

Simuler et épurer pour extraire les motifs sûrs et non redondants

Martine Cadot*
Alain Lelu**,***

*UHP/Loria, Nancy
Martine.Cadot@loria.fr,
<http://www.loria.fr/cadot>
**INRA, Crebi, Jouy en Josas
***UFC/Laseldi, Besançon
Alain.Lelu@jouy.inra.fr

Résumé. Nous présentons ici une chaîne de traitements permettant de n'extraire d'un tableau de données booléennes que les seuls 2-motifs valides statistiquement et de ne construire sur cette base que les seuls sur-motifs apportant un complément d'information. L'ensemble de ces motifs forme ainsi une représentation complète et non redondante des liaisons présentes au sein des données. Pour juger de la qualité de la liaison, positive ou négative, entre deux variables prises dans un ensemble de variables donné, pour un ensemble d'objets donné (paragraphe, patients, etc.), nous utilisons deux méthodes indépendantes s'appuyant sur une méthodologie statistique : d'un côté nous évaluons l'importance de la part de "variation" au sein des données imputable à la liaison (algorithme et indicateurs MIDOVA), de l'autre côté la significativité de cette liaison par rapport à des données de même type obtenues "par hasard" (méthode TourneBool de randomisation par échanges en cascade). Ces notions sont définies de façon théorique et utilisées sur un exemple réel formé de 193 textes caractérisés par 888 mots.

1 Introduction : problématique, principes de notre approche

En extraction de connaissances, la mise au jour des éléments les plus caractéristiques de ces données sous la forme de motifs (conjonctions de valeurs de plusieurs variables booléennes), permet classiquement de trouver les "pépites de connaissance" présentes - cf. le problème, maintes fois décrit, d'observation du panier de la ménagère, utile aux gestionnaires de grandes surfaces. Elle permet aussi de constituer de nouvelles variables pertinentes pour alimenter des traitements ultérieurs plus synthétiques : ainsi des méthodes linéaires de classification supervisées ou non, ou de projections cartographiques diverses, peuvent être enrichies par la prise en compte des non-linéarités qui résident dans ces conjonctions.

Dans les deux cas on cherche à minimiser le nombre de ces motifs, pour des raisons évidentes de réduction du temps humain d'interprétation, ou du temps machine des traitements suivants éventuels. La qualité des motifs extraits devient alors un enjeu essentiel : sont-ils ou

Simuler et épurer pour extraire des motifs pertinents

non le fruit du seul hasard ? Sont-ils ou non redondants ? Un motif de niveau k exprime-t-il davantage que ses sous-motifs de niveau $k - 1$?

Notre objectif vise un maximum de qualité dans l'extraction, à partir d'un tableau de P variables booléennes, des motifs de tous niveaux, depuis les 1-motifs que sont les variables elles-mêmes, jusqu'aux motifs de niveau supérieur. Nous proposons une démarche de qualité à deux volets : 1) validation par un test de randomisation, 2) suppression des redondances entre niveaux de motifs ; nous évitons ainsi les problèmes d'explosion combinatoire qui se posent habituellement et sont résolus de façon statistiquement insatisfaisante par des seuils ad-hoc. Cette démarche conjugue l'efficacité dans la réduction du volume de résultats (dans l'exemple que nous présentons, passage de quelque 400 000 à 4000 2-motifs statistiquement valides, puis 2276 3-motifs et 41 4-motifs) à l'efficacité de calcul (temps de calcul dévolu principalement à la validation des 2-motifs, et temps négligeable pour les motifs de niveau supérieur).

En apprentissage supervisé ou non, le problème de la sélection de motifs parmi les 2^P possibles a fait l'objet de nombreux travaux depuis Agrawal et Verkamo (1996). Ce problème est particulièrement crucial pour l'analyse non supervisée de corpus textuels, où le nombre de variables extraites - ce sont ici des mots - dépasse facilement la dizaine de milliers. Pire, l'extraction des mots composés, unités sémantiques véritables pouvant se compter par centaines de milliers, fait l'objet d'approches syntaxiques et/ou statistiques comportant de nombreux choix empiriques, et nécessitant une phase de validation manuelle quand un haut degré de qualité est exigé (Lelu A. (1997)).

