Mesure formelle de la robustesse des règles d'association

Yannick Le Bras[*,***] Patrick Meyer[*,***] Philippe Lenca[*,***] Stéphane Lallich[**],

[*]Institut Télécom, Télécom Bretagne,
UMR CNRS 3192 Lab-STICC,
Technopôle Brest Iroise CS 83818, 29238 Brest Cedex 3
{yannick.lebras || patrick.meyer || philippe.lenca}@telecom-bretagne.eu
[**]Université de Lyon, Laboratoire ERIC, Lyon 2, France
stephane.lallich@univ-lyon2.fr
[***]Université européenne de Bretagne, France

Résumé. Nous proposons dans cet article une définition formelle de la robustesse pour les règles d'association, s'appuyant sur une modélisation que nous avons précédemment définie. Ce concept est à notre avis central dans l'évaluation des règles et n'a à ce jour été que très peu étudié de façon satisfaisante. Il est crucial car malgré une très bonne évaluation par une mesure de qualité, une règle peut être très fragile par rapport à des variations légères des données. La mesure de robustesse que nous proposons dépend de la mesure de qualité utilisée pour évaluer les règles et du seuil d'acceptation minimal. Il est alors possible à partir de ces deux seuls éléments et de la valeur prise par la règle sur la mesure d'évaluer sa robustesse. Nous présentons plusieurs propriétés de cette robustesse, montrons sa mise en œuvre et illustrons celle-ci par les résultats d'expériences sur plusieurs bases de données pour quelques mesures. Nous donnons ainsi un nouveau regard sur la qualification des règles.

1 Introduction

Depuis sa définition originale par Agrawal et al. (1993) et l'algorithme APRIORI (Agrawal et Srikant, 1994) les motifs fréquents et les règles d'association ont suscité de très nombreux travaux algorithmiques (voir par exemple les synthèses proposées par Hipp et al. (2000), Goethals (2005) et Han et al. (2007)). Malheureusement, les nombreux algorithmes, déterministes et performants, du type d'APRIORI produisent de trop grandes quantités de règles. Par ailleurs l'intérêt d'une large proportion de ces règles est souvent discutable : Brin et al. (1997) mettent par exemple en évidence ce problème en montrant l'importance de l'étude des corrélations en complément du couple support confiance, c'est à dire la comparaison de la confiance à la probabilité du conséquent plutôt qu'à un seuil fixe. Face à ce problème à la fois quantitatif et qualitatif de nombreux efforts ont porté sur l'évaluation de l'intérêt des règles d'association afin de sélectionner idéalement toutes les règles intéressantes et uniquement celles-ci.

Une méthode très populaire pour évaluer l'intérêt des règles d'association consiste à quantifier cet intérêt à l'aide de mesures objectives (Piatetsky-Shapiro, 1991; Hilderman et Hamilton,

2000). Définies à partir de la contingence des règles, les mesures objectives permettent de classer les règles, mais les classements peuvent varier fortement d'une mesure à l'autre (Vaillant et al. (2004)). Les nombreuses mesures existantes et les propriétés de ces mesures ont ainsi suscité un grand nombre de travaux. Nous renvoyons le lecteur aux synthèses proposées par Lenca et al. (2004), Gras et al. (2004), Geng et Hamilton (2006), Lallich et al. (2007), Lenca et al. (2007), Geng et Hamilton (2007), Lenca et al. (2008), Suzuki (2008) et Guillaume et al. (2010).

On rappelle qu'une règle d'association $\mathtt{A} \rightarrow \mathtt{B}$, extraite d'une base de données $\mathcal{B}$, est déclarée de qualité, selon la mesure μ et le seuil σ_μ (souvent fixé par l'utilisateur), si $\mu(\mathtt{A} \rightarrow \mathtt{B}) \geq \sigma_\mu$. Ce mode de qualification des règles pose plusieurs questions légitimes : est-ce que la règle est le fruit du hasard, son évaluation est-elle significativement supérieure au seuil, serait-elle toujours valide si les données n'avaient pas été exactement ce qu'elles sont (i.e. si l'on souhaite prendre en compte le fait que les données sont bruitées ou évoluent) ou encore si le seuil d'acceptation avait été augmenté, même légèrement. Parallèlement, on peut se poser la question des règles, peut-être intéressantes, qui n'apparaissent pas à cause d'un seuil légèrement trop élevé. Ces questions débouchent sur la notion, intuitive, de robustesse d'une règle d'association i.e. de la sensibilité de son évaluation par rapport à des modifications, même mineures, de $\mathcal{B}$ et/ou σ_μ. Intuitivement encore, on sent bien que cette notion sera étroitement liée à l'ajout de contre-exemples et/ou à la perte d'exemples de la règle. L'étude des mesures en fonction, principalement, du nombre de contre-exemples prend ici un sens très important (Guillaume, 2000) : la décroissance des mesures en fonction du nombre de contre-exemples est un critère d'éligibilité (Lenca et al. (2003b)), tandis que la vitesse de décroissance dès l'apparition des premiers contre-exemples est une propriété qui peut être souhaitable ou non (Lenca et al. (2003a); Gras et al. (2004)). Nous renvoyons à Lenca et al. (2008) pour une étude de 20 mesures classiques sur ces deux caractéristiques.

Ainsi, si la question de la robustesse des règles d'association est légitime, celle-ci n'a fait l'objet, à notre connaissance, que de très peu de travaux. Ceux-ci se divisent principalement en trois grandes approches : la première est expérimentale et procède par simulation (Azé et Kodratoff (2002); Azé et al. (2003); Cadot (2005); Azé et al. (2007)), la seconde repose sur l'utilisation de tests statistiques (Lallich et Teytaud (2004); Rakotomalala et Morineau (2008); Cadot et Lelu (2007)), et la troisième est formelle et procède essentiellement par l'étude des dérivées des mesures (Lenca et al. (2006); Vaillant et al. (2006); Gras et al. (2007)). Notre proposition s'inscrit dans la lignée des méthodes formelles, mais se détache des précédentes par l'absence d'étude des dérivées.

Lenca et al. (2006) définissent la robustesse à partir du nombre d'exemples qu'une règle peut perdre tout en restant de qualité. Les auteurs proposent un premier modèle où les exemples sont transférés vers les contre-exemples. Ce modèle permet de développer un résultat formel pour 9 transformées linéaires et 4 transformées monotones de la confiance. Ce résultat illustre les différents comportements des 13 mesures étudiées par rapport à la table de contingence de la règle. Les auteurs mettent alors en évidence que l'origine des contre-exemples peut influencer beaucoup les variations des mesures et qu'une règle dominée au sens des mesures peut s'avérer plus robuste que la règle qui la domine. Ainsi Vaillant et al. (2006) étendent cette approche à trois modèles de contre-exemples. Une écriture des mesures selon les tables de contingence obtenues permet d'étudier le comportement des mesures à partir de leurs dérivées première et seconde. Le cas de dix mesures illustre les différents comportements. Par la suite, Gras et al. (2007) utilisent également le calcul différentiel pour l'étude de la stabilité de 4 mesures lorsque

les paramètres dont elles dépendent varient sensiblement au voisinage de leur observation.

