Du local au global : un nouveau défi pour l'analyse statistique implicative

Thomas Delacroix, Philippe Lenca, Stéphane Lallich

DU LOCAL AU GLOBAL : UN NOUVEAU DÉFI POUR L'ANALYSE STATISTIQUE IMPLICATIVE

Thomas DELACROIX[1], Philippe LENCA[2] et Stéphane LALLICH[3]

GOING GLOBAL:A NEW CHALLENGE FOR S.I.A.

RÉSUMÉ

L'analyse statistique implicative ainsi que la très grande majorité des approches qui ont été développées en extraction de règles sont confrontées à un certain nombre de difficultés dont la surabondance et le caractère redondant des règles extraites. Si de nombreux chercheurs en extraction de motifs ont essayé de proposer des solutions à ces problèmes, ils sont bien moins nombreux à en avoir cherché les causes. On présente ici l'une des causes principales de ces problèmes : le caractère local des approches en extraction de règles qui caractérise à la fois les modèles utilisés pour la mesure de l'intérêt des règles et les processus d'extraction de règles ; ainsi que son contre-pied : le choix d'une approche globale. On propose un certain nombre d'options permettant de faire ce choix dans le but d'ouvrir de nouvelles perspectives et d'accompagner la transition des approches courantes toujours très majoritairement locales vers des approches plus globales.

Mots-clés : *extraction de règles, mesure d'intérêt, modèles probabilistes, modèle d'entropie maximale, indépendance mutuelle contrainte, analyse statistique implicative.*

ABSTRACT

Statistical implicative analysis, as well as the vast majority of current approaches towards rule mining, are faced with a certain number of difficulties such as the issue of over abundant and redundant rules. While a large number of papers in pattern mining have addressed these issues, most have concentrated on finding solutions rather than identifying causes. We present here one of the main causes of these issues: the local nature of the various approaches in rule mining, which can qualify both the models used for the measure of rule interestingness and the mining process; as well as its counterpart: choosing a global approach. We suggest a certain number of options allowing to make this choice with the objective of opening up to new perspectives and encouraging the transition from a vast majority of current local approaches towards more global approaches.

Keywords : *rule mining, interestingness measures, probabilistic models, maximum entropy model, mutual constrained independence, statistical implicative analysis.*

1 Introduction

Depuis les trois dernières décennies, la recherche scientifique en extraction de connaissances et en apprentissage statistique a connu un véritable essor porté par un certain nombre de grandes thématiques, parmi lesquelles on retrouve les techniques de fouille reposant sur l'extraction de motifs et, en particulier, les règles de type $a \rightarrow b$ (Han et al., 2012, Klösgen and Zytkow, 2010, Friedman et al., 2008, Aggarwal et Han, 2014). Au sein de cette thématique, l'approche la plus diffusée a été, sans nul doute, celle de l'extraction de règles d'associations dans le cadre de l'analyse de panier de course qui

1 IMT Atlantique - UMR 6285 Lab-STICC, Brest, France, thomas.delacroix@imt-atlantique.fr
2 IMT Atlantique - UMR 6285 Lab-STICC, Brest, France, philippe.lenca@imt-atlantique.fr
3 Université de Lyon, Laboratoire ERIC, Lyon 2, France, stephane.lallich@univ-lyon2.fr

repose sur l'extraction d'itemsets et doit certainement une grande partie de son succès à sa simplicité (Agrawal et Srikant, 1994, Kotsiantis et Kanellopoulos, 2008, Zhao et Bhowmick, 2003, Fournier-Vigier et al., 2017). Toutefois, un certain nombre d'autres approches plus élaborées ont également été étudiées et l'analyse statistique implicative est l'une d'entre elles (Gras et al., 2009, 2013). Ce qui permet de regrouper des approches aussi différentes que l'analyse statistique implicative et l'extraction de règles d'associations, c'est bien évidemment le choix d'un motif fléché comme motif élémentaire à extraire des données (règle d'association dans un cas, quasi-implication dans l'autre), mais également le principe de sélection des motifs à extraire qui passe par l'utilisation d'une ou plusieurs mesures d'intérêt de la règle pour évaluer, règle par règle, de l'intérêt suffisant d'une règle justifiant son extraction. Dans le cas de l'analyse statistique implicative, il s'agit de l'intensité d'implication alors que dans le cas des règles d'association, il s'agit du support et de la confiance. De manière plus générale, un très grand nombre de mesures d'intérêt ont été envisagées pour l'extraction de règles (plus d'une soixantaine recensée entre Lallich et al. (2007), Geng et Hamilton (2009), Le Bras (2011) et Wu et al. (2012)) et un certain nombre de travaux ont proposé des classifications de ces mesures selon différents critères afin d'aider un utilisateur à choisir la mesure d'intérêt la plus adaptée à ses besoins (Lenca, 2004, Lenca et al., 2008, Le Bras, 2011, Wu et al., 2012). L'intensité d'implication n'a pas échappée à ce travail de comparaison des mesures d'intérêt des règles et on peut remarquer qu'elle satisfait de nombreux critères pertinents et exigibles d'une telle mesure (Gras et al., 2004, Gras et Couturier 2013). Mais, malgré toutes ses spécificités, l'analyse statistique implicative est limitée, au même titre que l'extraction de règles d'association, par le caractère intrinsèquement local d'une méthodologie de sélection d'une règle par la mesure de l'intérêt de cette règle. En effet, comme il sera détaillé par la suite, cette méthodologie peut être qualifiée de locale à deux titres. D'abord, parce que la mesure de l'intérêt d'une règle de type $a \rightarrow b$ se fait dans le cadre d'un modèle local de l'intérêt défini uniquement par des paramètres locaux de la règle. Ensuite parce que le processus d'extraction des règles par la mesure de l'intérêt se fait de manière locale, en ne considérant qu'une seule règle à la fois, plutôt que de mesurer l'intérêt de l'ensemble des règles extraites dans sa globalité.