Un ensemble de variables décrivant un ensemble d'observations n'a de sens en fouille de données qu'en tant que système, c'est à dire ensemble d'éléments liés par des relations de covariation de ses constituants. Un ensemble d'éléments isolés n'est pas un système, et les éléments non covariants d'un système, qui varieraient de façon propre et indépendante des autres éléments, ne feraient pas partie de ce système.

Certaines de ses covariations peuvent être dues au seul hasard : des descripteurs booléens fréquents ont beaucoup plus de chances d'être observés ensemble que des descripteurs rares, et inversement. Notre ensemble de traitements incorpore donc un test de randomisation par simulation de matrices de données de même structure que la matrice observée, ayant les mêmes sommes marginales que cette matrice.

D'autres covariations au sein de motifs de niveau K gardent - ou non - un potentiel de covariation pour des motifs de niveau $K + 1$: c'est le principe de notre algorithme Midova, qui construit itérativement une suite limitée - dans notre exemple au niveau 4 - de motifs non redondants de longueur croissante. Par exemple un motif non redondant de niveau 3 peut traduire une covariation différente de celle de ses sous-motifs de niveau 2, ce que les modèles statistiques classiques désignent sous le terme d'*interaction* (Winer B.J. (1991); Jakulin (2003)).

2 Valider statistiquement les liaisons entre variables exprimées par les 2-motifs

Alors que la notion de liaison entre deux variables ne pose pas de problème particulier, elle est plus délicate à définir entre plus de deux variables, comme le note François Bavaud dans Bavaud (1998). En effet la situation de non-liaison, ou indépendance, référence par rapport

à laquelle on mesure un écart (sur la significativité duquel on statue), peut prendre plusieurs définitions, dont les plus classiques sont :

- L'indépendance totale est définie par un effectif "fictif", obtenu à partir du produit des probabilités des marges ; ainsi dans le cas de 3 variables avec un effectif total N et des sommes marginales $n_{i..}$, $n_{.j.}$, $n_{..k}$: $\hat{n}_{ijk} = Np_{ijk}$, où $p_{ijk} = p_i p_j p_k$; $p_i = n_{i..}/N$, $p_j = n_{.j.}/N$, etc. Cette indépendance est inconditionnelle en ce sens qu'elle intègre l'indépendance de toutes les combinaisons de niveau inférieur. L'écart à cette hypothèse a été exploré par Brin et al. dans notre présent cadre d'extraction de motifs et de règles d'association : ces auteurs présentent avec prudence leur approche, qui utilise la distance à l'hypothèse d'indépendance (le classique test du Khi-deux), en indiquant clairement ses limites : *For association rules, these [validity] conditions [of the Khi2 test] will frequently be broken [...] The solution of the problem is to use an exact calculation for the probability, rather than the Khi2 approximation. The establishment of such a formula is still, unfortunately, a research problem [...]*.
- L'indépendance des variables d'un motif de niveau k peut n'être que conditionnelle, c'est à dire se produire en sus de dépendances (ou non) au sein des sous-motifs de niveau inférieur. Dans le présent travail, nous nous proposons de repérer les seules situations d'écart fort et incontestable à ce type d'indépendance, laissant ouverte la caractérisation fine de la situation d'indépendance.

Jusqu'ici, à notre connaissance, les seules autres tentatives de validation statistique de motifs ont été celles initiées par Régis Gras (Gras (1979)) ; celles-ci nécessitent que les données suivent des lois de répartition spécifiques, comme la loi normale ou celle de Poisson. Elles partagent également avec la tentative précédente d'utilisation du Khi2 l'inconvénient de ne pas tenir compte du contexte global : comme elles ne prennent en considération que les quelques colonnes du tableau Individus \times Variables concernées par le motif ou la relation d'implication, elles supposent une même loi de répartition quelles que soient les colonnes, et leur appliquent la même logique de seuil ; elles ne tiennent pas compte du fait qu'une valeur *un* dans une ligne comportant beaucoup de *uns* n'a pas la même signification que dans une ligne en comportant peu ; plus généralement, elles ignorent le problème statistique réputé difficile des *comparaisons multiples* (Jensen et Cohen (2000)). Dans cet ordre d'idées, J. Press (Press (2004)) développe une quinzaine de raisons pour lesquelles les tests statistiques habituels sont dans la plupart des cas inopérants en fouille de données : notamment beaucoup de variables de types et de distributions divers, et un nombre d'observations dépassant de beaucoup la trentaine qu'il suffit pour considérer "grands" les échantillons des statistiques habituelles.