Notre proposition, telle que présentée dans Le Bras et al. (2010), qui développe les idées présentées dans Lenca et al. (2006) et Vaillant et al. (2006), donne d'une part une définition précise de la notion de robustesse et d'autre part une mesure cohérente de la robustesse des règles d'association. Nous présentons en section 2 un rappel sur les règles d'association, la définition de la mesure de robustesse et son utilisation en pratique. Ensuite, en section 3 nous détaillons les expériences que nous avons menées ainsi que leurs résultats, et finalement nous concluons en section 4.

2 Vers une mesure formelle de la robustesse

Lors de travaux précédents (Le Bras et al., 2009b), nous nous sommes intéressés à un cadre formel d'étude des règles d'association et des mesures d'intérêt tel qu'initié par Hébert et Crémilleux (2007), dont le principal apport est d'associer une règle d'association à une projection dans le cube unité de $\mathbb{R}^3$.

2.1 Règles d'association et mesures d'intérêt

Notre approche s'appuie sur ce cadre, dont nous rappelons ici les notions principales. Notons r : $\mathtt{A} \rightarrow \mathtt{B}$ une règle d'association dans une base de données $\mathcal{B}$. Une mesure d'intérêt est une fonction qui associe à une règle d'association un nombre réel caractérisant l'intérêt que l'on peut porter à cette règle dans cette base de données. Dans cet article, nous nous intéressons exclusivement aux mesures d'intérêt objectives, c'est-à-dire aux mesures dont la valeur est déterminée par la table de contingence de r. La figure 1 présente une telle table de contingence, dans laquelle nous notons p_x la fréquence du motif $\mathtt{X}$.

	$\mathtt{B}$	$\overline{\mathtt{B}}$	
$\mathtt{A}$	p_{ab}	$p_{a\bar{b}}$	p_a
$\overline{\mathtt{A}}$	$p_{\bar{a}b}$	$p_{\bar{a}\bar{b}}$	$p_{\bar{a}}$
	p_b	$p_{\bar{b}}$	1

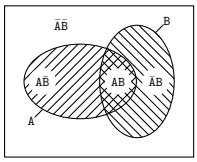


FIG. 1 – *Table de contingence de* r : $\mathtt{A} \rightarrow \mathtt{B}$

La table de contingence possédant trois degrés de liberté, une fois ces trois degrés choisis, il est possible de considérer les mesures comme des fonctions de $\mathbb{R}^3$ dans $\mathbb{R}$, et d'y appliquer tous les outils de l'analyse. Le Bras et al. (2009b) ont montré qu'il était possible d'établir un lien entre les propriétés algorithmiques de certaines mesures et leurs propriétés analytiques, notamment en ce qui concerne leurs variations. Pour pouvoir étudier les mesures comme de simples fonctions de trois variables, il est nécessaire de bien établir le domaine de définition. Ce domaine dépend de la paramétrisation choisie : à l'aide des exemples, des contre-exemples, ou encore de la confiance. Les présents travaux pourraient être menés d'une façon comparable pour chaque paramétrisation. Nous nous intéresserons ici au comportement des mesures vis-à-vis de la variation des contre-exemples des règles, dans la lignée des travaux de Vaillant et al. (2006) et Lenca et al. (2008), c'est-à-dire qu'une règle d'association r : $\mathtt{A} \rightarrow \mathtt{B}$ peut être caractérisée par les trois quantités $(p_{a\bar{b}}, p_a, p_b)$. Les mesures d'intérêt sont alors des fonctions

d'un sous-domaine $\mathcal{D}$ du cube unité de $\mathbb{R}^3$, dont la définition découle de l'interdépendance des quantités $(p_{a\bar{b}}, p_a, p_b)$. Celles-ci, en plus d'être comprises dans $[0, 1]$ sont soumises aux inégalités suivantes :

$$p_{a\bar{b}} \leq p_a \quad (1)$$
$$p_{a\bar{b}} \leq 1 - p_b \quad (2)$$
$$p_a - p_b \leq p_{a\bar{b}} \quad (3)$$

La définition exacte du domaine $\mathcal{D}$ est donc donnée par (Le Bras et al. (2009a)) :

$$\mathcal{D} = \left\{ (x, y, z) \;\middle|\; \begin{array}{ccccc} 0 & < & y & < & 1 \\ 0 & < & z & < & 1 \\ \max(0, y - z) & < & x & < & \min(y, 1 - z) \end{array} \right\}$$

où x (resp. y, z) représente $p_{a\bar{b}}$ (resp. p_a, p_b).

Lorsque l'on projette une règle dans $\mathbb{R}^3$, elle peut être étudiée comme un vecteur et on a alors la possibilité d'étudier un voisinage de la règle et d'observer le comportement d'une mesure sur ce voisinage. C'est sur cette idée que nous nous appuyons pour proposer une nouvelle caractérisation de la robustesse des règles d'association.

2.2 Une définition de la robustesse

Supposons que l'on cherche à évaluer les règles d'association extraites d'une base $\mathcal{B}$ à l'aide d'une mesure d'intérêt objective μ. L'utilisateur aura fixé un seuil, $\mu_{\min}$, au dessus duquel les règles sont jugées intéressantes. Les règles ainsi sélectionnées sont cependant dépendantes de plusieurs paramètres parmi lesquels :

- le seuil $\mu_{\min}$: l'utilisateur peut à tout moment modifier le seuil et faire ainsi apparaître, ou disparaître, un grand nombre de règles ;
- le bruit : la règle ne survivra peut-être pas à une variation des données, comme l'introduction de nouvelles transactions (taille de l'échantillon), ou bien la présence d'erreurs.

Nous proposons ici d'apporter une contribution à l'étude du second point, la fragilité de la règle par rapport aux variations des données. Vaillant et al. (2006) proposent plusieurs approches pour l'étude de la variation des mesures par rapport aux contre-exemples des règles. En s'appuyant sur différents modèles, ils proposent d'étudier les variations supportées par une règle afin qu'elle reste intéressante, ce qui ne permet pas de fournir un critère de robustesse général. Ces différents modèles nécessitent de connaître finement les variations des quantités du tableau de contingence, ce qui les rend difficilement applicables en pratique.