Or c'est à cause du caractère local de ce processus, à cause de cette limitation précisément, que quelques-uns des obstacles les plus importants en extraction de règles ne peuvent être véritablement franchis. Si le passage du local au global peut lui aussi receler un certain nombre d'embûches, le développement d'approches fondées sur le modèle d'entropie maximale (Jaynes, 1982, Berger et al., 1996, Mannila et al., 1999, Pavlov et al., 2000, Pavlov et al., 2003, Vreeken et Tatti, 2014) ou de sa réinvention sous le nom de modèle d'indépendance mutuelle contrainte dans le contexte des itemsets (Delacroix et al., 2015, Delacroix et al., 2017a, Delacroix et al., 2017b) permettent de l'envisager sérieusement. En particulier, l'analyse statistique implicative et sa communauté présentent un terrain particulièrement favorable pour expérimenter ce qui pourrait être un véritable changement de paradigme en extraction de règles, d'une approche locale à une approche globale.

2 Notations

Dans le souci d'une plus grande clarté, on fixe quelques termes et notations qui seront utilisés tout au long de cet article.

On considère un problème d'extraction de règles d'une base de données binaires portant sur l'observation de p attributs sur n individus. Les données peuvent donc être représentées par une matrice D de taille $n \times p$ à valeurs dans $\{0,1\}$ de telle manière que $D_{i,j} = 1$ si et seulement si le i-ème individu possède le j-ème attribut (ou, dit autrement, on a observé le j-ème attribut au cours de la i-ème observation). Les attributs seuls seront représentés par des lettres indexées : a_1, a_2, …, b_1, b_2, … ; et les expressions logiques sur les attributs (par exemple, la conjonction de trois attributs) par des lettres sans indices : a, b, … La conjonction et la disjonction sont notés par les symboles standards ∧ et ∨ et la négation d'une expression a est notée $\bar{a}$. Enfin, la fréquence relative de l'observation d'une expression logique d'attributs a est notée f_a et elle sera différenciée de la probabilité d'observation de a, étant donnée une observation quelconque, que l'on notera p_a (voir la section 3.2 et 3.3 pour plus de détails).

3 Modèles locaux – modèles globaux

Il est frappant de remarquer que l'intégralité des mesures d'intérêt référencées dans les différents travaux portant sur la caractérisation et l'étude des mesures d'intérêt de règles (Vaillant et al., 2004, Lallich et al., 2007, Lenca et al., 2008, Geng et Hamilton, 2009, Le Bras, 2011, et Wu et al., 2012) peuvent être décrites comme des fonctions de trois, voire quatre, mêmes variables. Ces variables sont :

- f_a : la fréquence relative des observations de a ;
- f_b : la fréquence relative des observations de b ;
- $f_{a \wedge b}$: la fréquence relative des observations de a conjointement à b ;
- n : la taille de l'échantillon total (plus rarement).

Il peut s'agir d'une fonction très simple, comme c'est le cas pour le support ou la confiance :

$$supp(a \to b) = f_{a \wedge b} \quad \text{et} \quad conf(a \to b) = {f_{a \wedge b}}/{f_a}.$$

Ou bien d'une fonction plus élaborée, comme dans le cas de l'intensité d'implication :

$$\varphi(a \to b) = \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{+\infty} e^{-\frac{t^2}{2}}\, dt \quad \text{où} \quad q(a,\bar{b}) = \frac{n(f_a - f_{a \wedge b}) - f_a(1-f_b)}{\sqrt{n f_a (1-f_b)}}.$$

Mais, dans tous les cas, cette fonction se limite au contexte local de la règle sans prendre en compte le fait que la règle s'inscrit dans un contexte plus général qui englobe chacune des règles. Cela revient à poser la question de l'intérêt de la règle en soi sans se poser la question de savoir si cet intérêt est compatible avec celui que l'on peut mesurer pour une autre règle. Or, si l'on ne se pose pas cette question, l'ensemble des intérêts mesurés localement ne présentera pas, a priori, de cohérence globale. Ce phénomène est particulièrement visible lorsque la mesure de l'intérêt s'appuie sur un modèle probabiliste sous-jacent, comme c'est le cas en analyse statistique implicative.

On commence d'abord par rappeler le principe de l'utilisation d'un modèle probabiliste pour l'extraction de connaissances. On présente, ensuite, ce qu'est une approche en extraction de règles s'appuyant sur des modèles probabilistes locaux, à travers l'exemple de l'analyse statistique implicative, et on en explicite les manquements. Enfin, on présente des modèles probabilistes globaux et on discute des possibilités d'une approche en extraction de règles s'appuyant sur des tels modèles.

3.1 Modèles probabilistes pour l'extraction de connaissances

Mesurer l'intérêt d'un motif en extraction de connaissances à travers un modèle probabiliste correspond, en règle générale, à mesurer l'éloignement entre une observation faite sur le motif et ce qui était initialement prévu par le modèle. Comme on se place dans un cadre probabiliste, on peut choisir de mesurer cet éloignement de manière à ce que cela représente un degré d'étonnement par rapport à la distribution décrite par le modèle. Un modèle statistique est un modèle probabiliste qui possède deux particularités. D'abord, le modèle est défini par un certain nombre d'observations préalables sur les données. Ensuite, l'observation faite sur le motif (dont on mesure l'éloignement au modèle) correspond à l'observation de la fréquence de ce motif dans les données. Ainsi, on peut mesurer le degré d'étonnement que l'on a à observer la fréquence d'apparition d'un motif au regard d'informations que l'on a déjà collectées sur les données et qui définissent le modèle. Un étonnement marqué signifie que ces informations (ou connaissances puisqu'il s'agit bien ici d'extraction de connaissances) ne permettent pas d'expliquer de manière suffisante la fréquence du motif en question. Dans ce cas, le motif peut être extrait, avec sa fréquence et la mesure de l'étonnement correspondante, de manière à compléter nos connaissances sur les données et réévaluer le modèle afin qu'il intègre ces nouvelles connaissances.