L'utilisation des tests de *randomisation* décrits par Manly (Manly (1997)) a contribué à notre réponse rigoureuse à ces objections. Ces tests ont pour origine le *test exact de Fisher*, répertoriant l'ensemble des combinaisons d'effectifs des cases du tableau des données compatibles avec la taille de la population, quand celle-ci ne dépasse pas la dizaine d'individus. L'augmentation de puissance des ordinateurs et le développement des techniques informatiques de simulation du hasard a permis aux statisticiens de développer ces tests ne nécessitant pas de connaître la loi de distribution des données, afin qu'on puisse les appliquer à des données de toutes tailles. Nous avons examiné dans Cadot et Napoli (2003) l'application de ce principe à la recherche de 2-motifs dans des données booléennes, au travers de quelques expériences, et conclu sur la nécessité de tirer des permutations respectant les sommes marginales du tableau de données initial. Notre algorithme rigoureux de permutations par *échanges rectangulaires*

Simuler et épurer pour extraire des motifs pertinents

pour validation des 2-motifs a été présenté à CSDA 2005 (Cadot (2005)); sa justification théorique se trouve dans Cadot (2006), basée sur la notion, originale à notre connaissance, d'*échanges en cascade* : nous montrons qu'avec un nombre fini d'échanges en cascade, on peut passer de toute matrice booléenne de marges données à toute autre de mêmes marges. Et ces échanges en cascades peuvent être obtenus par composition d'échanges rectangulaires.

Nous présentons ici en section 4 une application de cet algorithme, renommé plus brièvement "Tournebool", à des données réelles; celui-ci se limitant à l'extraction de 2-motifs valides, il intervient en préalable à l'extraction de motifs d'ordre supérieur, que nous examinons ci-après.

3 Algorithme et indicateurs MIDOVA

La stratégie la plus courante pour limiter l'explosion combinatoire des motifs consiste à choisir un seuil de support, et à n'extraire que ceux dont le support dépasse ce seuil (Agrawal et Verkamo (1996); Bastide (2000)). Pour les interpréter en terme de liaisons entre variables, on extrait de ces motifs des règles d'association, qu'on range selon leurs valeurs à divers indices de qualité (Lenca et al. (2003); Guillet (2004)), afin de se limiter à celles de meilleure qualité selon la sémantique couverte par ces indices.

Le choix préalable d'un seuil de support a des inconvénients gênants pour notre but.

1. Le seuil fixe de support ne permet pas de distinguer les cas d'associations fortes entre variables de faibles supports et d'associations fortuites entre variables de forts supports, ces dernières étant dues au seul effet des lois marginales de la matrice des données.
2. Il fait disparaître les oppositions entre variables, quels que soient leurs supports respectifs (le support du motif les liant étant faible, voire nul dans le cas de variables exclusives l'une de l'autre). Il fait également disparaître les associations positives entre variables rares, sa valeur étant fixée indépendamment du support des variables constituant les motifs. A cela s'ajoutent des inconvénients dus à l'extraction proprement dite, qui produit des motifs pour lesquels aucun souvenir des détails de leur composition n'est conservé, compromettant ainsi toute interprétation fine postérieure, et des inconvénients dus à la redondance en cas de variables avec des valeurs très proches (si A, B, C et D sont recouvrantes, les motifs AB, AC, AD, BD seront extraits ainsi que ABC, ABD, ACD, BCD et ABCD). Nous désirons une extraction de motifs sans ces inconvénients, mais fournissant un nombre raisonnable de motifs.

