inconsidérément ces deux seuils s'avère problématique :

- si le taux de croissance minimum requis est trop élevé, les motifs émergents extraits ont tendance à être trop spécifiques (i.e. trop longs),
- si le seuil de fréquence minimum est trop élevé, les motifs émergents fréquents ont un taux de croissance trop faible.

Nous définissons ici les motifs émergents forts qui sont les motifs ayant les meilleurs taux de croissance possibles. Concrètement, ils proposent un compromis entre la fréquence et le taux de croissance.

Définition 3 (motif émergent fort) *Un motif émergent fort X (ou strong emerging pattern, SEP en abrégé) de $\mathcal{D}_i$ est un motif émergent construit sur un motif fermé XC_i dans $\mathcal{D}_i$.*

Un des atouts des motifs émergents forts est que leur taux de croissance est immédiatement connu (cf. propriété 4). Nous donnons d'abord le lemme 1 qui facilite l'explicitation de cette propriété.

Lemme 1 *Si XC_i est un fermé de $\mathcal{D}_i$, alors XC_i est un fermé de $\mathcal{D}$.*

Preuve Aucune transaction de $\mathcal{D}_i \backslash \mathcal{D}$ ne contient l'item C_i. Si XC_i est fermé dans $\mathcal{D}_i$, les seules transactions de $\mathcal{D}$ contenant XC_i se trouvent dans $\mathcal{D}_i$ et $h(XC_i, \mathcal{D}) = XC_i$, donc XC_i est fermé dans $\mathcal{D}$.

La propriété 4 indique qu'il est immédiat d'obtenir le taux de croissance des motifs émergents forts.

Propriété 4 (SEPs : obtention de leur taux de croissance) *Si X est un motif émergent fort de $\mathcal{D}_i$, alors $GR_i(X)$ peut être obtenu directement avec les fréquences issues de la représentation condensée des motifs fermés fréquents de $\mathcal{D}$.*

Preuve Soit X un motif émergent fort, donc XC_i fermé de $\mathcal{D}_i$. Pour calculer $GR_i(X)$ il est nécessaire de calculer $\mathcal{F}(X, \mathcal{D}_i)$ et $\mathcal{F}(X, \mathcal{D})$. Par définition des $\mathcal{D}_i$, $\mathcal{F}(X, \mathcal{D}_i) = \mathcal{F}(XC_i, \mathcal{D})$ et le lemme 1 assure que XC_i est fermé dans $\mathcal{D}$ donc sa fréquence est fournie par la représentation condensée des motifs fermés de $\mathcal{D}$. Pour calculer $\mathcal{F}(X, \mathcal{D})$, deux cas se présentent : si X est fermé dans $\mathcal{D}$, sa fréquence est directement disponible. Sinon, XC_i étant fermé dans $\mathcal{D}$, la propriété 1 indique que X est un JEP : son taux de croissance est infini.

La propriété 5 met en évidence le fait que les motifs émergents forts ont les meilleurs taux de croissance possibles.

Propriété 5 (SEPs : EPs de taux de croissance maximum) *Soit un motif X ne contenant pas l'item C_i. Alors le SEP construit sur $h(X, \mathcal{D}_i)$ a un meilleur taux de croissance, i.e. on a $GR_i(X) \leq GR_i(h(X, \mathcal{D}_i) \backslash \{C_i\})$.*

Preuve Soit $Y = h(X, \mathcal{D}_i) \backslash \{C_i\}$. Grâce à la propriété de fermeture, $\mathcal{F}(X, \mathcal{D}_i) = \mathcal{F}(Y, \mathcal{D}_i)$. On peut alors écrire (équation 1) $GR_i(Y) = \alpha_i \times \frac{\mathcal{F}(X,\mathcal{D}_i)}{\mathcal{F}(Y,\mathcal{D}) - \mathcal{F}(X,\mathcal{D}_i)}$. L'extensivité de l'opérateur de fermeture permet d'écrire $X \subseteq h(X, \mathcal{D}_i)$ et $C_i \notin X$ donc $X \subseteq Y$ et $\mathcal{F}(X, \mathcal{D}) \geq \mathcal{F}(Y, \mathcal{D})$ puisque la fréquence est anti-monotone, ce qui montre que $GR_i(X) \leq GR_i(Y)$.