3.2 Modèles probabilistes locaux

En analyse statistique implicative, quand on cherche à calculer l'intensité d'implication d'une règle $a \to b$, le motif que l'on observe dans les données est celui de l'expression logique d'attributs $a \wedge \bar{b}$ qui correspond à la négation de l'implication de a vers b. On considère alors l'un des quatre modèles probabilistes suivants afin de confronter la fréquence observée $f_{a\wedge\bar{b}}$ à ce qui est prévu par le modèle. Il s'agit du modèle binomial, du modèle hypergéométrique, du modèle de Poisson ou du modèle gaussien. Malgré leurs différences, ces modèles reposent tous sur un même modèle probabiliste sous-jacent dans lequel on considère que, étant donnée une observation, (1) la probabilité p_a que l'on y observe a est égale à f_a, (2) la probabilité p_b que l'on y observe b est égale à f_b et (3) ces deux événements sont indépendants (la différence entre les quatre modèles portant sur les distributions du nombre d'observations de a, du nombre d'observations de b, et du nombre total d'observations). C'est d'ailleurs le modèle de base de n'importe quel test d'indépendance de deux attributs à partir de leur observation dans un échantillon. Maintenant, si l'on considère une règle $a \wedge b \to c$, le modèle sous-jacent correspondant dit qu'étant donnée une observation, la probabilité $p'_{a\wedge b}$ qu'on y observe $a \wedge b$ est égale à $f_{a\wedge b}$. Or, si $f_{a\wedge b} \neq f_a f_b$, alors cela n'est pas compatible avec l'hypothèse d'indépendance du modèle correspondant à $a \to b$: on a $p'_{a\wedge b} \neq p_{a\wedge b}$. Il ne peut donc pas exister, dans ce cas, de modèle plus global qui recouperait ces deux modèles locaux.

De manière générale, si l'on considère l'ensemble de tous les modèles locaux correspondant à chacune des règles de type $a \to b$, où a et b sont des conjonctions d'attributs, alors il existe un modèle global recoupant l'ensemble de ces modèles si et seulement si les fréquences relatives des observations dans les données sont égales aux probabilités données par le modèle d'indépendance mutuelle de tous les attributs, sauf éventuellement pour la fréquence relative de l'observation de la conjonction de tous les attributs. En pratique, cela n'arrive quasiment jamais dès que l'on considère plus de deux attributs. On note, en passant, qu'un modèle global tel qu'il est envisagé ici est entièrement défini par une mesure de probabilité sur l'espace probabilisable engendré par

les attributs (et possède donc $2^p - 1$ degrés de libertés). Ainsi, si l'on s'appuie sur ces modèles statistiques locaux pour faire de l'extraction de règles, on fait de multiples tests statistiques sur un même jeu de données mais avec des hypothèses différentes et incompatibles pour ces différents tests. Cela représente une fragilité dans les fondements même de la méthode et ainsi une limite dans la portée des résultats obtenus. Or, si un certain nombre de travaux ont traité spécifiquement la question des erreurs liées aux tests statistiques multiples et les moyens de correction de ces erreurs, le problème de l'incompatibilité des hypothèses des tests en est singulièrement absente (Bay et Pazzani, 2001, Lallich et al., 2007, Ge et al., 2003, Liu et al., 2011, et Webb, 2007). D'une manière générale, on peut voir les différents problèmes liés à l'utilisation de tests statistiques multiples comme étant reliés à un oubli systématique des étapes réalisés et des connaissances extraites, dès leur réalisation et dès leur extraction, et ceci tout au long du processus d'extraction. Pour chaque nouvelle règle, on recommence comme si l'on n'avait rien appris. On ne réévalue jamais les modèles, ni pour qu'ils soient compatibles en termes d'hypothèses, ni pour qu'ils intègrent les connaissances déjà acquises. C'est pour répondre à cette problématique générale de l'oubli que des modèles globaux évolutifs ont été développés (voir 3.3.1 et 3.3.2).

3.3 Modèles probabilistes globaux

La solution évidente pour répondre au problème lié aux hypothèses incohérentes est. de passer par un modèle global, c'est-à-dire un modèle probabiliste général pour le nombre d'occurrences de toutes les expressions logiques d'attributs, et d'en considérer des restrictions locales lorsque c'est nécessaire. Pour autant, ce n'est pas une solution miracle car, si elle répond à des problèmes théoriques, elle s'accompagne également d'un certain nombre de difficultés techniques dans sa mise en pratique.

On reprend le principe de la séparation du modèle probabiliste en deux parties, comme en analyse statistique implicative, avec d'une part une distribution sur le nombre d'observations et de l'autre un modèle probabiliste sous-jacent qui définisse chaque probabilité p_a, étant donnée une observation quelconque, d'observer l'expression logique a. La question de la définition du modèle probabiliste global se déplace alors entièrement au niveau de la définition de ce modèle probabiliste sous-jacent. Une fois le modèle défini, on peut, par exemple, facilement redéfinir l'indice d'implication (modèle gaussien) pour une règle $a \to b$ de la manière suivante :

$$\varphi(a \to b) = \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{+\infty} \mathrm{e}^{-\frac{t^2}{2}}\, dt \quad \text{où} \quad q(a,\bar{b}) = \frac{n f_{a\wedge\bar{b}} - p_{a\wedge\bar{b}}}{\sqrt{n p_{a\wedge\bar{b}}}}.$$

Employer un modèle global peut donc s'avérer tout aussi simple que de passer par des modèles locaux, à condition d'avoir défini au préalable le modèle global en question. Or c'est bien là que repose une des grandes difficultés de l'approche globale par rapport à une approche locale : il est a priori plus aisé de définir des modèles locaux que de définir un modèle global.

3.4 Propriétés des modèles globaux

Avant de présenter quelques exemples précis de modèles globaux, voici des éléments de classification des modèles.