Pour mesurer 1) le gain d'information d'un motif M par rapport à ses sous-motifs et 2) son potentiel de création de sur-motifs, nous nous appuyons sur les variations possibles de son support. On impose à ces variations de se faire en laissant les supports des sous-motifs de M inchangés. Pour calculer ces deux indices à partir du support s du motif M et de sa longueur L (le nombre de propriétés le constituant), nous calculons au préalable les bornes et le centre c de l'intervalle de variation du support s . Puis :

- *L'indice MIDOVA-g*. Pour la valeur du gain qu'il traduit, les considérations détaillées dans Cadot (2006) nous amènent à choisir la fonction $g = 2^{L-1}(s - c)$. On peut trouver dans l'exemple qui suit une illustration de ces considérations.
- *L'indice MIDOVA-r*. Une autre caractéristique essentielle dans notre optique est le "reste" de variabilité possible pour les sur-motifs, défini comme 2^{L-1} fois la différence entre le

support s et la borne (inférieure sg ou supérieure sd) la plus proche de s . Sa valeur commande la poursuite de l'algorithme ou son arrêt : $r = 2^{L-1} \min(|s - sg|, |s - sd|)$

3.1 Exemple de recherche de l'intervalle de variation

Prenons le cas de 60 sujets pour lesquels nous connaissons les valeurs de 3 propriétés A, B et C. Les supports respectifs de A, B, C, AB, AC, BC, ABC sont 35, 28, 40, 20, 27, 22 et 15. Les valeurs des 60 sujets pour les 3 propriétés sont représentées dans la figure 1 par un tableau d'incidence et par un diagramme de Venn. Comme il y a 3 propriétés, le tableau contient $2^3 = 8$

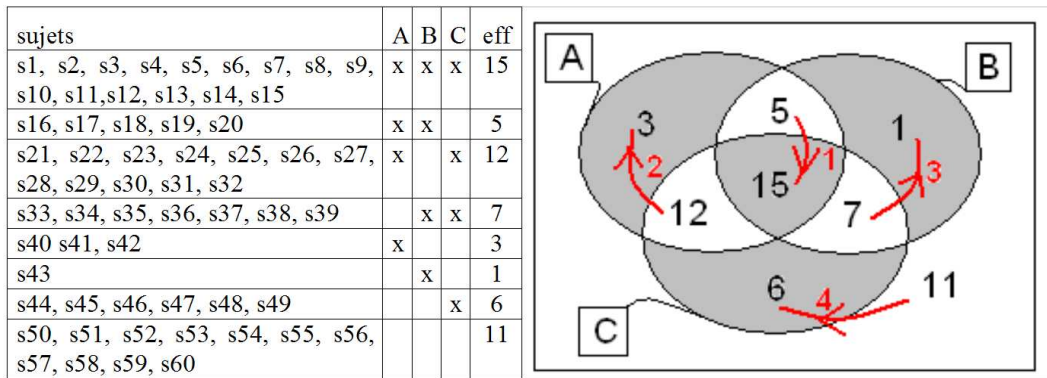


FIG. 1 – Répartition de 60 sujets selon 3 propriétés.