Illustrons la propriété 5 sur l'exemple. Le motif BC n'est pas un SEP de la classe 1 (car $h(BC, \mathcal{D}_1) \backslash \{C_1\} = ABC$), son taux de croissance est 1. On a bien $GR_1(AB) \leq GR_1(ABC) = 3$ et $\mathcal{F}(AB, \mathcal{D}_1) = \mathcal{F}(ABC, \mathcal{D}_1)$.

Les propriétés 4 et 5 montrent bien deux importants atouts des motifs émergents forts par rapport aux motifs émergents quelconques : d'une part, ils sont aisés à découvrir à partir d'une représentation condensée de motifs fermés fréquents, d'autre part, ils ont les meilleurs taux de croissance possibles. Remarquons que les motifs émergents basés sur X et $h(X, \mathcal{D}_i)$ ont la même fréquence, ils ont donc la même qualité vis à vis de ce critère. Cependant, le motif émergent fort issu de $h(X, \mathcal{D}_i)$ a un plus fort taux de croissance et présente ainsi un meilleur compromis entre fréquence et taux de croissance.

4 Caractérisation de plaques de silicium

Cette section suivante montre l'apport des motifs émergents forts dans le cadre d'une application industrielle avec la société Philips.

4.1 Présentation du problème

La fabrication de plaques de silicium est une tâche délicate. Un défaut dans le procédé de production peut entraîner une chute importante du taux de composants valides. Étant donnés le temps et le coût de fabrication d'un composant, il est d'une importance majeure de détecter et de limiter le plus rapidement possible les composants défectueux. Pour cela, des tests de qualité sont effectués dès la fabrication des plaques, puis une seconde série de vérifications a lieu après le montage. Il existe une double difficulté à la recherche d'une cause de défaillance au cours de la fabrication. La première provient du nombre important d'étapes mise en jeu au cours de la fabrication qui sont autant de facteurs potentiels de défaillance. La seconde difficulté réside dans le fait qu'un rendement acceptable lors de la première série de tests, avant le montage, n'implique pas automatiquement un rendement acceptable lors de la seconde série. Dans la pratique, il est long et onéreux de vérifier les hypothèses de diagnostic, ce qui rend capital la justesse de la détection des défaillances.

4.2 Prétraitement des données

L'objectif du prétraitement est d'associer pour chaque lot (un lot est composé de plusieurs plaques de silicium) l'outil utilisé à chaque étape. Dans cette expérience, nous nous sommes focalisés sur les données discrètes issues du *flow-chart* (suivi d'un lot au cours de sa production) qui sont utilisées pour caractériser les lots. Dans la suite, pour des raisons de confidentialité, les noms de ces étapes ont été modifiés. Les valeurs de rendement ont été réparties en trois classes quasi-homogènes correspondant à 3 niveaux de qualité (`Bonne`, `Moyenne` et `Mauvaise`). Au final, la caractérisation est effectuée sur une base de données composée de 609 items distincts (i.e. paires étape/équipement) et comportant 127 lignes (i.e. 127 lots). Chacune de ces lignes possède entre 119 et 127 items.

4.3 Résultats

La caractérisation des classes a été effectuée par une recherche des motifs fréquents avec une fréquence absolue de 10. La table 2 donne la fréquence relative et le nombre de motifs émergents forts (SEPs) pour chaque classe ainsi que la répartition des SEPs selon leur taux de croissance.