Modèles statiques – modèles évolutifs — Si l'utilisation d'un même modèle global permet de donner une cohérence au niveau des hypothèses des tests réalisés pour chacune

des règles, cela ne répond pas pour autant au problème de redondance parmi les règles extraites ni au problème de l'apparition de fausses découvertes lorsque l'on réalise des tests multiples sur la base d'un même modèle. C'est pourquoi il est préférable de pouvoir faire évoluer le modèle de manière à y intégrer une partie, du moins, des connaissances extraites des données. Certains modèles n'ont pas été développés en ce sens et ne possèdent pas ce caractère évolutif, on parlera ici de modèles statiques par opposition à la notion de modèle évolutif. Cette distinction est la même que celle qui est présentée dans Vreeken et Tatti (2014) sous les termes de modèles statiques et modèles dynamiques mais l'on a préféré le terme évolutif car il incorpore l'idée d'une certaine progression dans les changements qui s'opèrent sur le modèle.

Modèles objectifs – modèles subjectifs — De la même manière que l'on peut définir des mesures d'intérêt objectif et des mesures d'intérêt subjectif (Geng et Hamilton, 2006), on peut imaginer des modèles probabilistes pour l'intérêt qui intègrent dans leur définition une part de la subjectivité d'un utilisateur (Jaroszewicz et Simovici, 2004, proposent un modèle subjectif pour l'intérêt défini à partir d'un réseau bayésien). Toutefois, les modèles proposés par la suite sont uniquement des modèles statistiques objectifs. Ainsi les modèles présentés sont construits, d'une part, à partir d'un ensemble de paramètres qui correspondent chacun à un relevé statistique dans les données et, d'autre part, dans l'idée qu'en dehors de ces paramètres, la modélisation cherche à traduire une absence de toute autre connaissances des données (ce qui passe, pour chacun des modèles proposés ici, par l'utilisation dans la construction du modèle de distributions uniformes sur des ensembles satisfaisant des contraintes liées aux paramètres).

3.5 Exemples de modèles globaux

Modèles d'indépendance mutuelle — Il s'agit du modèle global le plus simple et le plus classique. Les paramètres sur lesquels il s'appuie sont les fréquences relatives empiriques des attributs du problème et il est construit sur l'hypothèse d'indépendance mutuelle entre chacun de ces attributs. Ce modèle se calcule très facilement et possède d'ailleurs un avantage notable sur ce plan : il n'est pas nécessaire de calculer entièrement le modèle de manière explicite pour connaître la valeur qu'il prend en une expression logique d'attributs, celle-ci peut s'exprimer directement. Toutefois ce modèle est très limité car il ne repose que sur p informations et il est, a priori, statique (on verra par la suite qu'il peut être considéré comme un cas particulier du modèle d'entropie maximale / modèle d'indépendance mutuelle contrainte et que, dans ce cas, on peut le faire évoluer).

Modèles des marges fixes — Le modèle des marges fixes est un modèle statistique qui a été envisagé dans le domaine de l'apprentissage statistique et de l'extraction de connaissances, notamment pour des applications en écologie statistique et en psychométrie (Cobb et Chen, 2003, Gionis et al., 2007, Hanhijärvi et al., 2009, et Vreeken et Tatti, 2014). Ce modèle peut être interprété comme un raffinement du modèle d'indépendance mutuelle. Les paramètres qui le définissent sont les fréquences relatives empiriques des attributs du problème ainsi que du nombre d'attributs observés pour chacun des individus. Ces paramètres correspondent respectivement aux sommes des coefficients de chacune des colonnes de la matrice binaire D représentant les données et aux sommes des coefficients de chacune de ses lignes. Il y en a donc $n+p$. Il existe deux possibilités pour définir un tel modèle : une définition analytique simple via le modèle de Rasch (Rasch, 1960) qui est en fait le modèle de l'entropie maximale pour le problème des marges fixes ; ou l'approximation d'un modèle exact via des méthodes de *randomization*. Pour ce deuxième cas, on peut utiliser le fait qu'il est possible de passer

de n'importe quelle matrice binaire satisfaisant les contraintes des marges à n'importe quelle autre matrice binaire satisfaisant ces mêmes contraintes par une série de permutations particulières de leurs coefficients afin de générer aléatoirement (au sens d'une distribution uniforme) une telle matrice via des méthodes MCMC (Chaînes de Markov – Monte Carlo). Bien que la méthode, permette d'obtenir une bonne approximation de la distribution exacte, elle est coûteuse en termes de temps de calcul dès que les données sont trop nombreuses ($n \times p < 10\ 000$ environ). Le modèle de Rasch, quant à lui, est bien plus simple à définir mais ne peut être considéré comme une bonne modélisation qu'à condition que le nombre d'observations *n* et le nombre d'attributs *p* soient tous les deux très grands. Dans tous les cas, il faut noter qu'il s'agit bien évidemment ici de modèles statiques. Ils ne permettent ainsi que d'apprécier l'étonnement que l'on peut avoir à découvrir certains motifs étant données les marges du problème. Enfin, il est important de remarquer que cette modélisation traite de manière totalement asymétrique la présence et l'absence d'un attribut. Cela peut être particulièrement adapté dans certains contextes. En didactique, par exemple, pour l'extraction de connaissances d'une base de données détaillant les réponses (positives ou négatives) d'étudiants à un examen, le fait d'avoir su répondre à une question n'est pas interchangeable avec le fait de s'être trompé, et le nombre de questions auxquelles un étudiant a répondu (i.e. les marges pour les lignes) peut être considéré comme un indicateur pertinent. Par contre, si on considère l'exemple d'un attribut de type sexe (masculin ou féminin), il est, a priori, très important que le choix de la représentation de l'une de ces modalités par un 0 et de l'autre par un 1 n'ait aucune incidence sur les conclusions que l'on peut tirer du modèle. En résumé, ce modèle est statique mais permet tout de même une extraction plus précise que le modèle d'indépendance mutuelle, à condition de vérifier que le modèle est bien adapté au problème qu'il modélise et de ne pas considérer une base de données trop volumineuse.