lignes, et le diagramme de Venn 8 zones. La zone où les trois propriétés sont simultanément vérifiées est grisée, ainsi que les autres zones dont le nombre de propriétés est impair, donc ici vérifiant une seule propriété. Les quatre zones restantes (blanches) sont celles où un nombre pair de propriétés (0 ou 2) sont vérifiées. Pour chercher l'intervalle de variation du support de ABC, à support constant de ses sous-motifs, on essaie d'abord d'augmenter ce support. En suivant la flèche 1, on déplace un sujet (par exemple s_{16}) en lui ajoutant la propriété C qu'il n'avait pas. Il passe ainsi d'une zone non grisée à une zone grisée. Lors de ce déplacement, le support de AB ne change pas. Par contre ceux de AC et de BC augmentent d'une unité chacun. On compense cette augmentation en suivant les flèches 2 et 3, qui déplacent par exemple les sujets s_{21} et s_{33} en leur retirant la propriété C. Ce déplacement a pour conséquence une diminution du support de C d'une unité, qu'on compense en déplaçant par exemple le sujet s_{50} selon la flèche 4. Ainsi, si on désire augmenter le support de ABC, qui est dans une zone grisée, sans modifier les supports de ses sous-motifs, il faut augmenter d'autant les effectifs des 3 autres zones grisées et diminuer d'autant chaque effectif d'une zone non grisée. Comme le plus petit effectif des zones non grisées est 5, le support de ABC ne peut pas augmenter de plus de 5. Et pour faire diminuer le support de ABC, on procède de façon inverse, ce qui fait qu'il ne peut pas diminuer de plus de 1, minimum des effectifs des zones grisées. Le support de M varie ainsi entre 14 et 20, sa valeur centrale étant 17. Il ne reste plus qu'à calculer g en remplaçant le support s par 15, la longueur L par 3, et le centre c par 17, ce qui donne

Simuler et épurer pour extraire des motifs pertinents

$g = 2^{L-1}(s - c) = 4(15 - 17)$, soit -8, ce qui veut dire qu'il faut déplacer 8 sujets pour faire passer le support de ABC de 17 à 15 sans changer les supports de ses sous-motifs. En résumé :

- Borne inférieure : $sg = s - \min(\text{zones d'arité impaire})$
- Borne supérieure : $sd = s + \min(\text{zones d'arité paire})$
- Centre : $c = (sg + sd)/2$

3.2 L'algorithme MIDOVA

- 0) Initialisation : on choisit une valeur e correspondant à un écart négligeable du support, dont l'unité est l'objet, donc en nombre d'objets. Cette valeur peut être 0, 1, 2 ou plus.
- 1) Au niveau 1 (motifs réduits à une variable), chaque motif a un support compris entre 0 et N . On calcule son écart au centre "neutre" $N/2$ de l'intervalle, et la part de variabilité qu'il laisse aux supports des sur-motifs ; cette part est le support s lui-même si $s > N/2$, ou $N - s$ dans le cas contraire. Les motifs pour lesquels cette part est négligeable (inférieure ou égale à e), quand ils ont un support s proche de zéro ($s \leq e$), ou proche de N ($s \geq N - e$), ont épuisé leur part de variabilité. On les élimine des motifs à fusionner pour le niveau suivant.
- 2) Niveau k (tant qu'il reste des motifs) : On combine les seuls motifs du niveau précédent qui sont combinables en un motif Mk de niveau k , et on en déduit le support s de Mk et l'intervalle de variation de celui-ci ($sg ; sd$). On calcule l'écart de s au centre de l'intervalle, et la part de variabilité qu'il laisse aux sur-motifs de Mk (c'est le reste MIDOVA-r : l'écart entre sa valeur et la borne la plus proche sg ou sd). Si cette part est négligeable, il a épuisé sa part de variabilité, on l'élimine des motifs à combiner à l'étape suivante.

Lorsque l'algorithme a convergé (ce qui se produit d'autant plus rapidement que e est grand), on interprète les motifs obtenus en terme de gain : positif si interaction positive, négatif dans le cas contraire d'exclusion entre la présence des variables.

4 Application

Les données sont constituées à partir des résumés des 193 premiers livres de la collection Gallimard-jeunesse, qui forment une encyclopédie touchant des sujets très variés, résumés caractérisés par la présence de 888 mots au terme d'une extraction de termes semi-automatique. Ces présences/absences peuvent être représentées par une matrice booléenne Docs×Mots contenant 6559 *uns*. Le nombre de mots par document varie entre 3 et 63, la répartition des documents selon leur nombre de mots suivant une distribution approximativement binomiale, avec beaucoup de documents ayant entre 30 et 40 mots (cf. figure 1). Le nombre de documents par mot varie entre 1 et 62, la répartition des mots selon leur fréquence se faisant selon une distribution inégalitaire, plus de 90% des mots figurant dans moins de 15 textes chacun.