Classe	Seuil de fréquence minimale (%)			Nombre de SEPs
`Mauvaise`	0.22			6532
`Moyenne`	0.27			8116
`Bonne`	0.22			14782
Classe	$GR \in [1, 2[$	$GR \in [2, 5[$	$GR \in [5, \infty[$	JEPs
`Mauvaise`	5442	896	194	0
`Moyenne`	6379	1438	112	287
`Bonne`	10750	3680	351	1

Tab. 2 – Répartitions des SEPs

Nous présentons ici uniquement une synthèse des résultats. Il y a environ 4000 fois moins de SEPs que de motifs émergents et 25 fois moins de SEPs que de motifs fermés. En fait, il s'est avéré que c'est la confrontation des SEPs de longueur 1 et 2 ayant les plus forts taux de croissance qui s'est montrée la plus intéressante pour ce problème. La table 3 indique les SEPs les plus utiles. Remarquons qu'il n'y a pas de SEP réellement caractéristique de longueur 1. Par exemple, le motif `E=727` n'est pas discriminant : non seulement son taux de croissance est proche de 1, mais en plus, ce motif est présent dans `Mauvaise` et `Bonne`. Au contraire, les SEPs de longueur 2 semblent pertinents. L'opposition entre le motif `E=727 A=284` (pour `Mauvaise`) et le motif `E=727 A=222` (pour `Bonne`) laisse suspecter un problème au niveau de l'étape `A` étant donné que `E=727` n'est pas un item discriminant. De plus, l'étape `A` ne comporte que deux équipements le 222 et le 284. Ce résultat tend à montrer la nécessité de modifier les réglages de l'équipement 284 (pour les faire tendre vers ceux de l'équipement 222). Après échanges avec les experts, ceux-ci nous ont confirmé après vérification que l'étape `A` que nous

soupçonnions problématique était la cause de la chute du rendement.

Motifs émergents forts de longueur 1			
Classe	Motif	GR	Fréquence
`Mauvaise`	`E=727`	1.01	100% (45)
`Moyenne`	`F=232`	1.03	100% (37)
`Bonne`	`E=727`	1.01	100% (45)
Motifs émergents forts de longueur 2 avec un GR > 1,5			
Classe	Motif	GR	Fréquence
`Mauvaise`	`E=727 A=284`	3.64	75.6 % (34)
`Moyenne`	`I=504 F=232`	1.84	91.9 % (34)
`Moyenne`	`L=490 F=232`	1.62	54.0 % (20)
`Bonne`	`E=727 B=288`	2.92	71.1 % (32)
`Bonne`	`E=727 A=222`	2.33	91.1 % (41)

TAB. 3 – Exemples de motifs émergents forts

Dans ce travail, les SEPs ont l'avantage de donner un taux de croissance précis contrairement aux EPs qui seraient trouvés par manipulation de bordures. Cette quantification est à la fois utile pour la sélection des EPs et pour le jugement des experts. Enfin, grâce à leur plus faible nombre, il est plus aisé de percevoir la caractérisation des lots par rapport aux équipements mis en évidence par les SEPs.

5 Conclusion

À la lumière de récents résultats sur les représentations condensées, nous nous sommes intéressés à l'extraction et à la caractérisation de motifs émergents. Nous avons défini une nouvelle caractérisation des jumping emerging patterns, une représentation condensée exacte de l'ensemble des motifs émergents d'une base de données et nous avons proposé les motifs émergents forts qui sont les motifs émergents ayant les meilleurs taux de croissance possibles. Outre la simplicité de leur extraction, les SEPs, peu nombreux, s'avèrent particulièrement utiles pour apporter une aide au diagnostic. Leur efficacité a permis de déceler, dans le cadre d'une collaboration industrielle, un équipement défaillant au sein d'une chaîne de production de plaques de silicium. Ces résultats prometteurs encouragent l'utilisation des motifs émergents forts. Plus généralement, une autre perspective est l'étude de l'apport de la représentation condensée des motifs émergents pour la classification.

Remerciements. Nous remercions la société PHILIPS et en particulier Gilles Ferru pour leur collaboration dans la préparation des données et l'interprétation des résultats obtenus. Ce travail a bénéficié du support de l'Action Spécifique STIC-CNRS Discovery Challenge.

Références

[Bailey *et al.*, 2002] J. Bailey, T. Manoukian, et K. Ramamohanarao. Fast algorithms for mining emerging patterns. In *proceedings of the Sixth European Conference on Principles Data Mining and Knowledge Discovery, PKDD'02*, pages 39–50, Helsinki, Finland, 2002. Springer.