Modèles d'entropie maximale / d'indépendance mutuelle contrainte — Les modèles d'entropie maximale permettent de décrire la distribution de probabilité la plus objective possible étant donnée une série de contraintes sur ce modèle (Jaynes, 1982, Berger et al., 1996, Mannila et al., 1999, Pavlov et al., 2000, Pavlov et al., 2003, Vreeken et Tatti, 2014). Le terme de modèle d'indépendance mutuelle contrainte correspond à une réinvention de ce concept pour le cas spécifique où les contraintes sont données par les fréquences relatives empiriques d'expressions logiques d'attributs ou itemsets généralisés (Delacroix et al., 2015, 2017a, 2017b). Pour définir un tel modèle, on peut donc prendre comme paramètres n'importe quel ensemble de fréquences relatives empiriques d'expressions logiques d'attributs. Par exemple, on peut fixer $p_{a_1} = f_{a_1}$, $p_{a_2} = f_{a_2}$, $p_{a_1 \wedge \overline{a_3}} = f_{a_1 \wedge \overline{a_3}}$ et $p_{a_2 \wedge a_3} = f_{a_2 \wedge a_3}$ comme contraintes et le modèle d'indépendance mutuelle contrainte permet de définir les valeurs restantes non fixées. C'est une généralisation naturelle du modèle d'indépendance mutuelle qui permet une bien plus grande flexibilité sur le choix des connaissances définissant le modèle. Il est bien évidemment évolutif car on peut y intégrer toute nouvelle connaissance sur la fréquence relative d'une expression logique.

4 Processus d'extraction locaux – processus d'extraction globaux

Comme annoncé en introduction, les différentes approches en extraction de règles peuvent être généralement caractérisées d'approches locales à plusieurs titres. En

particulier, ce n'est pas parce qu'une approche s'appuie sur un modèle probabiliste global pour évaluer l'intérêt des règles à extraire qu'il repose nécessairement sur un processus d'extraction global de ces règles. En effet, par définition, toutes les mesures d'intérêt des règles dont il est question dans la littérature (Vaillant et al., 2004, Lallich et al., 2007, Lenca et al., 2008, Geng et Hamilton, 2009, Le Bras, 2011, Wu et al., 2012) sont surtout des mesures d'intérêt d'une règle. Elles permettent ainsi d'évaluer l'intérêt d'une règle seule par rapport au modèle utilisé dans ce contexte. Que ce modèle soit local (défini par des paramètres propres à la règle dont l'intérêt est mesuré) ou global (défini par des paramètres communs à toutes les règles), le processus d'extraction d'une règle demeure lui fondamentalement localisé autour de cette règle car l'intérêt de celle-ci ne dépend que de cette règle et du modèle, l'intérêt des autres règles étant évalué séparément. En contraste, un processus global d'extraction de règles est tout à fait envisageable à condition de partir d'un modèle global, défini par un certain nombre de règles, et d'évaluer l'intérêt, non pas d'une ou de plusieurs règles par rapport à ce modèle, mais du modèle en soi et donc, par là, des règles qui le définissent. Dans le cas d'une approche statistique, l'approche globale rend obsolète la question des tests multiples réalisés pour chaque règle (Bay et Pazzani, 2001, Lallich et al., 2007, Ge et al., 2003, Liu et al., 2011, et Webb, 2007) puisque cette multiplicité de tests locaux est remplacée par un test global. Par ailleurs, un test global permet d'atténuer les difficultés inhérentes à l'utilisation de l'intensité d'implication pour des données sur un large échantillon de données (n élevé) qui tend alors vers un indicateur binaire (Delacroix, 2014). De la même façon que pour les modèles, une approche via un processus d'extraction local est plus facilement mise en pratique mais une approche via un processus d'extraction global est mieux fondée théoriquement. On propose donc plusieurs approches différentes locales puis globales correspondant chacune à des niveaux différents de compromis entre, d'un côté, une utilisation raisonnable des ressources de calcul (à la fois en temps et en mémoire) et, de l'autre, la volonté de faire reposer les résultats obtenus sur des fondements théoriques les plus solides possible.

4.1 Processus d'extraction locaux

Sans rentrer dans le détail des processus, on présente deux types de processus locaux différents, ainsi que plusieurs sous-types, qui peuvent être envisagés dans le contexte de l'extraction de règles. Dans tous les cas, on suppose que le ou les modèles considérés sont des modèles probabilistes globaux.

4.2 Processus à modèle statique

Il s'agit du type le plus basique. On ne considère qu'un seul modèle et on extrait toutes les règles dont la mesure de l'intérêt par rapport au modèle dépasse un certain seuil. Ainsi, l'intérêt de chaque règle est considéré individuellement et de manière identique par rapport au modèle (et donc à la connaissance des paramètres qui le définissent). Si l'on considère une mesure d'intérêt fondée sur un test statistique, comme c'est le cas de l'intensité d'implication, alors on effectue, de facto, une multitude de tests statistiques ce qui entraîne les difficultés précitées (Bay et Pazzani, 2001, Lallich et al., 2007, Ge et al., 2003, Liu et al., 2011, et Webb, 2007).

4.3 Processus à modèle évolutif

L'idée, ici, est de faire évoluer le modèle de manière à intégrer les connaissances qui ont étonné l'utilisateur au fur et à mesure du processus d'extraction. On propose, ci-dessous,

différentes manières d'aborder un tel processus. Ces dernières ne constituent en aucun cas une liste exhaustive mais a pour simple objectif de faire réfléchir sur les façons dont on peut moduler les approches. Les modèles considérés dans la suite sont les modèles d'entropie maximale / d'indépendance mutuelle contrainte décrits précédemment (voir 3.3.1).