Notre vision de la robustesse est différente et s'appuie sur la notion de règle limite. Ces règles limites peuvent être abstraites, au sens où elles ne sont pas nécessairement réalisées dans la base $\mathcal{B}$. Nous définissons une distance sur les règles, $d_2(r, r')$, qui est la distance entre les deux projections de r et r' dans $\mathcal{D}$.

Définition 1 (Règle limite)**.** Une règle limite est une règle d'association $r_{\min}$, éventuellement abstraite, telle que $\mu(r_{\min}) = \mu_{\min}$. Soit r une règle d'association, on note r^* une règle limite qui minimise $d_2(r, r_{min})$ dans $\mathbb{R}^3$. Formellement,

$$r^* \in argmin\{d_2(r, r_{min}) | r_{min} \text{ règle limite}\}$$

Ce sont des règles qui, si elles étaient réalisées, seraient sélectionnées de justesse (par rapport au seuil $\mu_{\min}$). Pour une règle r donnée, r^* n'est pas unique, mais son choix n'est pas déterminant pour la notion de robustesse que nous allons définir par la suite.

Puisqu'une règle limite est une règle d'association, $r_{\min}$, associée à $(x_{\min}, y_{\min}, z_{\min})$, le triplet est nécessairement un élément de $\mathcal{D}$. Ainsi, $d_2(r, r^*)$ n'est pas simplement la distance de r à la surface $\mathcal{S}$ d'équation $\mu = \mu_{\min}$, mais la distance à $\mathcal{S} \cap \mathcal{D}$. La figure 2 montre les deux cas possibles de calcul de la robustesse.

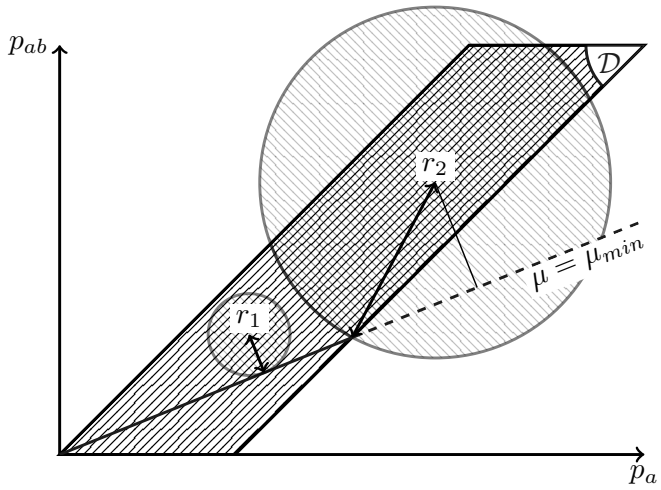


FIG. 2 – *Visualisation de la robustesse pour deux règles r_1 et r_2 à p_b fixé pour le cas particulier du plan $\mathcal{S}$ défini par la confiance.*

Définition 2 (Robustesse d'une règle)**.** Soit μ une mesure d'intérêt des règles d'association, et $\mu_{\min}$ un seuil prédéfini. Soit une base de données $\mathcal{B}$ et une règle d'association r sur cette base telle que $\mu(r) > \mu_{\min}$. On définit la robustesse de r par rapport à μ et $\mu_{\min}$ par :

$$\text{rob}_\mu(r, \mu_{\min}) = \frac{d_2(r, r^*)}{\sqrt{3}}$$

Le facteur important est le numérateur $d_2(r, r^*)$, la division par $\sqrt{3}$ est une normalisation de cette quantité pour la ramener à l'intervalle $[0, 1]$. D'autres normalisations sont évidemment envisageables. S'il n'y a pas d'ambiguïté, nous noterons cette robustesse $\text{rob}(r)$. Nous allons montrer dans le paragraphe suivant en quoi cette définition est une notion de robustesse, et quelques propriétés de la robustesse ainsi définie.

2.3 Propriétés de la robustesse

Commençons par justifier l'appellation de robustesse. Considérons une base $\mathcal{B}$ et une règle d'association r : $\texttt{A} \rightarrow \texttt{B}$ dans $\mathcal{B}$ telle que $\mu(r) > \mu_{\min}$. On note $(p_{a\bar{b}}, p_a, p_b)$ ses supports

associés. Introduisons du bruit dans la base $\mathcal{B}$ afin d'obtenir une base $\mathcal{B}'$ dans laquelle la règle r' : A $\rightarrow$ B est caractérisée par $(p'_{a\bar{b}}, p'_a, p'_b)$: les motifs restent identiques, mais leur support change. On suppose que l'on a des connaissances sur le bruit qui nous permettent d'assurer :

$$\begin{aligned} |p'_{a\bar{b}} - p_{a\bar{b}}| &\leq \frac{d_2(r, r^*)}{\sqrt{3}} \\ |p'_a - p_a| &\leq \frac{d_2(r, r^*)}{\sqrt{3}} \\ |p'_b - p_b| &\leq \frac{d_2(r, r^*)}{\sqrt{3}} \end{aligned}$$

Ainsi, $d_2(r, r') = \sqrt{|p'_{a\bar{b}} - p_{a\bar{b}}|^2 + |p'_a - p_a|^2 + |p'_b - p_b|^2} \leq d_2(r, r^*)$, et donc, par définition de r^*, $\mu(r') > \mu_{\min}$. rob(r) traduit donc la quantité de bruit acceptée par la règle tout en restant de qualité. C'est une notion de sécurité, qui permet d'affirmer que si le bruit est suffisamment contrôlé, la règle restera intéressante. L'inverse n'est cependant pas vrai car une règle peu robuste pourra évoluer de manière à devenir plus robuste.

Cette notion de robustesse est particulièrement facile à comprendre dans le cadre de bruit inséré par transactions. En effet, si l'on insère le bruit dans moins de rob$(r)\%$ des transactions, la règle r restera intéressante par rapport à μ_{min}. Cela peut correspondre à une base de données qui évolue, avec de nouvelles transactions, ou bien à des données insérées avec une possibilité d'erreur sur chaque transaction. Lorsque le bruit est inséré par attributs (Azé et Kodratoff, 2002; Azé et al., 2003), le contrôle est plus difficile à assurer.

Inversement, si l'on sait que la base de données contient un certain pourcentage de bruit, et que l'on extrait de cette base bruitée des règles dont la robustesse assure l'intérêt pour ce pourcentage de bruit, alors l'utilisateur est assuré que ces règles sont effectivement intéressantes dans la base *idéale* non bruitée. On peut par exemple penser à un système de capteurs physiques qui possèderaient des marges d'erreurs garanties par le constructeur, ou bien au cas de bases dont la qualité a été évaluée (Berti-Equille, 2007).