4.1 Le test de significativité des 2-motifs

Nous nous sommes d'abord intéressés aux associations de deux mots. Sur la base de 888 mots, il y en a $888 * 887 / 2$ soit 393 828. La seule information pertinente que nous avons retenue

sur ces associations de deux mots est le nombre de documents dans lesquels ils apparaissent simultanément. Nous avons décidé de ne garder que les associations de deux mots (2-motifs) figurant dans plus de textes qu'attendu par hasard, et celles dans moins de textes qu'attendu par hasard, en prenant un risque alpha inférieur ou égal à 5% (risque de se tromper en estimant qu'une association n'est pas due au hasard) dans cette décision. Ainsi c'est un test bilatéral que nous faisons, permettant d'établir un intervalle de confiance à 95% des valeurs du support en cas d'absence de liaison, les 2,5% à gauche de cet intervalle représentant les support trop petits pour être dus au hasard et les 2,5% à droite ceux trop grands pour être dus au hasard. Pour le réaliser nous générons au hasard et de façon indépendante 200 matrices booléennes ayant mêmes sommes marginales que la matrice de données. Et pour chaque association de deux mots, nous cherchons dans chaque matrice simulée combien de fois elle apparaît dans des documents. Nous disposons ainsi du support réel du motif de longueur 2 et de la liste de ses supports dans les matrices simulées.

Il y a peu de 2-motifs moins fréquents qu'attendus, mais bien davantage de plus fréquents qu'attendu. Par exemple le 2-motif *famille, rêve* a un support de 0 dans les données d'origine, ce qui signifie que ces 2 mots ne sont jamais dans un même texte, alors que dans 95% des matrices simulées, il apparaît avec un support compris entre 1 et 7. Ce qui indique une opposition significative entre ces deux mots dans notre corpus. De même, le 2-motif *peintre, ville* a un support de 2 dans les données d'origine, ce qui est peu par rapport au support de chacun (respectivement de 26 et 44) comme l'indiquent les données simulées qui font apparaître un intervalle à 95% de confiance de (3, 10) : dans cette collection encyclopédique, ces deux thèmes fréquents et distincts ont peu de recouvrement. Par contre le 2-motif *François 1er, Charles Quint* est peu fréquent (support de 3), mais plus fréquent qu'attendu (intervalle de confiance de (0,1)), ce qui s'explique par la rareté de chacun de ces deux mots (supports respectifs de 5 et 4).

Après la phase de test statistique d'échanges en cascade, il reste 4 000 motifs de longueur 2 significatifs avec un risque $\alpha=5\%$, de support s avec $0 \leq s \leq 46$.

4.2 Construction des motifs d'ordre supérieur valides avec l'algorithme Midova

Parmi les 4000 motifs de longueur 2, 3686 ont une valeur de r (reste selon Midova) supérieure à 1. Ils se combinent en 2276 motifs de longueur 3, dont 587 de r supérieur à 1, ces derniers créant 41 motifs de longueur 4, dont 2 de r supérieur à 1, trop peu nombreux pour produire des motifs de longueur supérieure. Ces motifs peuvent s'interpréter selon leur valeur de gain g . Voici quelques exemples d'interprétation.

Le 4-motif *archéologie, fouille, légende, site* a un indice r de 0, ce qui indique qu'il ne peut plus contribuer à un 5-motif. Son indice g est de 4, ce qui indique une liaison positive entre ces 4 mots plus informative que la liaison entre ces mots pris 3 à 3 et 2 à 2. Parmi ceux-ci, le 2-motif *fouille, légende* a un indice g de -2 qui indique une faible liaison négative. De même le 3-motif *pouvoir, puissance, Jérusalem* a un indice r de 0 et un indice g de 8, et le 2-motif *cinéma, film* a un indice r de 4 et un indice g de 9. Les plus fortes valeurs de g apparaissent pour les associations entre mots composés et leurs composants, par exemple *XXe siècle* et *XXe*, ou *chef* et *chef d'oeuvre*, *guerre mondiale* et *seconde guerre mondiale*. L'opposition maximale a lieu entre *Etats-Unis* et *Moyen-Age* avec $g=-24$.