Parcours simple avec ajout des règles intéressantes — Le principe ici est de considérer chacune des règles l'une après l'autre (l'ordre de parcours étant laissé au choix de l'utilisateur). Si une règle $a \rightarrow b$ est estimée étonnante par rapport aux données du modèle (i.e. si son intensité d'implication par rapport au modèle dépasse un certain seuil), alors on réévalue le modèle pour y intégrer comme paramètre la fréquence empirique qui est à la base de cette étonnement (i.e. $f_{a \wedge b}$). On peut commencer ce processus en partant du modèle neutre (correspondant à une absence totale de connaissances) ou du modèle d'indépendance mutuelle. À la fin du processus, les règles extraites sont toutes celles qui ont été intégrées au modèle. L'avantage de cette approche est qu'elle ne nécessite qu'un seul parcours de l'ensemble des règles envisagées. Par contre, elle dépend du choix de l'ordre de parcours des règles. Par ailleurs, il n'est pas impossible qu'une règle qui a déjà été testée négativement (c'est-à-dire que la mesure de son intérêt par rapport au modèle au moment où elle a été testée était inférieure au seuil fixé) puisse être considérée étonnante par rapport à une évolution postérieure du modèle, or cela ne sera jamais testé.

Ajout de la règle la plus intéressante — Ici, à chaque étape, on fait évoluer le modèle en y intégrant la règle pour laquelle la mesure de l'intérêt par rapport au modèle est à la fois la plus forte et au-dessus du seuil fixé. Une des difficultés majeures de cette approche est qu'elle peut être particulièrement coûteuse en termes de puissance de calcul puisqu'il faut, a priori, parcourir l'ensemble des règles à chaque étape afin de déterminer la plus intéressante d'entre elles.

Ajout de la règle la plus intéressante par couche — Ceci est une adaptation du cas précédent de manière à ne pas avoir à considérer l'ensemble de toutes les règles à chaque étape mais simplement un sous-ensemble (la couche). On extrait alors les règles en les intégrant au modèle couche par couche. Par exemple, on peut définir chaque couche C_m comme étant l'ensemble des règles $a_1 \wedge ... \wedge a_i \rightarrow b_1 \wedge ... \wedge b_j$ telles que $i, j \geq 1$ ainsi que $i + j = m$. On commence alors par C_2 et on fait évoluer le modèle en y intégrant, à chaque étape, la règle $a_i \rightarrow b_j$ dont la mesure de l'intérêt est la plus élevée, tout en étant au-dessus du seuil, et cela jusqu'à ce qu'il n'existe plus de règles dont la mesure de l'intérêt dépasse le seuil dans cette couche. On passe alors à la couche C_3 et on répète le processus jusqu'à arriver à la dernière couche (ou une couche jugée suffisamment profonde).

Ajout par paquet des règles — Il s'agit ici d'une adaptation possible des cas précédents. Au lieu de faire évoluer le modèle règle par règle, on le fait évoluer en y intégrant plusieurs règles à la fois. Cela permet de limiter le nombre de fois qu'il est nécessaire de redéfinir le modèle ce qui représente normalement une part non négligeable de l'utilisation des ressources de calcul. Ainsi, on peut combiner cela avec l'approche précédente, par exemple, en extrayant à chaque étape les dix règles les plus intéressantes de la couche courante.

4.4 Processus d'extraction globaux

Comme cela a été décrit précédemment, il ne s'agit pas dans le cas d'un processus d'extraction global d'évaluer l'intérêt d'une règle par rapport à un modèle mais d'évaluer l'intérêt d'un ensemble de règles définissant un modèle. On part pour cela du principe qu'un ensemble de règles est intéressant s'il permet, à partir de peu de règles, de décrire de manière adéquate les données. Dans le cadre de l'extraction de motifs à l'aide de modèles d'indépendance mutuelle contrainte, plusieurs solutions ont été envisagées pour évaluer ceci dont un test du χ^2 d'adéquation du modèle aux données, ainsi que deux mesures non statistiques de l'éloignement du modèle aux données, normalisées de manière à pénaliser les modèles définis par le plus de paramètres (Delacroix et al., 2017b).

Dans tous les cas, on parlera de mesure d'adéquation du modèle aux données. Il faut noter que, quelques soit la mesure choisie, la comparaison du modèle aux données doit se faire sur l'ensemble des 2^p atomes de l'espace probabilisé sous-jacent (i.e. les probabilités p_a où a est une conjonction composée, pour chacun des p attributs, de cet attribut ou de sa négation). Cela limite sérieusement le nombre d'attributs qui peuvent être considérés puisqu'il faut donc déterminer, pour chaque modèle défini dans le processus, les valeurs qu'il prend en ces 2^p atomes.

Le meilleur modèle — L'approche la plus directe en théorie est de prendre le meilleur modèle, c'est-à-dire celui dont la mesure d'adéquation est la plus élevée parmi tous les modèles. Les règles à extraire sont alors celles qui définissent ce modèle. Cette solution est, en pratique, quasi irréalisable car il faudrait, pour en connaître le meilleur, mesurer l'adéquation aux données de chaque modèle qui peut être défini par un sous-ensemble de l'ensemble des règles (il y en a, a priori, 2^{2^p} distincts).

Une approximation via un algorithme glouton — Si l'approche précédente n'est pas réalisable en pratique, la solution que l'on obtient demeure tout de même celle que l'on souhaite atteindre. Comme il s'agit d'un problème de recherche de maximum global, on peut essayer de l'atteindre, ou du moins de l'approcher par un maximum local, via un algorithme glouton. Pour cela, on part du modèle neutre et on le fait évoluer règle par règle, en choisissant à chaque fois la règle qui améliore le mieux la mesure d'adéquation du modèle jusqu'à obtenir un modèle correspondant à un maximum local (c'est-à-dire que si on rajoute une règle à ce modèle, la mesure de son adéquation aux données diminue). Bien que cette approche est moins demandeuse en ressources de calcul que la précédente, elle reste, en pratique, tout de même trop coûteuse.