Propriété 1. *La robustesse $rob(r)$ présente des caractéristiques analytiques intéressantes :*
- *la robustesse d'une règle est un réel de $[0, 1]$;*
- *rob$_\mu(r, \mu_{\min}) = 0$ si r est une règle limite, c'est-à-dire si $\mu(r) = \mu_{\min}$;*[1]
- *si la mesure μ, vue comme fonction de 3 variables, est continue de $\mathcal{D} \subset \mathbb{R}^3$ dans $\mathbb{R}$, alors la robustesse est décroissante par rapport à μ_{min} ;*
- *la robustesse est continue par rapport à r.*

Ces propriétés permettent de déduire des comportements attendus de la notion de robustesse. Ainsi, plus le seuil est fixé haut, moins les règles sont robustes, et plus il est important d'avoir des données fiables. D'autre part, deux règles dont les projetés sont proches ont des robustesses équivalentes.

2.4 Évaluer la robustesse

Le calcul de cette robustesse fait naturellement appel à un calcul de distance à une surface sous certaines contraintes. Il existe un certain nombre de mesures pour lesquelles le calcul de

[1] Il faut noter que la valeur rob$_\mu(r, \mu_{\min}) = 1$ est une valeur théorique qui correspond à une configuration très particulière de r, $\mu_{\min}$ et de μ. En pratique dans nos expériences nous n'avons pas rencontré cette valeur.

la distance se ramène à un calcul de distance à un plan. Nous nous intéressons ici uniquement à ces mesures, que nous appelons mesures planes. Les mesures plus complexes (e.g. Klosgen, force collective, spécificité relative) demandent de recourir à des techniques d'analyse numérique, et ne seront pas développées ici.

Définition 3 (Mesure plane)**.** Une mesure d'intérêt μ est dite plane si la surface définie par $\mu(r) = \mu_{min}$ est un plan.

C'est en particulier le cas de mesures telles que Sebag-Schoenauer, taux d'exemples-contre-exemple, Jaccard, contramin, precision, recall, spécificité. Dans ce cas, la géométrie euclidienne permet de calculer la distance au plan $\mathcal{P} : ax + by + cz + d = 0$ d'une règle r de coordonnées (x_1, y_1, z_1) est donnée par :

$$d(r_1, \mathcal{P}) = \frac{|ax_1 + by_1 + cz_1 + d|}{\sqrt{a^2 + b^2 + c^2}}$$

Il reste cependant à prendre en compte que r^* doit appartenir au domaine $\mathcal{D}$. C'est donc en fait la distance au polygone intersection $\mathcal{P} \cap \mathcal{D}$ qui nous intéresse réellement. Encore une fois, ce calcul est aisément réalisable : il suffit pour cela de déterminer les points formant les sommets de ce polygone (convexe), puis de calculer la distance à chaque côté (en tant que segment). La distance au périmètre du polygone sera la plus petite de ces distances. On obtient donc l'algorithme suivant de calcul de la robustesse dans le cas d'une mesure plane :

- Trouver $r^\perp$, projection orthogonale de r sur $\mathcal{P}$;
- Si $r^\perp \in \mathcal{D}$, $r^* = r^\perp$ et renvoyer $d_2(r, r^*)$;
- Sinon, renvoyer la distance au périmètre du polygone intersection.

Nous avons décidé lors de nos expérimentations de nous concentrer sur ce type de mesures, car elles permettent d'obtenir des résultats précis ne faisant pas appel à des algorithmes d'approximation. L'étude dans le cas de mesures non planes fera l'objet de travaux ultérieurs.

Exemple 1. Les mesures suivantes sont planes. Leur ligne de niveau $\mu = \mu_0$ définit les plans suivants :

- mesure de confiance : $x - (1 - \mu_0)y = 0$;
- mesure de Sebag-Shoenauer : $(1 + \mu_0)x - y = 0$;
- taux d'exemples-contre-exemples : $(2 - \mu_0)x - (1 - \mu_0)y = 0$;
- mesure de Jaccard : $(1 + \mu_0)x - y + \mu_0 z = 0$.

Approfondissons le cas de la confiance. Dans une paramétrisation par les contre-exemples, le plan défini par le seuil de confiance $conf_{min}$ est $\mathcal{P} : x - (1 - conf_{min})y = 0$. La distance à ce plan d'une règle r de projection (x_1, y_1, z_1) et de confiance $conf(r) > conf_{min}$ sera alors donnée par

$$d = y_1 \frac{conf(r) - conf_{min}}{\sqrt{1 + (1 - conf_{min})^2}} \tag{4}$$

La robustesse dépend donc, à $conf_{min}$ fixé, de deux paramètres : y_1, le support de l'antécédent et $conf(r)$, la mesure de la règle. Ainsi, deux règles ayant la même confiance peuvent avoir des robustesses très différentes. De même, deux règles ayant la même robustesse peuvent avoir des confiances différentes. Il ne sera donc pas étonnant d'observer des règles de mesure faible, et de robustesse élevée, tout comme des règles de mesure élevée, mais de robustesse très faible. En effet, il est possible de découvrir une règle qui soit très intéressante, mais très fragile.

Exemple 2. Considérons une base fictive de 100000 transactions. On note n_x le nombre d'occurrences du motif `X`. Dans cette base, on trouve une première règle r_1 : `A` $\rightarrow$ `B` telle que $n_a = 100$ et $n_{a\bar{b}} = 1$. Sa confiance est de 99%. Mais sa robustesse comme définie précédemment au seuil de confiance 0.8 est $\text{rob}(r_1) = 0.0002$. Une seconde règle r_2 : `C` $\rightarrow$ `D` présente les caractéristiques suivantes : $n_c = 50000$ et $n_{c\bar{d}} = 5000$. Sa confiance n'est que de 90% mais sa robustesse de 0.05. Elle présente pourtant plus de contre-exemples proportionnellement à son antécédent que r_1 et pourrait être jugée, à tort selon la robustesse, moins fiable.

Dans le premier cas, la règle limite la plus proche a comme caractéristiques $n^*_a = 96$ et $n^*_{a\bar{b}} = 19$. La règle originale ne supporte donc que de très petites variations sur les lignes de la base de données. La seconde règle a quant à elle une plus proche règle limite de paramètres $n_c = 49020$ et $n_{c\bar{d}} = 9902$ et la règle originale accepte donc de l'ordre du millier de changements. La règle r_2 est donc beaucoup moins sensible au bruit que la règle r_1. Pourtant, c'est cette règle r_1 qui présentait le plus fort intérêt selon la mesure de confiance.