5 Conclusion : efficacité de la chaîne de traitement proposée

L'action combinée du test statistique et de Midova a réduit l'explosion de la pyramide des motifs en largeur (2-motifs) par un facteur 144 et en hauteur par un facteur d'au moins 2, comme on peut le voir dans la figure 2. La courbe pointillée en haut à gauche est le résultat de l'extraction de toutes les associations de 2 mots, que ceux-ci figurent ou non dans un même texte (s0r0). Elle indique clairement le caractère exponentiel de cette démarche naïve. Les 3 courbes pointillées rouges montrent le résultat de l'approche *A Priori* (le seuil de support est de 2), combinée ou non avec un filtrage par le test d'échange en cascade (haz05 pour le risque de 5%, haz01 pour 1%, haz1 sans le test). Les 3 courbes bleues représentent les résultats de l'approche Midova ($r \geq 2$) combinée ou non avec le test. Les 3 courbes noires combinent l'approche *A Priori* et Midova, mais ne permettent pas d'obtenir les oppositions. Il en ressort clairement que notre approche permet d'obtenir les oppositions partielles ou totales que ne ressortent pas des autres méthodes, ainsi qu'une condensation non redondante des liaisons entre variables. Notre approche fournit une quantité de motifs bien inférieure et de moindre complexité, valides statistiquement, et d'interprétation plus riche, le tout pour un temps de calcul de la construction des k-motifs ($k > 2$) négligeable par rapport à celui de la validation des 2-motifs. Nous comptons approfondir et optimiser ces aspects d'efficacité de calcul afin de passer à l'échelle de corpus plus importants que l'exemple déjà conséquent traité ici.

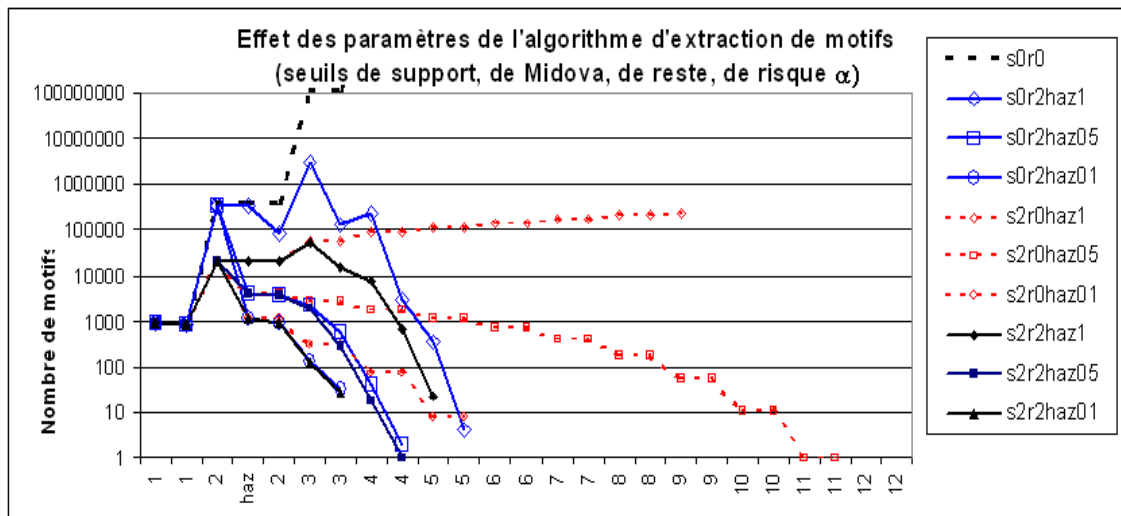


FIG. 2 – Comparaison des efficacités respectives de l'utilisation combinée de Midova et du test d'échanges en cascade avec la méthode classique de type *A Priori*, avec seuil de support, où on trouve en abscisse les longueurs des motifs et en ordonnée leur nombre. Chaque longueur est présente deux fois, la première pour les motifs extraits, la seconde pour ceux dont le reste r est non nul, l'étape de hasard étant indiquée par la valeur "haz". Pour chaque courbe, on a indiqué le seuil de support, de Midova- r , et le risque α (ainsi s0r2haz05 indique un seuil de support de 0, de reste de 2, et le risque α de 5%)

Et ces principes ne sont pas limités aux données binaires. Nous avons ainsi commencé à définir un gain pour les règles d'association floue qui prolonge celui que nous venons de définir pour les RA classiques (Cuxac et al. (2005)).