Une approximation via un algorithme glouton par couche — Il s'agit d'une adaptation de l'approche ci-dessus, semblable dans l'idée au processus d'ajout de règle par couche décrit précédemment (voir 4.1.2). Dans ce cas, on cherche la règle à ajouter dans une couche et on passe à la couche suivante seulement s'il n'y a plus de règle dans la couche qui permette d'améliorer la mesure de l'adéquation du modèle aux données. On peut également envisager de s'arrêter dès que l'on atteint une certaine couche que l'on estime suffisamment profonde.

Dans le cadre de l'extraction d'itemsets intéressants, des essais ont été menés sur des bases de données tronquées de manière à ne garder que 3 ou 4 attributs et ainsi pouvoir déterminer exactement le meilleur modèle d'indépendance mutuelle contrainte en termes de mesure d'adéquation aux données. Dans les différents cas traités, les deux approximations décrites ici ont permis d'obtenir la meilleure solution (Delacroix et al., 2017b). Ainsi, la dernière approche qui est largement moins demandeuse en termes de ressources de calcul semble devoir être privilégiée.

5 Approches intermédiaires

Si l'utilisation d'approches locales, tant au niveau du modèle que du processus d'extraction, est l'une des causes principales du problème de la redondance dans le domaine de l'extraction de motifs, les approches globales sont difficiles à mettre en œuvre car elles sont très coûteuses en termes de puissance de calcul. Les modèles d'entropie maximale tels que le modèle d'indépendance mutuelle contrainte sont certes les modèles permettant de décrire le plus précisément les données mais ce sont aussi ceux qui requiert le plus de ressources de calcul (Pavlov, 2003). L'utilisation de processus globaux font ensuite s'additionner les besoins en puissance de calculs. Toutefois, des approches intermédiaires sont envisageables et représentent vraisemblablement une perspective très intéressante pour la recherche à venir en extraction de motifs. En effet, il existe un monde de possibilités entre, d'une part, les approches dans lesquelles l'extraction se fait de manière extrêmement localisée autour de chacune des règles à extraire et qui ont prédominé jusqu'à présent et, d'autre part, les approches complètement globales proposées dans cet article. La découverte de ce monde reste pour le moment largement à faire et on ne s'y aventurera pas dans cet article. Toutefois, une piste est proposée dans le but d'inciter à la réflexion dans ce domaine.

Une approche intermédiaire — On commence par découper le problème global en sous-problèmes ne comprenant que 5 attributs. Cela correspond à autant de sous-problèmes qu'il y a de combinaisons de 5 éléments parmi p. Si on prend $p = 100$, on en obtient environ 75 millions, ce qui reste raisonnable. Chacun de ces sous-problèmes correspond au problème d'extraire des règles intéressantes depuis la base de données tronqué de manière à ne garder que 5 attributs. On détermine donc pour chacun d'entre eux l'ensemble des règles intéressantes comprenant ces 5 attributs (ou moins) en considérant l'algorithme glouton par couche décrit précédemment et en fixant un seuil commun à tous les sous-problèmes, de manière à ce que, si le seuil n'est pas atteint, on ne retienne aucune règle. Enfin, on comptabilise pour chaque règle la fréquence à laquelle elle a été considérée comme intéressante relativement au nombre de sous-problèmes auxquels elle peut être rattachée. Cette fréquence relative peut alors être considérée comme un indicateur de l'intérêt de la règle. Une approche similaire est décrite dans le contexte de l'extraction d'itemsets intéressants et testée dans Delacroix et al. (2017b).

6 Conclusion

On a cherché, dans cet article, à mettre en lumière le caractère local de l'approche de fouille développée en analyse statistique implicative, à la fois au niveau des modèles pour mesurer l'intensité d'implication d'une quasi-implication et au niveau du processus d'extraction s'appuyant sur cette intensité d'implication. Ce caractère local est d'ailleurs largement partagé par d'autres approches en extraction de règles et il est vraisemblablement la cause d'un certain nombre de problèmes liés à la surabondance et la redondance des règles extraites. Or d'autres approches plus globales sont possibles et restent, pour le moment, relativement peu explorées. Pour faire évoluer les pratiques en matière d'extraction de règles de ces approches locales vers des approches plus globales, il subsiste deux obstacles majeurs. D'une part, les approches globales sont toujours plus gourmandes en termes de ressources de calcul et, d'autre part, elles reposent sur des outils mathématiques complexes qui peuvent nécessiter un certain temps d'appropriation. Même s'il est envisageable de passer par des approches semi-globales pour écarter les

problèmes de puissance de calcul, il est important que des chercheurs s'intéressent à des approches plus globales et les mettent en pratique afin qu'elles se développent. La communauté de l'analyse statistique implicative paraît alors tout indiquée pour relever ce défi. En effet, ses membres ont fait le choix d'une approche non triviale de la question de l'extraction de règles qui focalise avant tout sur la qualité des informations qui sont extraites plutôt que sur la quantité des données qui peuvent être traitées. Aborder l'extraction de quasi-implications via une approche statistique plus globale serait ainsi en total cohérence avec ce principe fondateur en analyse statistique implicative. C'est la raison pour laquelle on propose le développement intégral d'une analyse statistique implicative globale en tant que nouveau défi à relever pour le prochain rendez-vous de la communauté de l'analyse statistique implicative. En particulier, la solution à apporter à ce défi devra intégrer des outils de représentation graphique des connaissances extraites semblables à ceux qui existent aujourd'hui en analyse statistique implicative et qui permettent une lisibilité aisée de celles-ci dans leur ensemble (comme le graphe implicatif).

Références

[1] Agrawal, C. C., & Han, J. (Eds.). (2014). *Frequent pattern mining*. Springer.

[2] Agrawal, R., & Srikant, R. (Sept. 1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).

[3] Bay, S. D., & Pazzani, M. J. (2001). Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, *5*(3), 213-246.

[4] Berger, A. L., Pietra, V. J. D., & Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. Computational linguistics, 22(1), 39-71.