Il convient donc de se poser la question du réel intérêt d'une règle : comment doit-on arbitrer entre une règle d'association très bien évaluée par les mesures, mais de robustesse très faible, et une règle moins bien évaluée, mais dont la robustesse nous assure une plus grande fiabilité vis-à-vis du bruit ?

2.5 Applications pratiques de la robustesse

La robustesse définie précédemment peut avoir deux applications immédiates. La première concerne la classification de règles : cette mesure permet de comparer deux règles entre elles et donc d'établir un préordre sur l'ensemble de règles concerné. La seconde concerne le filtrage des règles situées au-dessus d'un certain seuil fixé par l'utilisateur. Cependant, au même titre que le seuil de mesure, le seuil de robustesse peut s'avérer délicat à fixer. La figure 3 montre

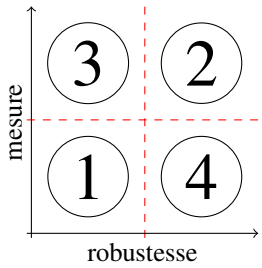


FIG. 3 – *Zones remarquables entre robustesse et mesure*

que l'on peut distinguer quatre comportements types. S'il est facile d'arbitrer entre deux règles (robuste/intéressante) (2) et (fragile/peu intéressante) (1), la tâche est moins évidente entre deux règles (robuste/peu intéressante) (4) et (fragile/intéressante) (3). Vaut-il mieux avoir une règle très intéressante, mais très dépendante du bruit dans les données, ou bien est-il préférable d'avoir une règle très robuste, qui supportera des changements dans les données, mais dont la mesure est proche du seuil fixé ? Les réponses à cette question dépendent évidement de la situation pratique et de la confiance que l'utilisateur a dans la qualité des données.

Dans la suite, nous montrons que les graphiques robustesse/mesure font apparaître un grand nombre de règles robustes, mais dominées en terme de mesure par des règles moins robustes.

3 Mise en oeuvre de la robustesse

Nous présentons ici les résultats obtenus sur 4 bases et pour 5 mesures planes. Dans un premier temps, nous présentons le protocole expérimental choisi, puis nous étudions les graphiques obtenus afin de mettre en évidence les liens entre mesure et robustesse. Enfin, nous analysons l'effet du bruit sur les règles d'association.

3.1 Protocole expérimental

3.1.1 L'extraction des règles

Comme expliqué précédemment, nous nous intéressons au cas de mesures planes. Nous en avons retenues 5 : confiance, Jaccard, Sebag-Shoenauer, taux d'exemples-contre-exemples, et la spécificité. Le tableau 1 rappelle leur écriture en fonction des contre-exemples, ainsi que le plan qu'elles définissent.

nom	formule	plan
confiance	$\frac{p_a - p_{a\bar{b}}}{p_a}$	$x - (1-\mu_0)y = 0$
Jaccard	$\frac{p_a - p_{a\bar{b}}}{p_b + p_{a\bar{b}}}$	$(1+\mu_0)x - y + \mu_0 z = 0$
Sebag-Shoenauer	$\frac{p_a - p_{a\bar{b}}}{p_{a\bar{b}}}$	$(1+\mu_0)x - y = 0$
spécificité	$\frac{1 - p_b - p_{a\bar{b}}}{1 - p_a}$	$x - \mu_0 y + z = 1 - \mu_0$
taux exemples-contre-exemples	$1 - \frac{p_{a\bar{b}}}{p_a - p_{a\bar{b}}}$	$(2-\mu_0)x - (1-\mu_0)y = 0$

TAB. 1 – *Les mesures planes retenues avec leur écriture par rapport aux contre-exemples, le plan défini par une valeur* μ_0

Pour effectuer nos expériences, nous nous sommes appuyés sur 4 bases de données usuelles (Asuncion et Newman, 2007). Census a été discrétisée, et nous en avons extrait, ainsi que de Mushroom, des règles de classe, c'est-à-dire où le conséquent est contraint. Les bases Chess et Connect ont été binarisées afin d'en extraire des règles d'association sans contrainte. Les règles ont ensuite été extraites, grâce à l'implémentation d'APRIORI de Borgelt et Kruse (2002), de manière à obtenir des règles de support non nul, de confiance supérieure à 0.8 et de taille variable en fonction de la base. L'ensemble de ces informations est synthétisé dans la table 2. Nous avons ainsi obtenu des règles intéressantes, sans exclure les pépites de connaissance, mais tout en gardant un nombre de règles raisonnable. Notons que Borgelt (2003) présente de manière détaillée les caractéristiques de ces différentes bases.

3.1.2 Calcul de la robustesse

Pour chaque ensemble de règles, et chaque mesure, nous avons appliqué la même méthode de calcul de la robustesse des règles d'association extraites des bases. Dans un premier temps,

base	attributs	transactions	type	taille	# règles
census	137	48842	classe	5	244487
chess	75	3196	sans contrainte	3	56636
connect	129	67557	sans contrainte	3	207703
mushroom	119	8124	classe	4	42057

TAB. 2 – *Bases de données utilisées dans nos expériences. L'avant-dernière colonne fixe la taille maximum des règles extraites.*

nom	seuil
confiance	0.984
Jaccard	0.05
Sebag-Shoenauer	10
spécificité	0.5
taux exemples-contre-exemples	0.95

TAB. 3 – *Les mesures planes retenues et le seuil choisi.*

nous avons sélectionné uniquement les règles dont la mesure était supérieure à un seuil prédéfini. Nous avons choisi de fixer ce seuil définitivement pour toutes les bases aux valeurs indiquées table 3. Ces seuils ont été fixés après observation du comportement des mesures sur les règles extraites de la base Mushroom, afin d'obtenir des règles intéressantes et des règles inintéressantes dans des proportions équilibrées.