[Boulicaut *et al.*, 2003] J. F. Boulicaut, A. Bykowski, et C. Rigotti. Free-sets : a condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery journal*, 7(1) :5–22, 2003. Kluwer Academics Publishers.

[Calders et Goethals, 2002] T. Calders et B. Goethals. Mining all non-derivable frequent itemsets. In T. Elomaa, H. Mannila, et H. Toivonen, editors, *proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'02)*, pages 74–85. Springer, 2002.

[Crémilleux et Boulicaut, 2002] B Crémilleux et J. F. Boulicaut. Simplest rules characterizing classes generated by delta-free sets. In *22nd Int. Conf. on Knowledge Based Systems and Applied Artificial Intelligence (ES'02)*, pages 33–46, Cambridge, UK, December 2002.

[De Raedt *et al.*, 2002] L. De Raedt, M. Jäger, S. D. Lee, et H. Mannila. A theory of inductive query answering. In *proceedings of the IEEE Conference on Data Mining (ICDM'02)*, pages 123–130, Maebashi, Japan, 2002.

[De Raedt et Kramer, 2001] L. De Raedt et S. Kramer. The levelwise version space algorithm and its application to molecular fragment finding. In *IJCAI*, pages 853–862, 2001.

[Dong *et al.*, 1999] Guozhu Dong, Xiuzhen Zhang, Limsoon Wong, et Jinyan Li. CAEP : Classification by aggregating emerging patterns. In *Discovery Science*, pages 30–42, 1999.

[Dong et Li, 1999] Guozhu Dong et Jinyan Li. Efficient mining of emerging patterns : Discovering trends and differences. In *Knowledge Discovery and Data Mining*, pages 43–52, 1999.

[Han *et al.*, 1997] E-H. S. Han, G. Karypis, V. Kumar, et B. Mobasher. Clustering based on association rule hypergraphs. In *proceedings of the workshop on Research Issues on Data Mining And Knowledge Discovery, SIGMOD 97*, 1997.

[Li *et al.*, 2000] Jinyan Li, Guozhu Dong, et Kotagiri Ramamohanarao. Making use of the most expressive jumping emerging patterns for classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2000.

[Li et Ramamohanarao, 2000] Jinyan Li et Kotagiri Ramamohanarao. The space of jumping emerging patterns and its incremental maintenance algorithms. In *Proc. 17th International Conf. on Machine Learning*, pages 551–558. Morgan Kaufmann, San Francisco, 2000.

[Li et Wong, 2001] J. Li et L. Wong. Emerging patterns and gene expression data. In *Genome Informatics 12*, pages 3–13, 2001.

[Mannila et Toivonen, 1997] Heikki Mannila et Hannu Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3) :241–258, 1997.

[Mitchell, 1980] T. Mitchell. Generalization as search. *Artificial Intelligence, vol. 18, n 2 p. 203-226*, 1980.

[Pasquier *et al.*, 1999] Nicolas Pasquier, Yves Bastide, Rafik Taouil, et Lotfi Lakhal. Discovering frequent closed itemsets for association rules. *Lecture Notes in Computer Science*, 1540 :398–416, 1999.

[Zaki, 2000] M. Zaki. Generating non-redundant association rules. In *proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, ACM SIGKDD'00*, pages 34–43, 2000.

[Zhang *et al.*, 2000] Xiuzhen Zhang, Guozhu Dong, et Kotagiri Ramamohanarao. Exploring constraints to efficiently mine emerging patterns from large high-dimensional datasets. In *Knowledge Discovery and Data Mining*, 2000.

Summary

Emerging patterns (EPs) are associations of features whose frequencies increase significantly from one class to another. EPs emphasize the contrast between data classes and have been proven useful to build powerful classifiers and help establishing diagnosis. Because of the huge search space, mining and representating EPs is a hard and complex task for large datasets. We propose here an exact condensed representation of EPs (i.e., all EPs and their growth rates are directly obtained from the condensed representation). The use of recent results on condensed representations of frequent closed patterns is on the core of this work. From this condensed representation, we give also a method to provide EPs with the highest growth rates. Such EPs, called strong emerging patterns, were successfully used in collaboration with the Philips company to identify the failures of a production chain of silicon plates.