Références

- Agrawal, H. Mannila, R. S. H. T. et A. I. Verkamo (1996). *Advances in Knowledge Discovery and Data Mining*, Chapter Fast Discovery of Association Rules, pp. 307–328. Menlo Park, California : AAAI Press /MIT Press.
- Bastide, Y. (2000). *Data mining : algorithmes par niveau. techniques d'implantation et applications*. Ph. D. thesis, Université Blaise Pascal, Clermont-Ferrand.
- Bavaud, F. (1998). *Modèles et données : une introduction à la Statistique uni-, bi- et trivariée*. Paris ; Montréal (Qc) : L'Harmattan.
- Brin, S., R. Motwani, et C. Silverstein. Beyond market baskets : Generalizing association rules to correlations. In *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15*, Tucson, Arizona, USA.
- Cadot, M. (2005). A simulation technique for extracting robust association rules. In *CSDA 2005*, Chypre.
- Cadot, M. (2006). *Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association*. Ph. D. thesis, Université de Franche-Comté.
- Cadot, M. et A. Napoli (2003). Une optimisation de l'extraction d'un jeu de règles s'appuyant sur les caractéristiques statistiques des données. Volume 16, pp. 631–656.
- Cuxac, P., M. Cadot, et C. François (2005). Analyse comparative de classifications : apport des règles d'association floues. In *EGC 2005*, pp. 519–530.
- Gras, R. (1979). *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques*. Ph. D. thesis, Université de Rennes I.
- Guillet, F. (2004). Mesure de qualité des connaissances en ecd.
- Jakulin (2003). Attribute interactions in machine learning. Master's thesis, University of Ljubljana, Slovenija.
- Jensen, D. D. et P. R. Cohen (2000). Multiple comparisons in induction algorithms. *Machine Learning* 38(3), 309–338.
- Lelu A., Tisseau-Pirot A.-G., A. A. (1997). Cartographie de corpus textuels évolutifs, un outil pour l'analyse et la navigation. *Hypertextes et Hypermédias 1*.
- Lenca, P., P. Meyer, P. Picouet, et B. Vaillant (2003). Aide multicritère à la décision pour évaluer les indices de qualité des connaissances. In *EGC 2003*, pp. 271–282.
- Manly, B. (1997). *Randomization, Bootstrap and Monte Carlo methods in Biology*. Texts in Statistical Science. Boca Raton, Florida, USA : Chapman & Hall/CRC.
- Press, J. S. (2004). *Statistical Data Mining and Knowledge Discovery*, Chapter The role of Bayesian and frequentist multivariate modeling in statistical Data Mining, pp. 309–338. Boca Raton, US : Chapman & Hall/CRC.

Simuler et épurer pour extraire des motifs pertinents

Winer B.J., Brown D.R., M. K. (1991). *Statistical principles in experimental design* (third edition ed.).

Summary

Our goal is twofold: 1) we want to mine the only statistically valid 2-itemsets out of a boolean datatable, 2) on this basis, we want to build the only higher-order non-redundant itemsets compared to their sub-itemsets. For the first task we have designed a randomization test (*Tournebool*) respectful of the structure of the data variables and independant from the specific distributions of the data. In our test set (193 texts and 888 terms), this leads to a reduction from 400,000 2-itemsets to 4000 significant ones, at the 95% confidence interval. For the second task, we have devised a hierarchical stepwise procedure (*MIDOVA*) for evaluating the residual amount of variation devoted to higher-order itemsets, yielding new possible positive or negative high-order relations. On our example, this leads to 2300 3-itemsets, 41 4-itemsets, and no higher-order ones, in a computationally efficient way.