[5] Cobb, G. W., & Chen, Y. P. (2003). An application of Markov chain Monte Carlo to community ecology. The American Mathematical Monthly, 110(4), 265-288.

[6] Delacroix, T. (2014). A renewed approach to the foundations of SIA theory: generalizing SIA to incorporate multiple behavior hypotheses. Thoughts on the implicative intensity. Educação Matemática Pesquisa, 16(3).

[7] Delacroix, T., Boubekki, A., Lenca, P., & Lallich, S. (2015). Constrained independence for detecting interesting patterns. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA),* (pp. 1052-1061). IEEE.

[8] Delacroix, T., Lenca, P., & Lallich, S. (2017). Computing the mutual constrained independence model. ASMDA 2017.

[9] Delacroix, T., Lenca, P., & Lallich, S. (2017). What to expect from a set of itemsets. À paraître.

[10] Fournier-Viger, P., Lin, J. C. W., Vo, B., Chi, T. T., Zhang, J., & Le, H. B. (2017). A survey of itemset mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.

[11] Friedman, J., Hastie, T., & Tibshirani, R. (2008). *The elements of statistical learning* (Vol. 1, 2nd edition). Springer, Berlin: Springer series in statistics.

[12] Ge, Y., Dudoit, S., & Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test*, *12*(1), 1-77.

[13] Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, *38*(3), 9.

[14] Gionis, A., Mannila, H., Mielikäinen, T., & Tsaparas, P. (2007). Assessing data mining results via swap randomization. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(3), 14.

[15] Gras, R., & Couturier, R. (2013). Spécificités de l'Analyse Statistique Implicative par rapport à d'autres mesures de qualité de règles d'association. *Educação Matemática Pesquisa*, *15*(2).

[16] Gras, R., Couturier, R., Blanchard, J., Briand, H., Kuntz, P., & Peter, P. (2004). Quelques critères pour une mesure de qualité de règles d'association. *Revue des nouvelles technologies de l'information RNTI E-1*, 3-30.

[17] Gras, R., Regnier, J. C., & Guillet, F. (2009). Analyse statistique implicative : Une méthode d'analyse de données pour la recherche de causalités (p. 510). Cépaduès Editions.

[18] Gras, R., Régnier, J. C., Marinica, C., & Guillet, F. (2013). L'analyse statistique implicative Méthode exploratoire et confirmatoire à la recherche de causalités (p. 522). Cépaduès Editions.

[19] Han, J., Pei, J., & Kamber, M. (2012). *Data mining: concepts and techniques* (3rd edition). Elsevier.

[20] Hanhijärvi, S., Ojala, M., Vuokko, N., Puolamäki, K., Tatti, N., & Mannila, H. (2009). Tell me something I don't know: randomization strategies for iterative data mining. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 379-388). ACM.

[21] Jaroszewicz, S., & Simovici, D. A. (2004). Interestingness of frequent itemsets using bayesian networks as background knowledge. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 178-186). ACM.

[22] Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. Proceedings of the IEEE, 70(9), 939-952.

[23] Klösgen, W., & Zytkow, J. M. (2010). *Handbook of data mining and knowledge discovery* (2nd edition). Oxford University Press, Inc..

[24] Kotsiantis, S., & Kanellopoulos, D. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, *32*(1), 71-82.

[25] Lallich, S., Teytaud, O., & Prudhomme, E. (2007). Association rule interestingness: Measure and statistical validation. In *Quality measures in data mining* (pp. 251-275). Springer Berlin Heidelberg.

[26] Lallich, S., Vaillant, B., & Lenca, P. (2007). A probabilistic framework towards the parameterization of association rule interestingness measures. *Methodology and Computing in Applied Probability*, *9*(3), 447-463.

[27] Le Bras, Y. (2011). Contribution à l'étude des mesures de l'intérêt des règles d'association et à leurs propriétés algorihmiques (Thèse de doctorat).

[28] Lenca, P., Meyer, P., Vaillant, B., & Lallich, S. (2008). On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European journal of operational research*, *184*(2), 610-626.

[29] Lenca, P., Meyer, P., Vaillant, B., Picouet, P., & Lallich, S. (2004). Evaluation et analyse multicritère des mesures de qualité des règles d'association. *Revue des Nouvelles Technologies de l'Information (Mesures de Qualité pour la Fouille de Données)*, 219-246.

[30] Liu, G., Zhang, H., & Wong, L. (2011). Controlling false positives in association rule mining. *Proceedings of the VLDB Endowment*, *5*(2), 145-156.

[31] Mannila, H., Pavlov, D., & Smyth, P. (1999). Prediction with local patterns using cross-entropy. In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 357-361). ACM.

[32] Pavlov, D., Mannila, H., & Smyth, P. (2000). Probabilistic models for query approximation with large sparse binary data sets. In Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence (pp. 465-472). Morgan Kaufmann Publishers Inc..

[33] Pavlov, D. N., Mannila, H., & Smyth, P. (2003). Beyond independence: Probabilistic models for query approximation on binary transaction data. IEEE Transactions on Knowledge and Data Engineering, 15(6), 1409-1421.

[34] Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Danmarks paedagogiske Institut.

[35] Vaillant, B., Lenca, P., & Lallich, S. (2004, October). A clustering of interestingness measures. In Discovery science (Vol. 3245, pp. 290-297).

[36] Vreeken, J., & Tatti, N. (2014). Interesting patterns. In *Frequent pattern mining* (pp. 105-134). Springer International Publishing.

[37] Webb, G. I. (2007). Discovering significant patterns. *Machine learning*, *68*(1), 1-33.

[38] Wu, J., Zhu, S., Xiong, H., Chen, J., & Zhu, J. (2012). Adapting the right measures for pattern discovery: A unified view. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *42*(4), 1203-1214.

[39] Zhao, Q. & Bhowmick, S. S. (2003). Association Rule Mining: A Survey. *CAIS* (Technical report 2003116). Nanyang Technological University, Singapore.