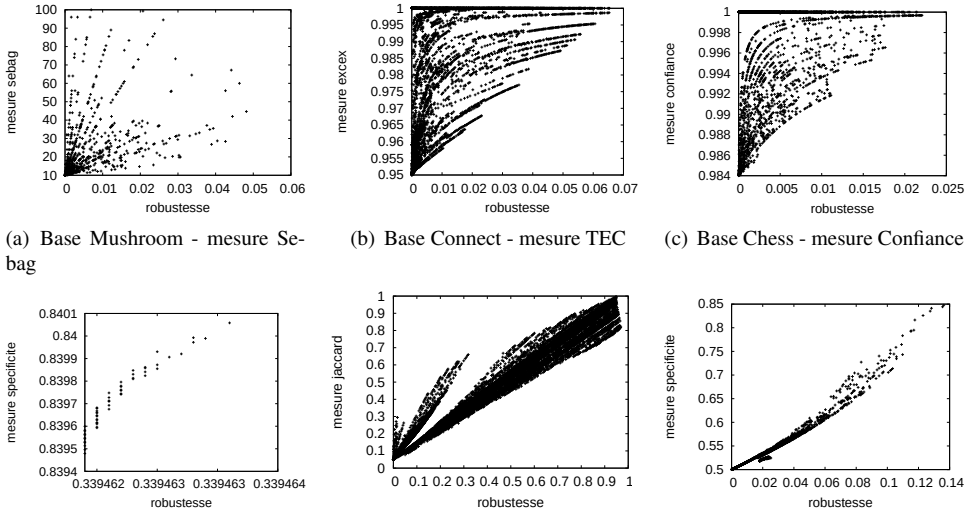
Nous avons ensuite implémenté un algorithme s'appuyant sur la description faite dans la section 2.4 pour le cas spécifique des mesures planes, calculant la robustesse d'une règle par rapport à une mesure et son seuil. Nous obtenons en sortie une liste de règles avec leur support, leur robustesse et leur mesure. La complexité de cet algorithme dépend essentiellement du nombre de règles à analyser. Nous ne nous attarderons pas ici sur la complexité de tels algorithmes, car il s'agit simplement d'analyser un ensemble de règles. Ces résultats nous permettent d'obtenir des graphiques mesure/robustesse que nous analyserons dans la partie 3.2.

3.1.3 Insertion du bruit

Comme indiqué précédemment, nous analysons l'influence du bruit sur les règles en fonction de leur robustesse. Nous avons donc mis en place une procédure d'insertion de bruit dans une base de données. Notre choix s'est porté sur un bruit introduit par ligne. Nous avons décidé d'introduire du bruit dans 5% des lignes de chaque base en sélectionnant les lignes bruitées de manière aléatoire, et en modifiant de manière aléatoire les valeurs des attributs de ces lignes (tirage équiprobable sans remise parmi les valeurs apparues). Une fois le bruit inséré nous calculons les nouveaux supports des règles de l'ensemble initial. Nous extrayons les règles intéressantes au sens des mesures données, et évaluons leur robustesse. L'étude du bruit est discutée dans la partie 3.3

3.2 Analyse de la robustesse

Nous avons obtenu, pour chaque base et chaque mesure, des données nous permettant de visualiser la mesure d'une règle en fonction de sa robustesse. La figure 4 propose un échantillon représentatif des résultats, dans le sens ou l'allure des graphiques est sensiblement la même pour toutes les bases, pour une mesure donnée. Plusieurs points peuvent être relevés.



(a) Base Mushroom - mesure Sebag

(b) Base Connect - mesure TEC

(c) Base Chess - mesure Confiance

(d) Base Census - mesure Spécificité

(e) Base Connect - mesure Jaccard

(f) Base Mushroom - mesure Spécificité

FIG. 4 – *Valeur de la mesure en fonction de la robustesse pour différents couples base/mesure.*

Dans un premier temps, la mesure possède un caractère globalement croissant avec la robustesse. Cependant si l'on observe précisément, il est bien visible qu'un très grand nombre de règles sont dominées au sens de la mesure par des règles pourtant moins robustes. Cela est particulièrement marqué dans le cas de la mesure de Sebag, puisqu'une règle de mesure de Sebag de valeur 100 peut être beaucoup moins robuste (10^{-4}) qu'une règle de mesure 20 ($2 \cdot 10^{-3}$). La seconde suportera vingt fois plus de changements que la première.

Ensuite, nous observons des lignes de niveau dans la plupart des cas. Sebag et Jaccard présentent des droites de niveau, la Confiance et TEC présentent des courbes concaves, et la Spécificité semble présenter des courbes convexes.

Traitons le cas particulier des courbes relatives à la confiance. Une telle démonstration peut se faire pour les autres mesures. L'équation (4) montre l'écriture de la robustesse en fonction de la mesure, où y représente p_a. Exprimons $\mu(r)$ en fonction de d et de x en utilisant le fait que $p_a = \frac{p_{a\bar{b}}}{1-conf(r)}$:

$$\mu(r) = \frac{\mu_{min} + \sqrt{1 + (1-\mu_{min})^2} * \frac{d}{x}}{1 + \sqrt{1 + (1-\mu_{min})^2} * \frac{d}{x}} \tag{5}$$

Ainsi, à x constant, c'est-à-dire à nombre de contre-exemples constant, les règles se trouvent sur une courbe bien définie, concave et croissante. Les lignes de niveau observées dans le cas de la confiance sont donc formées par des règles ayant le même nombre de contre-exemples.

Un comportement paraît récurrent quel que soit la mesure : il ne semble pas exister de règle qui soit à la fois très proche du seuil de mesure et très robuste. Seule Sebag se distingue un peu de ce comportement. Nous pensons que cela est fortement lié au fait que les mesures étudiées ici sont des mesures planes. En effet, dans ce cas, la variation des lignes de niveau est constante et ne présente pas de fortes pentes.

3.3 Etude de l'influence du bruit

Nous allons ici étudier les liens entre l'introduction de bruit et l'évolution des ensembles de règles par rapport à la robustesse. Nous avons décidé de créer 5 bases bruitées à partir de chaque base initiale, puis pour chaque base bruitée (méthode présentée 3.1.3), d'étudier la robustesse des règles qui sont conservées, et des règles qui ont disparues. Pour valider notre notion de robustesse, nous attendons de ces expériences d'observer une robustesse plus faible dans l'ensemble des règles disparues que dans l'ensemble des règles conservées. La table 4

(a) mesure TEC

base	disparues	conservées
census	0.83e-6	0.79e-6
chess	1.16e-3	0.96e-2
connect	5.26e-4	7.72e-3
mushroom	9.4e-5	6.6e-4

(b) mesure de Sebag

base	disparues	conservées
census	1.53e-6	1.53e-6
chess	1.63e-3	1.72e-2
connect	8.38e-4	1.42e-2
mushroom	1.28e-4	1.22e-3

(c) mesure de spécificité

base	disparues	conservées
census	0	0.19
chess	7.23e-5	8.76e-2
connect	0	1.2e-1
mushroom	2.85e-4	1.37e-2

(d) mesure de confiance

base	disparues	conservées
census	2.61e-7	2.61e-7
chess	5.59e-4	3.77e-3
connect	2.16e-4	2.73e-3
mushroom	5.51e-5	2.34e-4

(e) mesure de Jaccard

base	disparues	conservées
census	0	0
chess	3.2e-4	1.69e-1
connect	1.94e-3	1.43e-1
mushroom	3.20e-4	1.90e-2

TAB. 4 – *Comparaison entre les robustesses moyennes des règles disparues et conservées pour les différentes mesures*

montre les résultats obtenus en faisant la moyenne des robustesses des règles au sein des différents ensembles de règles obtenus, sur les 5 bruitages.

Dans la plupart des cas apparaît un facteur 10 entre la robustesse des règles conservées et des règles disparues. Seul le cas de la base de données Census pour les mesures TEC, Sebag et confiance ne confirme pas ce résultat, mais le comportement de Census ne contredit pas pour autant notre théorie. En effet, les robustesses initiales issues de la base de données Census sont de l'ordre de 10^{-6}, et sont donc vulnérables à 5% de bruit (de l'ordre de $10^{-}2$). Il est donc normal que toutes les règles soient susceptibles de devenir inintéressantes.

A l'opposé, la mesure de spécificité fait apparaître un comportement commun à la base Census et à la base Connect. Pour ces deux bases, aucune règle ne disparaît lorsque l'on introduit 5% de bruit. Si l'on regarde la moyenne de la robustesse des règles conservées, on s'aperçoit qu'elle est bien supérieure aux 5%, ce qui signifie que toutes les règles sont protégées. Dans le cas de la base Census, la plus petite mesure de spécificité relevée est de 0.839, donc bien au dessus du seuil fixé. Il n'est donc pas étonnant que les règles de la base Census

soient protégées du bruit. Dans le cas de la base Connect, la moyenne des mesures observées est de 0.73 avec un écart type de 0.02. la plus petite mesure de spécificité est de 0.50013 et correspond à une robustesse de $2.31e-5$. Pourtant elle a bien été sauvée dans les 5 tirages de bruit effectués. Cela permet de souligner le fait que notre définition de la robustesse correspond à la définition d'un périmètre de sécurité autour de la règle. Si la règle change et sort de ce périmètre, son évolution peut se faire librement dans l'espace sans atteindre la surface seuil. Cependant, le risque persiste.

4 Conclusion

La robustesse des règles d'association est un sujet important, qui n'a été que peu traité par des approches formelles. Réussir à caractériser la robustesse d'une règle, c'est s'offrir une assurance sur son intérêt, et donc être capable de donner des informations sécurisées à l'utilisateur. Nous avons proposé dans cet article une nouvelle notion de robustesse opérationnelle, dépendante d'une mesure et d'un seuil, et nous avons montré en quoi cette notion traduisait l'intuition naturelle du mot robustesse.

Nous traitons le cas particulier des mesures planes qui autorisent une caractérisation formelle de la notion de robustesse. Les résultats des expériences menées illustrent la théorie proposée. Nous envisageons de mettre en place un protocole de calcul de la robustesse sur une mesure quelconque, ce qui nécessite d'avoir recours à des méthodes numériques. Enfin, l'application de notre approche pour des tâches de classification est une perspective intéressante.

Références

Agrawal, R., T. Imieliski, et A. Swami (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD International Conference on Management of Data, Washington, D.C., United States*, pp. 207–216.

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules in large databases. In *20th International Conference on Very Large Data Bases, Santiago de Chile, Chile*, pp. 478–499.

Asuncion, A. et D. Newman (2007). UCI machine learning repository, http ://www.ics.uci.edu/∼mlearn/MLRepository.html.

Azé, J., S. Guillaume, et P. Castagliola (2003). Evaluation de la résistance au bruit de quelques mesures quantitatives. *n^o spécial RNTI Entreposage et fouille de données*, 159–170.

Azé, J. et Y. Kodratoff (2002). Evaluation de la résistance au bruit de quelques mesures d'extraction de règles d'association. In *2nd Extraction et Gestion des Connaissances conference, Montpellier, France*, pp. 143–154.

Azé, J., P. Lenca, S. Lallich, et B. Vaillant (2007). A study of the robustness of association rules. In *The 2007 Intl. Conf. on Data Mining, Las Vegas, Nevada, USA*, pp. 163–169.

Berti-Equille, L. (2007). Measuring and modelling data quality for quality-awareness in data mining. In *Quality Measures in Data Mining*, pp. 101–126.

Borgelt, C. (2003). Efficient implementations of APRIORI and ECLAT. In *1st Workshop on Frequent Item Set Mining Implementations*.

Borgelt, C. et R. Kruse (2002). Induction of association rules : APRIORI implementation. In *15th Conference on Computational Statistics, Berlin, Germany*, pp. 395–400.

Brin, S., R. Motwani, et C. Silverstein (1997). Beyond market baskets : Generalizing association rules to correlations. In *ACM SIGMOD International Conference on Management of Data, Tucson, Arizona, USA*, pp. 265–276.

Cadot, M. (2005). A simulation technique for extracting robust association rules. In *Computational Statistics & Data Analysis*, Limassol, Chypre.

Cadot, M. et A. Lelu (2007). Simuler et épurer pour extraire les motifs sûrs et non redondants. In *3rd Workshop on Qualité des Données et des Connaissances, Namur, Belgium* , pp. 15–24.

Geng, L. et H. J. Hamilton (2006). Interestingness measures for data mining : A survey. *ACM Computing Surveys 38*(3, Article 9).

Geng, L. et H. J. Hamilton (2007). Choosing the right lens : Finding what is interesting in data mining. In *Quality Measures in Data Mining*, pp. 3–24.

Goethals, B. (2005). Frequent set mining. In *The Data Mining and Knowledge Discovery Handbook*, pp. 377–397.

Gras, R., R. Couturier, J. Blanchard, H. Briand, P. Kuntz, et P. Peter (2004). Quelques critères pour une mesure de qualité de règles d'association - un exemple : l'intensité d'implication. *n^o spécial RNTI Mesures de qualité pour la fouille de données*, 3–31.

Gras, R., J. David, F. Guillet, et H. Briand (2007). Stabilité en A.S.I. de l'intensité d'implication et comparaisons avec d'autres indices de qualité de règles d'association. In *3rd Workshop on Qualité des Données et des Connaissances, Namur Belgium* , pp. 35–43.

Guillaume, S. (2000). *Traitement des données volumineuses*. Ph. D. thesis, U. de Nantes.

Guillaume, S., D. Grissa, et E. Mephu Nguifo (2010). Propriétés des mesures d'intérêt pour l'extraction des règles. In *Atelier QDC/EGC*, pp. 15–28.

Han, J., H. Cheng, D. Xin, et X. Yan (2007). Frequent pattern mining : current status and future directions. *Data Mining and Knowledge Discovery*, 55–86.

Hébert, C. et B. Crémilleux (2007). A unified view of objective interestingness measures. In *5th Intl. Conf. on Machine Learning and Data Mining, Leipzig, Germany*, pp. 533–547.

Hilderman, R. J. et H. J. Hamilton (2000). Applying objective interestingness measures in data mining systems. In *4th European Conference on Principles of Data Mining and Knowledge Discovery, Lyon, France*, pp. 432–439.

Hipp, J., U. Güntzer, et G. Nakhaeizadeh (2000). Algorithms for association rule mining — a general survey and comparison. *SIGKDD Explorations*, 58–64.

Lallich, S. et O. Teytaud (2004). évaluation et validation de l'intérêt des règles d'association. *n^o spécial RNTI Mesures de qualité pour la fouille de données*, 193–218.

Lallich, S., O. Teytaud, et E. Prudhomme (2007). Association rule interestingness : Measure and statistical validation. In *Quality Measures in Data Mining*, pp. 251–275.

Le Bras, Y., P. Lenca, et S. Lallich (2009a). On optimal rules discovery : a framework and a necessary and sufficient condition of antimonotonicity. In *13th Pacific-Asia Conference on*

Knowledge Discovery and Data Mining, Bangkok, Thailand, pp. 705–712.

Le Bras, Y., P. Lenca, S. Lallich, et S. Moga (2009b). Généralisation de la propriété de monotonie de la all-confidence pour l'extraction de motifs intéressants non fréquents. In *5th Workshop on Qualité des Données et des Connaissances, Strasbourg, France*, pp. 17–24.

Le Bras, Y., P. Meyer, P. Lenca, et S. Lallich (2010). Mesure de la robustesse de règles d'association. In *6th Workshop on Qualité des Données et des Connaissances, Hammamet, Tunisie*, pp. 27 – 38.

Lenca, P., S. Lallich, et B. Vaillant (2006). On the robustness of association rules. In *2nd IEEE International Conference on Cybernetics and Intelligent Systems and Robotics, Automation and Mechatronics, Bangkok, Thailand*, pp. 596 – 601.

Lenca, P., P. Meyer, P. Picouet, et B. Vaillant (2003a). Aide multicritère à la décision pour évaluer les indices de qualité des connaissances – modélisation des préférences de l'utilisateur. In *3rd Extraction et Gestion des Connaissances conference, Lyon, France*, pp. 271–282.

Lenca, P., P. Meyer, P. Picouet, B. Vaillant, et S. Lallich (2003b). Critères d'évaluation des mesures de qualité en ECD. *n^o spécial RNTI Entreposage et fouille de données*, 123–134.

Lenca, P., P. Meyer, B. Vaillant, et S. Lallich (2008). On selecting interestingness measures for association rules : user oriented description and multiple criteria decision aid. *European Journal of Operational Research 184*, 610–626.

Lenca, P., P. Meyer, B. Vaillant, P. Picouet, et S. Lallich (2004). Évaluation et analyse multicritère des mesures de qualité des règles d'association. *n^o spécial RNTI Mesures de qualité pour la fouille de données*, 219–246.

Lenca, P., B. Vaillant, P. Meyer, et S. Lallich (2007). Association rule interestingness measures : experimental and theoretical studies. In *Quality Measures in Data Mining*, pp. 51–76.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, pp. 229–248.

Rakotomalala, R. et A. Morineau (2008). The TVpercent principle for the counterexamples statistic. In *Statistical Implicative Analysis, Theory and Applications*, pp. 449–462. Springer.

Suzuki, E. (2008). Pitfalls for categorizations of objective interestingness measures for rule discovery. In *Statistical Implicative Analysis, Theory and Applications*, pp. 383–395. Springer.

Vaillant, B., S. Lallich, et P. Lenca (2006). Modeling of the counter-examples and association rules interestingness measures behavior. In *The 2006 Intl. Conf. on Data Mining, Las Vegas, Nevada, USA*, pp. 132–137.

Vaillant, B., P. Lenca, et S. Lallich (2004). A clustering of interestingness measures. In *7th International Conference on Discovery Science, Padova, Italy*, pp. 290–297.

Annexe 1

Propriété 2 (Décroissance par rapport à μ_{min})**.** *Soit μ une mesure d'intérêt continue en tant que fonction de $\mathbb{R}^3$ dans $\mathbb{R}$.*
Soit r une règle d'association dans une base $\mathcal{B}$, et μ_1 et μ_2 deux seuils tels que $\mu(r) > \mu_2 >$

μ_1. *Alors*

$$rob_\mu(r, \mu_1) > rob_\mu(r, \mu_2).$$

Démonstration. Considérons $r_1^* \in \arg\min\{d_2(r, r_{min}) | \mu(r_{min}) = \mu_1\}$, et plaçons nous sur le segment $s = [r, r^*]$. La mesure μ étant continue, sa restriction au segment s est continue. On applique alors le Théorème des Valeurs Intermédiaire : puisque $\mu(r) > \mu_2 > \mu_1 = \mu(r^*)$, il existe un point $r_2 \in s$ tel que $\mu(r_2) = \mu_2$. On obtient donc l'inégalité

$$d_2(r, r_2) < d_2(r, r_1^*).$$

Notons $r_2^* \in \arg\min\{d_2(r, r_{min}) | \mu(r_{min}) = \mu_2\}$. Par définition, nous obtenons donc l'inégalité

$$d_2(r, r_2^*) < d_2(r, r_2)$$

Revenons alors à la définition de la robustesse et remarquons que $d_2(r, r_1^*) = rob_\mu(r, \mu_1)$ et que $d_2(r, r_2^*) = rob_\mu(r, \mu_2)$. Les inégalités obtenues précédemment impliquent donc

$$rob_\mu(r, \mu_2) = d_2(r, r_2^*) < d_2(r, r_2) < d_2(r, r_1^*) = rob_\mu(r, \mu_1).$$

En conclusion, nous avons montré que si μ est continue et si $\mu(r) > \mu_2 > \mu_1$ alors $rob_\mu(r, \mu_1) > rob_\mu(r, \mu_2)$. □

Annexe 2

Propriété 3. *La robustesse est continue par rapport à r.*

Démonstration. La robustesse est définie par une distance à une partie dans un espace métrique. A ce titre, elle est continue par rapport à r. □

Summary

In this article we give a formal definition of the robustness of association rules, based on a model from our previous work. We think that it is a central concept in the evaluation of the rules which has only been studied unsatisfactorily up to now. It is crucial because we have observed that a good rule (according to a given quality measure) might turn out as a very fragile rule with respect to small variations in the data. The robustness measure that we propose here depends on the selected quality measure, the value taken by the rule and the minimal acceptance threshold chosen by the user. We present a few properties of this robustness, detail its use in practice and show the outcomes of various experiments. All in all, we present a new perspective on the evaluation of association rules.